ACII 2009

Affective Computing and Intelligent Interaction

PROCEEDINGS OF THE DOCTORAL CONSORTIUM 2009

Amsterdam, Sept 10-12, 2009

Alessandro Vinciarelli, Catherine Pelachaud,
Roddy Cowie and Anton Nijholt (eds.)

CIP GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Vinciarelli A., Pelachaud C., Cowie R., Nijholt A.

Affective Computing and Intelligent Interaction
Proceedings of the Doctoral Consortium 2009
A. Vinciarelli, C. Pelachaud, R. Cowie, A. Nijholt (eds.)
Amsterdam, Universiteit Twente, Faculteit Elektrotechniek, Wiskunde en Informatica

ISSN 0929–0672
CTIT Workshop Proceedings Series WP09-13

trefwoorden: Affective computing, Affective signal processing, Emotions, Appraisal theory, Social Network Analysis, Social signal processing, Non-verbal communication, Facial expression, Facial behaviors analysis, Intelligent agents, Embodied conversational agents, Dialogue acts, Human-computer interaction, Human-robot interaction, Brain-Computer interfaces, Meeting reordings, Broadcast data, Recommender systems.

Book orders:
Ms. C. Bijron
University of Twente
Faculty of Electrical Engineering, Mathematics and Computer Science
P.O. Box 217
NL 7500 AE Enschede
tel: +31 53 4893740
fax: +31 53 4893503
Email: bijron@cs.utwente.nl

Druk- en bindwerk: Ipskamp Drukkers, Enschede.

Preface

This volume collects the contributions presented at the ACII 2009 Doctoral Consortium, the event aimed at gathering PhD students with the goal of sharing ideas about the theories behind affective computing; its development; and its application.

Published papers have been selected out a large number of high quality submissions covering a wide spectrum of topics including the analysis of human-human, human-machine and human-robot interactions, the analysis of physiology and nonverbal behavior in affective phenomena, the effect of emotions on language and spoken interaction, and the embodiment of affective behaviors.

The participants have actively contributed to the success of the event not only with their articles, but also with their presentations and the refreshing discussions during which they have compared their approaches, discussed future research problems, and received feedback from the international community.

We hope the Doctoral Consortium has been a chance to formulate interesting research questions, to develop collaborative relationships with other members of the ACII community, and to acquire awareness of the state-of-the-art in our vibrant domains.

The Doctoral Consortium included the presentation of the first ”Fiorella de Rosis” Award as well. The award is given by the HUMAINE Association to commemorate one of the outstanding figures in the field of emotion and computing. She was a founder member of the Association, and co-chair of the first ACII Doctoral Consortium (in Lisbon, 2007). Her research made the fundamental point that emotion needs to be integrated into logical models of argument. She was also one of the field’s idealists, always ready to speak out when she felt that others were settling for the least awkward solution rather than the best. She combined intellect and conviction with genuine warmth, and her death in 2008 was deeply felt throughout the community.

We take this opportunity to thank all the people that have helped to make this Doctoral Consortium possible, the General Chairs of ACII 2009, the members of the Program Committee, and the reviewers. Furthermore, we acknowledge the European Network of Excellence SSPNet (www.sspnet.eu) that has supported the participation of some of the students. The editors are grateful to Hendri Hondorp who did the final technical editing of the proceedings.

Alessandro Vinciarelli, Catherine Pelachaud, Roddy Cowie and Anton Nijholt — Amsterdam, September 2009

Doctoral Consortium Committee

Roddy Cowie (Queens University Belfast, United Kingdom)
Catherine Pelachaud (CNRS, France)
Alessandro Vinciarelli (Idiap Research Institute, Switzerland)

Program Committee

Shazia Afzal (University of Cambridge)
Barbara Caputo (Idiap Research Institute)
Ginevra Castellano (Queen Mary University London)
Alfred Dielmann (Idiap Research Institute)
Didier Grandjean (University of Geneva)
Hatice Gunes (Imperial College London)
Jennifer Hanratty (Queen's University Belfast)
Dirk Heylen (University of Twente)
Kostas Karpousis (Technical University of Athens)
Margaret McRorie (Queen's University of Belfast)
Daniela Romano (University of Sheffield)
Ioana Vasilescu (LIMSI-CNRS)
Gualtiero Volpe (University of Genova)

Extra Reviewers

Philip Garner (Idiap Research Institute), Marcello Mortillaro (University of Geneva),
Andrei Popescu-Belis (Idiap Research Institute), Hiroshi Shimodaira (University of Edinburgh).

Sponsors

humaine [1]

Social Signal Processing Network [2]

[1] http://emotion-research.net/

[2] http://wcms.inf.ed.ac.uk/sspnet/

Contents

Doctoral Consortium Papers

Dialogue Act Recognition and the Role of Affect 1
Nicole Novielli, Carlo Strapparava

The Expression of Joy and Frustration in English Conversation 9
Changrong Yu, Jiehan Zhou

Affective Support in Narrative-Centered Learning Environments 17
Jennifer Robison

Social Network Analysis in Multimedia Indexing: Making Sense of People in Multiparty Recordings 25
Sarah Favre

Emotive and Personality Parameters in Multimedia Recommender Systems 33
Marko Tkalčič, Jurij Tasič and Andrej Košir

A Unified Features Approach to Human Face Image Analysis and Interpretation 41
Zahid Riaz, Suat Gedikli, Micheal Beetz, Bernd Radig

Gaining Rapport by Voicing Appropriate Emotional Responses Based on User State 49
Jaime C. Acosta

Synthesis of Nonverbal Listener Vocalizations 57
Sathish Pammi

Toward Natural Human-Robot Interaction: Exploring Facial Expression Synthesis on an Android Robot 65
Laurel D. Riek

Endowing Artificial Agents with Emotional Autobiographical Memories 73
Davi D'Andréa Baccan, Luís Macedo

Non-verbal Behaviour and Attribution of Mental States 81
Sylwia Hyniewska, Susanne Kaiser, Catherine Pelachaud

Neurophysiological Assessment of Affective Experience 89
Christian Mühl

Affect driven Creativity Support Tools 97
Priyamvada (Pia) Tripathi

List of authors 105

Dialogue Act Recognition and the Role of Affect *

Nicole Novielli
Dipartimento di Informatica, University of Bari
via Orabona, 4 - 70125 Bari, Italy
novielli@di.uniba.it

Carlo Strapparava
FBK-irst, Istituto per la Ricerca Scientifica e Tecnologica
via Sommarive, 18 - I-38050 Povo Trento, Italy
strappa@fbk.eu

Abstract

We study the task of automatic labeling dialogues with the proper speech acts, relying on empirical methods and simply exploiting lexical semantics of the utterances. We investigate the relationship between affective factors and linguistic realization of dialogue acts: we present some preliminary results about the role that affect plays in dialogue act disambiguation and we propose a discussion about open problems.

1 Introduction

In natural conversations people can ask for information, express their opinions, state some facts, agree or disagree with their partner through sequences of Dialogue Acts (Core and Allen 1997). Regardless of the used language or the domain in which the discussion takes part, communicative goals are the main factor influencing the linguistic realization of such a series of acts. Though, the speaker affective states and attitudes have been proved to significantly affect her communicative behavior and language (see e.g. Bosma and André 2004; de Rosis et al. 2007)

In this perspective, opinions deserve a specific discussion. Humans, in fact, may express their opinions in several ways: they may patently shiver, close the windows or say 'Cold today, isn't it?', to manifest their opinion that the temperature is not adequate. Considerable efforts are being made towards inferring goals from observation of nonverbal behavior (see, e.g., Gray and Breazeal 2005). The process of inferring the communicative goal of our interlocutor is particularly complex and can be schematically represented as follows: (i) identification of the meaning of the words used; (ii) identification of the proposition expressed in light of the meaning and the rest of the situation in which the utterance takes place and (iii) identification of further implicatures over and above the proposition expressed (Gauker 1994).

The long-term goal of our study is to exploit the relationship between the communicative intention of a Dialogue Act (DA) and its linguistic realization. In particular, we aim at defining linguistic profiles for DAs through a similarity study in latent semantic spaces automatically acquired from dialogue corpora. To ensure the independence of our DA profiles from the language used, the application domain and other important features such as the interaction mode, we focus our experiments on two different corpora of natural dialogues.

Even if prosody and syntactic features surely play a role in the linguistic realization of dialogue acts, (Jurafsky et al. 1998; Stolcke et al. 2000; Warnke et al. 1997), in our study we aim at simply exploiting lexical semantics of utterances. With the advent of the Web, a large amount of material about natural language interaction (e.g. blogs, chats, conversation transcripts) has become available, raising the attractiveness of empirical methods of analysis on this field. And

[*] This paper is dedicated to Fiorella de Rosis: this research would have not begun without her encouragement.

still, language will be one of the most common communication media with smart environments and what the speakers use to convey their messages. Moreover, words are just what we have at disposal when we consider texts found on the Web.

DA profiles can be useful for both generation and recognition purposes. There is a large number of applicative scenarios that could benefit from automatic dialogue act processing and deep understanding of the conversational structure: e.g. conversational agents for monitoring and supporting human-human remote conversations, blogs, forum and chat log analysis for opinion mining, automatic meeting summarization and so on. In particular, one of the long-term goals of our research is to exploit conversational analysis techniques for interpersonal stances modeling by mean of analysis of the dialogue pattern (Martalo et al. 2008).

We propose an experimental study about automatic labeling of natural dialogues with the proper speech acts. In particular, the research described in this paper represents a preliminary step towards the definition of an unsupervised approach. Evaluation displays encouraging results, and supports our assumption of independence of DA profiles from the language and the application domain in which the conversations take part.

However, from a first error analysis, the discrimination among communicative acts such as statements and opinions is difficult to be resolved relying on a simple DA profiling. This suggests to go further towards incorporating affect information in the process. In this paper we explore the affective load of sentences for dialogue acts disambiguation especially for opinion recognition.

2 Dialogue Corpora

In this paper we exploit two corpora, both annotated with DA labels. According to our goal of developing a recognition methodology as much general as possible, we selected two corpora which are different in the content and in the used language: the Switchboard (Godfrey et al. 1992), a collection of transcriptions of spoken English telephone conversations about general interest topics, and an Italian corpus of dialogues in the healthy-eating domain (Clarizio et al. 2006).

Speaker	**Dialogue Act**	**Utterance**
A	OPENING	*Hello Ann.*
B	OPENING	Hello Chuck.
A	STATEMENT	*Uh, the other day, I attended a conference here at Utah State University on recycling*
A	STATEMENT	*and, uh, I was kind of interested to hear cause they had some people from the EPA and lots of different places, and, uh, there is going to be a real problem on solid waste.*
B	OPINION	Uh, I didn't think that was a new revelation.
A	AGREE /ACCEPT	*Well, it's not too new.*
B	INFO-REQUEST	So what is the EPA recommending now?

Table 1: An excerpt from the Switchboard corpus

The Switchboard corpus is a collection of English human-human telephone conversations (Godfrey et al. 1992), involving couples of randomly selected strangers: they were asked to select a general interest topic and to talk informally about it. Full transcripts are distributed by the Linguistic Data Consortium. A part of this corpus is annotated (Jurafsky et al. 1997) with DA labels (overall 1155 conversations, for a total of 205,000 utterances and 1.4 million words)[1].

The Italian corpus had been collected in the scope of some previous research about Human-ECA (Embodied Conversational Agent) interaction: a Wizard of Oz tool was employed (Clarizio et al. 2006) in which the application domain and the ECA's appearance may be settled at the beginning of the simulation. The ECA played the role of an artificial therapist and the users were free to interact with it in natural language, without any particular constraint. This corpus is about healthy eating and contains overall 60 dialogues, 1448 users' utterances and 15,500 words.

[1] ftp.ldc.upenn.edu/pub/ldc/public_data/swb1_dialogact_annot.tar.gz

Label	Description	Example	Ita	En
INFO-REQUEST	Utterances that are pragmatically, semantically, and syntactically questions	*'What did you do when kids were growing up?'*	34%	7%
STATEMENT	Descriptive, narrative, personal statements	*'I usually eat a lot of fruit'*	37%	57%
S-OPINION	Directed opinion statements	*'I think he deserves it.'*	6%	20%
AGREE-ACCEPT	Acceptance of a proposal, plan or opinion	*'That's right'*	5%	9%
REJECT	Disagreement with a proposal, plan, or opinion	*'I'm sorry no'*	7%	.3%
OPENING	Dialogue opening or self-introduction	*'Hello, my name is Imma'*	2%	.2%
CLOSING	Dialogue closing (e.g. farewell and wishes)	*'It's been nice talking to you.'*	2%	2%
KIND-ATT	Kind attitude (e.g. thanking and apology)	*'Thank you very much.'*	9%	.1%
GEN-ANS	Generic answers to an Info-Request	*'Yes', 'No', 'I don't know'*	4%	4%
total cases			1448	131,265

Table 2: The set of labels employed for DA annotation and their distribution in the two corpora

Labelling. Dialogue Acts (DA) are well studied in linguistic (Austin 1962; Searle 1969) and computational linguistics (Core and Allen 1997; Traum 2000) since long time. A DA can be identified with the communicative goal of a given utterance (Austin 1962). A plethora of labels and definitions have been used to address this concept: *speech act* (Searle 1969), *adjacency pair part* (Schegloff 1968), *game move* (Power 1979); Cohen and Levesque (1995) focus more on the role speech acts play in interagent communication. Traditionally, the NLP community has employed DA definitions with the drawback of being domain or application oriented. In the recent years, some efforts have been made towards unifying the DA annotation (Traum 2000).

In this study we refer to a domain-independent framework for DA annotation, the DAMSL architecture (Dialogue Act Markup in Several Layers) by Core and Allen (1997). In particular the Switchboard corpus employs a revision (Jurafsky et al. 1997). Table 2 shows the set of labels employed with their definitions and examples: it maintains the DAMSL main characteristic of being a domain-independent framework and it is also consistent with the annotation rationale applied in the labelling of the Switchboard corpus with SWBD-DAMSL. Thus, the original SWBD-DAMSL annotation had been automatically converted into the categories included in our markup language as described in (Novielli and Strapparava 2009). Also we did not consider the utterances formed only by non-verbal material (e.g. laughter).

3 Exploiting the Lexical Semantics of DAs

Recently, the problem of DA recognition has been addressed with promising results: Poesio and Mikheev (1998) combine expectations about the next likely dialogue 'move' with information derived from the speech signal features; Stolcke et al. (2000) employ a discourse grammar, formalized in terms of Hidden Markov Models, combining also evidences about lexicon and prosody; Keizer et al. (2002) make use of Bayesian networks for DA recognition in dutch dialogues; Grau et al. (2004) consider naive Bayes classifiers as a suitable approach to the DA classification problem.

Regardless of the model they use (discourse grammars, models based on word sequences or on the acoustic features or a combination of all these) the mentioned studies are developed in a supervised framework. Unfortunately, it is not always easy to have large training material at disposal, partly because of manual labeling effort and moreover because often it is not possible to find it. For this reason we decided to explore the possibility of using a fully unsupervised methodology. This paper is a preliminary contribution in this direction.

3.1 The Unsupervised Approach

Schematically, our unsupervised methodology is: (i) building a semantic similarity space in which words, set of words and text fragments can be represented homogeneously, (ii) finding seeds that properly represent dialogue acts and considering their representations in the similarity space, and (iii) checking the similarity of the utterances.

To reduce the data sparseness, we use a POS-tagger and morphological analyzer (Pianta et al. 2008) for preprocessing the corpora and we use lemmata instead of tokens in the format *lemma#POS*. No feature selection is performed, keeping also stopwords. In addition, we augment the features of each sentence with a set of linguistic markers, defined according to the semantic of the DA categories (Novielli and Strapparava 2009).

To get a similarity space with the required characteristics, we use Latent Semantic Analysis (LSA), a corpus-based measure of semantic similarity (Landauer et al. 1998). In LSA, term co-occurrences in a corpus are captured by means of a dimensionality reduction operated by a singular value decomposition (SVD) on the term-by-document matrix **T** representing the corpus.

LSA can be viewed as a way to overcome some of the drawbacks of the standard vector space model (sparseness and high dimensionality). In fact, the LSA similarity is computed in a lower dimensional space, in which second-order relations among terms and texts are exploited. The similarity is then measured with the standard cosine similarity. Note also that LSA yields a vector space model that allows for a *homogeneous* representation (and hence comparison) of words, sentences and texts. For representing a word set or a sentence in the LSA space we use the *pseudo-document* representation technique, as described by Berry (1992). In practice, each text segment is represented in the LSA space by summing up the normalized LSA vectors of all the constituent words, using also a *tf.idf* weighting scheme (Gliozzo and Strapparava 2005).

Label	**Seeds**
INFO-REQ	Question_mark
S-OPINION	Verbs which directly express opinion or evaluation (guess, think, suppose, affect)
AGREE-ACC	yep, yeah, absolutely, correct
OPENING	Expressions of greetings (hi, hello), words and markers related to self-introduction formula
KIND-ATT	Lexicon which directly expresses wishes (wish), apologies (apologize), thanking (thank) and sorry-for (sorry, excuse)

Table 3: Some example of set of seeds

	Italian						**English**					
	SVM			**LSA**			**SVM**			**LSA**		
Label	prec	rec	f1	prec	rec	f1	prec	rec	f1	prec	rec	f1
INFO-REQ	.92	.99	.95	.96	.88	.92	.92	.84	.88	.93	.70	.80
STATEMENT	.85	.68	.69	.76	.66	.71	.79	.92	.85	.70	.95	.81
S-OPINION	.28	.42	.33	.24	.42	.30	.66	.44	.53	.41	.07	.12
AGREE-ACC	.50	.80	.62	.56	.50	.53	.69	.74	.71	.68	.63	.65
REJECT	-	-	-	.09	.25	.13	-	-	-	.01	.01	.01
OPENING	.60	1.00	.75	.55	1.00	.71	.96	.55	.70	.20	.43	.27
CLOSING	.67	.40	.50	.25	.40	.31	.83	.59	.69	.76	.34	.47
KIND-ATT	.82	.53	.64	.43	.18	.25	.85	.34	.49	.09	.47	.15
GEN-ANS	.20	.63	.30	.27	.38	.32	.56	.25	.35	.54	.33	.41
micro	.71	.71	.71	.66	.66	.66	.77	.77	.77	.68	.68	.68

Table 4: Evaluation of the supervised and unsupervised methods on English and Italian corpora

The methodology is unsupervised[2] as we do not exploit any training material. The seeds are general and language-independent since they are defined by considering only the communicative goal and the specific semantic of each dialogue act, just avoiding as much as possible the overlapping between seed groups. Since our aim is to design an approach which is as general as possible, we do not consider domain words that could make easier the classification in the specific corpora. Table 3 shows some example of set of seeds with the corresponding DAs. The seeds are the same for both languages, which is coherent with our goal of defining a language-independent method. We run the SVD using 400 dimensions respectively on the English and Italian unlabeled corpus. Starting from a set of seeds (words) representing the DAs, we build the corresponding vectors in the LSA space and then we compare the utterances to find the communicative act with the highest similarity.

[2] Or minimally supervised, since providing hand-specified seeds can be regarded as a minimal sort of supervision.

We compare the performance of our approach with a 'ceiling' results represented by the performance of Support Vector Machines (Vapnik 1995)[3]. We randomly split the two corpora into 80/20 training/test partitions. SVMs have been used in a large range of problems, including text classification, image recognition and medical applications, and they are regarded as the state-of-the-art in supervised learning. We got .71 and .77 of F1 respectively for the Italian and English corpus. Table 4 shows the performance for each DA. To allow comparison, the performance is measured on the same test set partitions for both experiments. Since we are evaluating an unsupervised approach, we consider random DA selection (11%) as baseline.

3.2 Error Analysis

After conducting an error analysis, we noted that many utterances are misclassified as STATEMENT. One possible reason is that statements usually are quite long and there is a high chance that some linguistic markers that characterize other dialogue acts are present in statements.On the other hand, looking at the corpora we observed that many utterances which appear to be linguistically consistent with the typical structure of statements have been annotated differently, according to the actual communicative role they play. In the following example, a statement-like utterance (by speaker B) is annotated differently because of its context (speaker A's move):

A: *'In fact, it's easier for me to say, uh, the types of music that I don't like are opera and, uh, screaming heavy metal.'* STATEMENT

B: *'The opera, yeah, it's right on track.'* AGREE-ACCEPT

For similar reasons, we observed some misclassification of S-OPINION as STATEMENT, which is the main cause of decrease of the performance of our method. In fact, most part of the S-OPINION utterances in our corpora (92% of the English data set and 25% of the Italian one) are misclassified as statements (the better performance in the opinion recognition for the Italian corpus is probably due to the restricted domain and, hence, lexicon, of this second data set). The only significative difference between the two labels seems to be the wider usage of 'slanted' and affectively loaded lexicon when conveying an opinion.

Another source of confounding is the misclassification of the OPENING as INFO-REQUEST. The reason is not clear yet, since the misclassified openings are not question-like. Eventually, there is some confusion among the back-channel labels (GEN-ANS, AGREE-ACC and REJECT) due to the inherent ambiguity of common words like *yes*, *no*, *yeah*, *ok*.

Recognition of such cases could be improved (i) by enabling the classifiers to consider not only the lexical semantics of the given utterance (local context) but also the knowledge about a wider context window (e.g. the previous n utterances), (ii) by enriching the data preprocessing (e.g. by exploiting information about lexicon polarity and subjectivity parameters). These are both directions we intend to follow in our future research.

4 Exploiting Affective Load for Dialogue Act Disambiguation

Sensing emotions from text is a particularly appealing task of natural language processing (Strapparava and Mihalcea 2007; Pang and Lee 2008): the automatic recognition of affective states is becoming a fundamental issue in several domains such as human-computer interaction or sentiment analysis for opinion mining. Recently there have been several attempts to integrate emotional intelligence into user interfaces (Conati 2002; Picard and Klein 2001; Clarizio et al. 2006). A first attempt to exploit affective information in dialogue act disambiguation has been made by Bosma and André (2004), with promising results. In their study, the recognition of emotions is based on sensory inputs which evaluate physiological user input.

In this Section we present the results of a qualitative study aimed at investigating the relationship between the affective load of a given utterance and its communicative goal (i.e. its DA label). To the best of our knowledge, this is the first attempt to study the relationship between the communicative act of an utterance and its affective load by applying lexical similarity techniques to textual input.

[3] We used SVM-light package (Joachims 1998) under its standard configuration

4.1 Method

We calculate the affective load of each DA label using the methodology described in (Strapparava and Mihalcea 2008). The idea underlying the method is the distinction between *direct* and *indirect* affective words. For direct affective words, authors refer to the WordNet Affect (Strapparava and Valitutti 2004) lexicon, an extension of the WordNet database (Fellbaum 1998) which employs six basic emotion labels (anger, digust, fear, joy, sadness, surprise) to annotate WordNet synsets. LSA is then used to learn, in an unsupervised setting, a vector space from the British National Corpus[4]. As said before, LSA has the advantage of allowing homogeneous representation and comparison of words, text fragments or entire documents, using the pseudo-document technique exploited in Section 3.1. In the LSA space, each emotion label can be represented in various way. In particular, we employ the 'LSA Emotion Synset' setting, in which the synsets of direct emotion words are considered. The affective load of a given utterance is calculated in terms its lexical similarity with respect to one of the six emotion labels. The overall affective load of a sentence is then calculated as the average of its similarity with each emotion label.

4.2 Results

Results are shown in Table 5 (a) and confirm our preliminary hypothesis about the use of slanted lexicon in opinions. In fact, S-OPINION is the DA category with the highest affective load. Opinions are immediately followed by KIND-ATT due to the high frequency of politeness formulas in these utterances (see Table 5 (b) for example utterances).

Label	Affective Load
S-OPINION	.1439
KIND-ATT	.1411
STATEMENT	.1300
INFO-REQ	.1142
CLOSING	.0671
REJECT	.0644
OPENING	.0439
AGREE-ACC	.0408
GEN-ANS	.0331

(a)

S-OPINION
You know, but, gosh uh, it's getting pathetic now, absolutely pathetic. They're just horrid, you'll have nightmares, you know. That's no way to make a decision on some terrible problem. They are just gems of shows. I mean, really, fabulous in every way . And, oh, that is so good. Delicious. They have some delicious, delicious things
KIND-ATTITUDE
I'm sorry, I really feel strongly about this. Sorry, now I'm probably going to upset you. I hate to do it on this call.

(b)

Table 5: Affective load of DA labels (a) and examples of slanted lexicon (b)

5 Discussion and Future Work

This contribution is a preliminary step towards our long-term goal of defining an unsupervised methodology for automatically annotating interactions with the proper speech acts, by simply exploiting the lexical semantics of individual dialogue turns. The methodology has to be independent from some important features of the corpus being analyzed, such as the language and the application domain. Moreover, it will embed some form of emotional intelligence in order to better disambiguate dialogue acts.

S-OPINION	**STATEMENT**
adjectives	*nouns*
obstinate (.67) overloaded(.65) pathetic(.53) satisfying(.50) dirty(.50) ridiculous(.47)	jumbo(.48) gomphrena(.48) rhapsody(.48) milliliter(.48)
verbs	*adjectives*
disqualify(.63) hurt(.40)	nonstop(.48) outboard(.48) bohemian(.48)

Table 6: The lexical similarity for S-OPINION and STATEMENT

[4] http://www.hcu.ox.ac.uk/bnc/

In this work, we have studied how lexical semantics of dialogue turns can be exploited to automatically annotate dialogues with the proper speech acts, using an unsupervised approach. The methodology consists of defining a very simple and intuitive set of seeds that profiles the specific dialogue acts, and subsequently performing a similarity analysis in a latent semantic space. The performance of the unsupervised experiment has been compared with a supervised state-of-art technique such as Support Vector Machines. The results are quite encouraging and highlight the role played by lexical semantic in profiling the communicative goal of a dialogue turn.

On the other side, the method shows a lack of performance in disambiguating between objective and subjective statements (opinions). In fact, opinions are conveyed through a statement-like structure and the main difference between the two labels seems to be the wider use of slanted lexicon in expressing attitudes and preferences. To verify whether DA profiles could be improved with additional features, we conducted a similarity study on the whole annotated corpus. Results show that S-OPINIONs are more similar to slanted adjectives with a non-neutral a priori polarity while STATEMENT are shown to be similar to nouns or adverbs which do not directly refer to attitudes or evaluations (see Table 6). In addition, we performed a qualitative study about the affective load of utterances, exploiting a state-of-the-art technique in checking the affective content sentences. The experimental results show that a relationship exists between the affective load and the communicative goals of utterances.

Regarding future developments, we will investigate how to include in the framework a wider context (e.g. the previous n utterances), as well as new linguistic markers (i.e. enriching the preprocessing techniques). In particular, it would be interesting to exploit the role of slanted or affective-loaded lexicon to deal with the misclassification of opinions as statements. Along this perspective, DA recognition could serve also as a basis for conversational analysis aimed at improving a fine-grained opinion mining in dialogues.

To conclude, there is a huge number of applications which could benefit from DA annotation of dialogues in both human-human and human-computer interaction scenarios (e.g. meeting summarization, communication in multi-agent systems, analysis of chat or blog transcripts and so on). In our previous research, we focused on how to exploit conversational analysis techniques for long-term attitude modeling. In particular, we investigated, with promising results, how affective factors influence dialogue patterns and whether this effect may be described and recognized by HMMs (Martalo et al. 2008). The long-term goal is to analyze the possibility of using this formalism to classify the user behavior for adaptation purposes.

References

Austin, J. (1962). *How to do Things with Words*. Oxford University Press, New York.

Berry, M. (1992). Large-scale sparse singular value computations. *International Journal of Supercomputer Applications*, 6(1).

Bosma, W. and André, E. (2004). Exploiting emotions to disambiguate dialogue acts. In *IUI '04: Proceedings of the 9th international conference on Intelligent user interfaces*, pages 85–92, New York, NY, USA. ACM.

Clarizio, G., Mazzotta, I., Novielli, N., and deRosis, F. (2006). Social attitude towards a conversational character. In *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication*, pages 2–7, Hatfield, UK.

Cohen, P. R. and Levesque, H. J. (1995). Communicative actions for artificial agents. In *in Proceedings of the First International Conference on Multi-Agent Systems*, pages 65–72. AAAI Press.

Conati, C. (2002). Probabilistic assessment of user's emotions in educational games. *Applied Artificial Intelligence*, 16:555–575.

Core, M. and Allen, J. (1997). Coding dialogs with the DAMSL annotation scheme. In *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Cambridge, MA.

de Rosis, F., Batliner, A., Novielli, N., and Steidl, S. (2007). 'You are Sooo Cool, Valentina!' Recognizing Social Attitude in Speech-Based Dialogues with an ECA. In Paiva, A., Prada, R., and Picard, R. W., editors, *Affective Computing and Intelligent Interaction*, LNCS, pages 179–190, Berlin-Heidelberg.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Gauker, C. (1994). *Thinking Out Loud: An Essay on the Relation between Thought and Language*. Princeton University Press.

Gliozzo, A. and Strapparava, C. (2005). Domains kernels for text categorization. In *Proc. of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 56–63, University of Michigan, Ann Arbor.

Godfrey, J., Holliman, E., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 517–520, San Francisco, CA. IEEE.

Grau, S., Sanchis, E., Castro, M. J., and Vilar, D. (2004). Dialogue act classification using a bayesian approach. In *Proceedings of the 9th International Conference Speech and Computer (SPECOM-2004)*, pages 495–499, Saint-Petersburg, Russia.

Gray, J. and Breazeal, C. (2005). Toward helpful robot teammates: a simulation-theoretic approach for inferring mental sate of others. In *Proceedings of the AAAI Workshop on Modular Construction of Human-Like Intelligence*, Pittsburgh.

Joachims, T. (1998). Text categorization with Support Vector Machines: learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*.

Jurafsky, D., Shriberg, E., and Biasca, D. (1997). Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-01, University of Colorado Institute of Cognitive Science.

Jurafsky, D., Shriberg, E., Fox, B., and Curl, T. (1998). Lexical, prosodic, and syntactic cues for dialog acts. In *Proceedings of ACL/COLING 98*, pages 114–120, Montreal.

Keizer, S., op den Akker, R., and Nijholt, A. (2002). Dialogue act recognition with bayesian networks for dutch dialogues. In Jokinen, K. and McRoy, S., editors, *Proceedings 3rd SIGdial Workshop on Discourse and Dialogue*, pages 88–94, Philadelphia, PA.

Landauer, T., Foltz, P., and Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25.

Martalo, A., Novielli, N., and de Rosis, F. (2008). Attitude display in dialogue patterns. In *AISB 2008 Convention on Communication, Interaction and Social Intelligence*, Aberdeen, Scotland.

Novielli, N. and Strapparava, C. (2009). Towards unsupervised recognition of dialogue acts. In *NAACL HLT 2009, Student Research Workshop*.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pianta, E., Girardi, C., and Zanoli, R. (2008). The TextPro tool suite. In *Proceedings of LREC-08*, Marrakech, Morocco.

Picard, R. W. and Klein, J. (2001). Computers that recognise and respond to user emotion: Theoretical and practical implications. Technical report, MIT Media Lab.

Poesio, M. and Mikheev, A. (1998). The predictive power of game structure in dialogue act recognition: Experimental results using maximum entropy estimation. In *Proceedings of ICSLP-98*, Sydney.

Power, R. (1979). The organisation of purposeful dialogues. *Linguistics*, 17:107–152.

Schegloff, E. (1968). Sequencing in conversational openings. *American Anthropologist*, 70:1075–1095.

Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, London.

Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Ess-Dykema, C. V., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Strapparava, C. and Mihalcea, R. (2007). SemEval-2007 task 14: Affective Text. In *Proceedings of the 4^{th} International Workshop on Semantic Evaluations (SemEval 2007)*, pages 70–74, Prague.

Strapparava, C. and Mihalcea, R. (2008). Learning to identify emotions in text. In *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560, New York, NY, USA. ACM.

Strapparava, C. and Valitutti, A. (2004). WordNet-Affect: an affective extension of WordNet. In *Proceedings of LREC*, volume 4, pages 1083–1086.

Traum, D. (2000). 20 questions for dialogue act taxonomies. *Journal of Semantics*, 17(1):7–30.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.

Warnke, V., Kompe, R., Niemann, H., and Nöth, E. (1997). Integrated dialog act segmentation and classification using prosodic features and language models. In *Proceedings of 5th European Conference on Speech Communication and Technology*, volume 1, pages 207–210, Rhodes, Greece.

The Expression of Joy and Frustration in English Conversation

Changrong Yu
University of Oulu, English Philology
changrong.yu@oulu.fi

Jiehan Zhou
University of Oulu, Department of Electrical and Information Engineering
jiehan.zhou@ee.oulu.fi

Abstract

This paper studies the expression of emotion in two different scenarios, narrative story and argument, in naturally occurring English conversation. The different ways of expressing joy and frustration are examined through linguistic features, sequential positioning, prosody (pitch, loudness, and speed) and embodied actions. It is suggested that joyful display can have a positive effect on the whole conversation. It explores if the two different types of frustration, active and passive frustration, are associated with special types of linguistic, paralinguistic features and embodied actions.

Keywords: Joy, frustration, emotion expression, linguistic, paralinguistic, embodied actions

1 INTRODUCTION

Human society relies heavily on the free and easy communication among its members by conversation. Some conversations bring us joy. We align with the speaker and are happy with his/her story. However, some conversations make us frustrated because we cannot persuade the recipients to accept our opinions or attitudes. This paper studies the expression of emotion in two different emotional scenarios, narrative story and arguments, from the same speakers.

In spontaneous English conversation, we express our emotions through two cues, verbal and non-verbal. For example, what I look like when I sound angry is one of the non-verbal cues. In conversation, people use verbal cues together with all kinds of embodied actions such as facial expressions, gestures, body movements, actions and physiological cues to express emotional states, attitudes, and intentions, and to communicate interpersonal relations, to influence the perception of other interlocutors and to obtain some goals as well as to influence the behavior of others [1]. This paper focuses on the linguistic features (e.g., syntax, lexical information and semantics) associated with the acoustic features (e.g., pitch, tempo, hesitations, speaking rate). In addition, this paper studies how the non-verbal cues and verbal cues interact with each other in expressing emotions.

Why do we communicate joy and frustration and how do we communicate them? For answering this question, we have studied naturally occurring conversation primarily from the perspective of interactional linguistics and conversation analysis.

The paper aims at finding out how people jointly achieve emotion comprehension and how they fail emotion perception in conversation by analysis of the linguistic features and the accompanied embodied actions in English conversation. The remainder of the paper is organized as follows. Section 2 presents data and research objectives. Section 3 briefly overviews related work on emotion approaches. Section 4 studies expression of joy. Section 5 studies expression of frustration. Conclusion is drawn in Section 6.

2 DATA AND OBJECTIVES

The data comes from the videotape and transcript called ‘Never in Canada’ collected in the Department of English, University of Oulu, in 2003. The three speakers are around 23 years old when the data was collected. They are all exchange students at Oulu University. Jason and Mary are from United States, Sophia is from Canada. The emotional expression of joy happens in a narrative story. The joyful data is a narrative story of a personal experience, named by “no offense, we just don't do that, in Canada”. Another data is an argument from the same speakers, Jason and Mary, which are from the same video data as the joyful one. The corpus data was transcribed using the conventions in Du Bois et al. [23] (. The data is transcribed into intonation units, or stretches of speech uttered under a single intonation contour, such that each line represents one intonation unit [24].

This paper investigates the various forms of emotion expression in these two conversations. Speakers do not only use linguistic channels to express their emotions, but also do so through paralinguistic channels and embodied actions, like facial expressions, gestures and postures. As a result, we first study how speakers display their emotions through linguistic features, sequential positioning, prosody (pitch, loudness, and speed) and embodied actions. Second, we explore how verbal expression for emotion is unavoidably accompanied by embodied actions, and examine how vocal and kinesic expressions are causes and effects of emotional display. Third, we find out the emotional sequence of the recipient at the other speaker's turn in arguments. Finally we seek if joy and frustration are associated with special types of paralinguistic features or embodied actions.

3 RESEARCH BACKGROUND OF EMOTION IN SOCIAL INTERACTION

The expression of emotion in English conversation and discourse has not been systematically explored. However several approaches are now studying emotion in discourse and talk-in- interaction.

Within social psychology, some researchers have specified the prototypes of the five basic emotions, i.e., love, joy, surprise, anger, sadness, and fear [2, 3][4]. The approach studies emotion in discourse and social life. These mentioned researchers elicited the prototype of emotions from written experience. Subjects were given questionnaire to write their emotion episodes in which they experienced the five basic emotions in real life [5]. This typical study method is often used by social psychologists [6][7]. The subjects need to write in detail the cause of their emotion, and in their feeling and thinking in as much detail as they can, languages they used, physical actions and so on. In line with their study, we try to elicit the prototype of emotions if there is. However, our study stands on the foot of naturally occurring data. The approach of social psychology has extended the network of emotion being influenced by social, moral, cultural and psychological factors. Gottman [8] proposed affective reciprocity in emotional interaction, by studying the audio-recorded data from marital conversation of dissatisfied couples. Discursive psychology (DP) has been profoundly influenced by conversation analysis (CA) which offered the approach for dealing with interactional materials. DP has started to study evaluative expressions in naturally occurring interaction as part of varied social practices, considering what such expressions are doing, rather than their relationship to attitudinal objects or other putative mental entities; they study how evaluation is situated sequentially and rhetorically (Wiggins & Potter 2003). This approach goes beyond the function of emotion signs.In sociolinguistics, Chafe [1], emphasizes that emotion is present in everyday conversation. Emotion is what gives communication life. Coordination between partners in conversation occurs at many levels, and they are all grounds for emotion. Emotion is thus identified as intersubjective.

The method used in our study is drawn from linguistic-interactional approaches to emotions and conversation analysis (CA). It is necessary to give a brief introduction on CA. CA is an established approach to studying human interactions, and is applied disciplines such as sociology, linguistics, anthropology, communications, and social psychology. The main method is the close study of recordings, either audio or virtual transcripts of naturally occurring conversational interaction.

Emotions are not addressed in the field of CA. In CA, the term of recipients are used for the substitution of traditional speaker and listener. Sacks's insights provide a good starting point for incorporating grammar into a theory of social action and also for an analysis of social interaction (e.g. [10-12][13, 14].

The study of the connection between dialogic nature of language and the grammatical features, such as the epistemic stance (see [15]) is shown by interactional linguists. This approach proposes that emotion should be studied in discourse and emotion is interpersonal and intersubjectively achieved in conversation. Sandlund [16] studied the dates of academic talk-in-interaction in terms of sequential environment, their interactional elicitors, their management and closing, by using the conversation analytic approach. She studied basically three themes: frustration, embarrassment and enjoyment, and within each, assortments of practices for doing emotions were found. Frustration was primarily located in the context of violations of activity-specific turn-taking norms. Enjoyment was found to be collaboratively pursued between and within institutional activities. The findings indicate that emotion displays can be viewed as transforming a situated action and opening up alternative.

4 EXPRESSION OF JOY

In this data, Jason is the story-teller. Jason's narration starts at 5.04 minutes of their whole conversation, and his narration goes across more than 120 intonation units in this 2.21 minutes episode. We are unable to provide the whole transcripts due to the limitations of the length. In this episode, Jason tells his friends in which situation and how he told people that he was Canadian. While he and some other exchange students were waiting for the taxi in a long queue at four o'clock in the morning at negative twenty-degrees, he shouted out: <*VOX this is the dumbest, fucking thing, I have ever seen, in my entire life VOX*>, and then <*VOX no offense, (0.7) we just don't do that, in Canada VOX*>. Afterwards, they walked up the roads and hailed a taxi instead of waiting for their turn in the gigantic queue. Jason's action in the story gains excited laughter and compliments from the other two recipients. We study how the speakers (narrator and recipients) coordinate coherently and joyfully in the process.

Of the two recipients- Mary and Sophia, Sophia has heard the story once, but she still encourages Jason to re-tell the story to Mary. Since the story is still new for Mary, she makes more evaluation than Sophia. Sophia always affiliates with Mary in the conversation. We use Anvil [29] to track these three speakers' emotions, and the starting time and the end time of the emotion expression were recorded. After we annotated each speaker's emotion, then we save these annotations into a table for comparing the mutual emotional interaction of the recipients. The analysis suggests that the positive emotion and the compliments of the recipients help push Jason's narration to the climax. Our finding is similar to the theory of emotion coincidence, emotion contagion and empathy in terms of cognitive and sociology. The expression of an emotional state in one person often leads to the experience or expression of a similar emotion in another person [1][17].

From the generated annotation track, we obtain emotion inter-correlations between Jason's narration and Mary's emotion, as seen in Table 1. We find out that the scale of Mary's emotion deepens and becomes more and more positive with the progress of Jason's story. Their emotion interacts and mirrors with each other very well. In the prelude of the story, Mary is reactive (surprise), then in the stage of preface and development, she becomes quite positive in line 446, *[It sounds like] a joke*, then her laughter in sequences 450, 502, 510 and 512, is a sign of acceptance and offers the story-teller, Jason, a relaxing and encouraging atmosphere. Following the laughter, she gives her evaluation in sequence 519, *(.) @ °Nice°*, and in the climax of Jason's narrative, her emotion becomes lively positive, which is shown by her compliments and empathy, starting in sequence 523, *That's funny.(.) Yeah, I guess I wouldn't--(.) stand in a line like that. [I would] [^{3}go to a different^3]---*. The above analysis shows that the emotion flow of Jason's narration well mirrors Mary's emotion flow. Jason's narration cannot achieve such a positive effect without the collaboration of Mary. And the recipients show different degrees of affiliation with each other during the interaction.

The result is meanwhile similar to the cognitive finding, with positive emotions resulting from goal congruence and producing more creative and variable actions [18]. The annotation tool can help us understand how people jointly achieve emotion comprehension, why assessments or evaluations occur, where they occur, how they occur, and what is evaluated in conversation.

Table 1. Emotion inter-correlation between Jason's narration and the two recipients

Joyful cues of Jason	Joyful cues of Mary and Sophia
Prelude: *Oh you,didn't hear about that story?* Surprise with proud	Reactive: *You gotta, tell the story. "You told people [you were Canadian]?"* surprise, curiosity, eager encouragement and curiosity from two recipients
Preface and development: *Oh, this is great. (.) The greatest Saturday night,...",* Jason's willingness for sharing his story with recipients ; Linguistic features of Jason's expression: vivid lexical choices, rising tone and lengthened prosody	positive reaction: tease, acceptance, spontaneous laughter
Climax: dramatized reiteration	Lively positive emotion : *(.) @ °Nice°.That's funny,* compliment with exciting laughter
Denouement: self-evaluation of the whole story	Positive verbal evaluation

Jason’s recounting ends in a joyful atmosphere. The narration together with its humorous exploitation evokes lots of laughter. Jason masters the sensitive topic of declaring himself as Canadian and strikes a confident balance between humor and seriousness.

Mary evaluates Jason’s story actively by her non-verbal cues (laughter, gesture, facial expression), verbal cues (rising intonation). Her verbal compliments are expressed in the phrase and sentence level.

Here is the development of Mary’s emotion: developing from surprise, curiosity to tease (disbelief), acceptance, excitement (contentment), complimenting evaluation and finally declaring to do the same. Let us analyze Mary’s emotion flow extracted from the interaction, example 1- excerpt of Mary’s conversation.

```
399  MAR:You told people [you were Canadian]?
413  MAR:[In Fin]land]?
442  MAR:[Spaniard],
446  MAR:[It sounds like] a joke.
450  MAR:[@@@@@]
472  MAR:So,
473      everybody just takes their turn?
477  MAR:=So it's not competitive?
478      [at all]?
502  MAR:@@@@@@@ [(h)]
510  MAR:[@@@@@@]
512  MAR:(.)(h)@@@@@
513       So did you get in li--
514       Did you jump in line?
515       for the cab?
519  MAR:(.) @°Nice°.
523  MAR:That’s funny.
         (.) Yeah,
         I guess I wouldn’t--
         (.) stand in a line like that.
         [I would]
530  MAR:[3go to a different3]--
          No.
537  MAR: No.
          You have to—
          You have to get out there and
539  MAR:@@@
550  MAR:[Yeah].
553  MAR:Yeah.
         (1.2)
```

We can observe that Mary’s emotion becomes more and more positive with the progress of Jason’s story. At the end, she not only compliments Jason’s behaviors, but also declares that she would have done the same if she were in the same situation. Mary’s laughter displays her alignment with Jason, which is one of the obvious emotional cues other than the lexical and syntactic cues. In this joyful conversation, the two recipient’s emotions are expressed by direct verbal compliments, but far more often they are expressed through other cues, mainly laughers and curious facial expressions.

5 EXPRESSION OF FRUSTRATION

Emotions in arguments have been proposed or studied by researchers in linguistic-interactional approach, sociolinguistics, conversation analysis, sociology and discourse analysis. Scholars who study arguments or conflicts in naturally occurring data put their focus on the sequential positioning and turn-taking in interaction from the analytic and interpretative perspectives [25-26][16]. Pomerantz [26] suggests that agreements are performed with a minimization of gap between the prior turn’s completion and the agreement turn’s initiation; disagreement components are frequently delayed within a turn or over a series of turns. Schiffrin [27] proposes that turn taking becomes more competitive during verbal conflict. Overlaps and interruptions are frequent. The oppositional turns can be performed in a mitigated or aggravated manner [25]. [25] has analyzed the sequential organization of turns and the sequential organization of closing. The authors examine video-recording of young girls playing hopscotch, they effectively display emotional ‘stance’ toward actions by their co-participants through precise coordination of pitch elevation, intonation, syntactic choice, timing and gesture. Vuchnich [28] has proposed three different closings of arguments, which are win, loss and stand-off or withdrawal. Our study aims to find out the ways of expressing frustration from the naturally occurring English conversation.

The data of argument is around 4 minutes long. The argument is if Mr. Bush will be re-elected as president in 2004. Jason holds the opinion that Bush will be re-elected because a war always gets politicians re-elected. However, Mary is strongly against this opinion. The argument is an instance of oppositional argument, in which two or more speakers openly engage in disputing over a position across a series of turns [19]. The argument follows the sequence of action and opposition. Jason is the action party who proposes his opinions first. Mary is the oppositional party.

In the data, the speakers try to defend their opinions in the argument. And both of them have experienced setback and frustration. Neither of them is able to persuade r or succumb to each other.

There are many different structures for accomplishing oppositional turns at talk. These include disagreement, challenge, denial, accusation, threat, and insult [19]. While as the opposition party, Mary's emotion is conveyed by her turn-taking and the feature of prosody. Taking the conversational floor has become very competitive, so it leads to frequent overlaps and interruptions in this segment. Overlaps and interruptions are the emotional cues of competition and frustration. Overlap is used as a way to give Jason pressure. The main five features of Mary's emotion display are as follows:

- Loud pitch for argument
- Interruption of the prior turn of the action party
- Immediate negation forms for disagreement in TCU, TCU is the basic unit of one turn, which can be words, phrases and clauses.
- Wh-questions with falling pitch contour
- Using sarcasm to show frustration

Example 2 is the transcript of the argument. Table 2 summarizes the expression of emotion of Mary.

Example 2- I study politics. It's my life.

```
10.    JAS:I think I'm going to move--
11.        to Finland.
12.        for at least,
13.        two more years.
14.        (0.9)
15.    MAR:[Why].
16.    JAS:[Course] he'll get reelected,
17.    so then I have to stay [2(there) XXX (years)2].
18.    MAR:                             [2No,
19.                                     he's not going2],
20.         to get reelected.
21.         (.)
22.    JAS:He is.
23.    MAR:No,
24.        He's not.
25.    JAS:He is.
26.    MAR:No.
27.        (.) Why are you saying that.
28.    JAS:I study politics.
29.        [It's my life].
30.    MAR:[I study politics] too man,
31.        And he's not getting re  [2elected2].
32.    JAS:                         [2It's2] my
                          life.
33.        (0.7)
34.         A war always gets--
35.         (.) politicians reelected.
```

Table 2. Active frustration expression of Mary

Emotional cues	Syntax	Prosody	embodied Action	Emotion display
(.) Why are you saying that (see line 26,27).	Interrogative with falling pitch contour	higher pitch, loudness, falling contour	looking straight in Jason's face, bending upper body forward towards Jason	Active frustration, criticism
[I study politics] too man, And he's not getting re[2elected2](see line30,31).	overlap, other repetition	higher pitch, loudness, fall-rising contour	looking straight in Jason's face, bending upper body forward towards Jason	Active frustration, sarcasm

From the above data, we find out that expression of emotion in interaction is not isolated; it is interwoven by all the emotional channels. Emotion is the combination of the linguistic, paralinguistic and kinesic features. We generalize that the expression of frustration can be conveyed by active behaviors. In the above case of active frustration, the frustrated person "lashes out" verbally and physically at an intended target.

Mary's frustration is regarded in the above data as the active type, or active frustration, because she defends her opinions actively by some aggressive behaviors, such as the tendency to lean forward towards the target of anger. However, frustration can also be conveyed by passive actions. Example 3 is the continuation of the argument.

Example 3 - A war always gets politician reelected.

```
36.     MAR:Not i--
37.          if there's so many body     bags,
38.          that-- it--
39.          covers the White  [House]
40.     JAS:                 [The war] in Iraq,
41.         would never cause that many body
    bags.
42.     MAR:You have no idea.
43.     JAS:I [XXXX].
44.     MAR:  [Nobody knows].
45.             (1.3)
46.     JAS:I think it would be--
47.         It's a technological war,
48.         so it wouldn't be a problem.
49.         (.)
50.         I-- if--
51.     MAR:It's not like,
52.         in ninety-one,
53.         when they had all the support.
54.     JAS:If,
55.         (1.3)
56.         If.
57.         (.)
58.         They lost a lot of casualties.
59.         He would have to,
60.         go against,
61.         his own policy,
62.         and then pull out,
63.         and then he'd be a hero for
             pulling out,
64.         and he'd still get reelected,
65.         but the odds of him,
66.         (1.2)
67.         even having a body bag problem,
68.         before his reelection occurred,
69.         would be,
70.         slim.
71.         (2.7)
72.     SOP:When's the next elections?
73.     JAS:Two Th[ousand Four].
74.     MAR:      [Two Thousand] Four,
75.         (1.6)
```

Tannen [20] argues that silences and pauses actually display tension and high emotion. In Jason's turn, starting from line 63, first Mary shakes her head twice, then puts her one hand under her chin and starts avoiding eye contact with Jason. Her passive embodied actions include these consecutive actions: head shaking, putting one hand under her chin, bending down her head, blinking her eyes and turning away, turning her body orientation away from Jason, disappointed facial expression, fingers writing and moving on the table. She becomes more frustrated shown by her passive embodied. Here we make a contrast between Jason's assertive arguments and Mary's emotional embodied actions. Table 3 presents a comparison of their emotion expression.

Table 3. Comparison between Jason's assertive argument and Mary's embodied sequence

Jason's assertive argument	Jason's emotion	Mary's embodied sequence	Mary's emotion
From line54 to line 62	confident	Listening to Jason with eye contact	calm
From line63 to line 71	Assertive, speaking with quicker tempo	Eye blinking, avoiding eye contact, disappointed facial expression, head shaking, fingers writing and moving on the table, turning her body orientation with left hand supporting her chin.	Passive frustration

One significant feature of Mary's embodied action is that she brings out two hands to convey or enhance her emotion expression, starting in Jason's line 63. Before line 63, both her hands are put on her lap under the table. After Jason's if/then turn Mary begins using her prosody, and the facial expression together with the hand gestures to assist her emotion expression.

When frustration is expressed passively, it is characterized by evasive behaviors and tension. Mary's passive frustration lasts till Jason finishes his turn with a long pause. There is a 2.7- second pause in line 71 without anybody taking the floor. In this segment, Mary's passive frustration is expressed by her embodied action. Mary's frustration leads to the change of body posture, and body orientation. This passive frustration suggests that Mary is suffering a setback.

6 CONCLUSION

The expression of Emotion is co-constructed by both verbal and non-verbal cues. However, Joy and frustration are expressed by different linguistic, paralinguistic features and embodied actions.

The joyful expression in the first data is carried out by verbal cues with curiosity, interest, and excitement and compliment. In addition to that, the joy is expressed by loud laughter, smile and concentrated facial expression. Emotion contagion is obvious among speakers. All the speakers collaborate joyfully to push the story to highlight. And the finding supports the positive emotion theory of Fredrickson. She has argued that positive emotions have a complementary effect: they broaden people's momentary thought-action repertoires, widening the array of the thoughts and actions that come to mind: to play and create when experiencing joy, to explore when experiencing interest, to savor and integrate when experiencing contentment, and to combine play, exploration, and savoring when experiencing love [21].

Speakers convey frustration by virtue of paralinguistic features and embodied actions simultaneously at the other speaker's turn in argument. The expression of emotion in interaction is not isolated; it is interwoven by all the emotional channels. It is hard to link physiology and emotions [22]. The expression of frustration can be conveyed by active behaviors. In the above cases of active frustration, the frustrated person "lashes out" verbally or physically at an intended target. Frustration is associated with special types of paralinguistic features and embodied actions in the data. Active frustration is closely related to competitive turn-taking, overlaps, interruptions of the aspects of sequential positioning. And active frustration is often conveyed by high pitch and loudness. In addition, the speakers prefer to use aggressive facial expression and hand gestures to enhance their frustration. Speakers mirror each other's frustration easily. As a result, the argument can become aggravated during the expression of active frustration. When frustration is a passive emotion, it is characterized by evasive behaviors and tension. In our data, the speakers are not distracted by passive emotional information transmitted from the facial expression embodied actions of their oppositional party. They carry on the argument even if they discern the frustration of their opponent. During the whole argument, the recipients do not consider the passive emotion of their opponents; instead they take the passive frustration as a sign of their own victory and try to take the opportunity to achieve their goal.

ACKNOWLEDGEMENT

This work was partially supported by the project Ubiquitous Computing and Diversity of Communication (MOTIVE) founded by the Academy of Finland's Research Program.

REFERENCES

[1] Chafe, W. (1994). Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing. Chicago: University of Chicago Press., 1994,

[2] Fehr, B. and Russell, J.A.(1984). "Concept of emotion viewed from a prototype perspective," Journal of Experimental Psychology: General, pp. 464-486, 1984.
[3] Fehr, B., Russell, J.A. and Ward, L. M. (1982). "Prototypicality of Emotions: A reaction time study," Bulletion of the Pshchonomic Society, pp. 253-264, 1982.
[4] Shaver P., Schwartz J., Kirson D. and O'Connor C. (2003). Emotion Knowledge: Further Exploration of a Prototype Approach. Emotions in Social Psychology: Essential Readings (in W. Gerrod Parrott Ed.). Social Psychology, vol. 42, 2003, pp. 513–531.
[5] Schwartz J. and Shaver, P. R. (1987). "Emotions and emotion knowledge in interpersonal relations," in Advances in Personal Relationships , vol. 1, W. Jones and D. Perlman, Eds. Greenwich, CT: JAI Press, 1987, pp. 197-241.
[6] Averill, J. R. (1982). Anger and Aggression:An Essay on Emotion. New York: Springer-verlag, 1982,
[7] Scherer K. R. and Wallbott H. G.(1994). "Evidence for universality and cultural variation of differential emotion response patterning," Journal of Personality and Social Psychology, vol. 66, pp. 310–328, 1994.
[8] Gottman J. M. and Levenson R. W.(1985). "A valid procedure for obtaining self-report of affect in marital interaction " Journal of Consulting and Clinical Psychology, vol. 53, pp. 151-60., 1985.
[9] Wiggins Sally and Potter Jonathan(2003). "Attitudes and evaluative practices: Category vs.item and subjective vs. objective constructions in everyday food assessments " British Journal of Social Psychology, vol. 42, pp. 513–531, 2003.
[10] Sacks, H.(1992). Lectures on Conservation. , vol. 1, Oxford: Blackwell, 1992,
[11] Sacks,H.(1974). "An analysis of the course of a joke's telling," in Explorations in the Ethnography of Speaking R. Bauman and J. Scherzer, Eds. Cambridge: Cambridge University Press, 1974, pp. 337-353.
[12] Goodwin M. H. and Goodwin C.(2000)., "Emotion within situated activity," in Communication: An Arena of Development. Http://www.Sscnet.Ucla.edu/clic/cgoodwin/00emot_act.Pdf Budwig, N.,Ina Uzgiris and James Wertsch, Ed. Stamford: Ablex Publishing Corporation, 2000, pp. 33–53.
[13] Jefferson, G. (1984)."On the organization of laughter in talk about troubles," in Structures of Social Actions: Studies in Conversational Analysis Anonymous London: Cambridge Univeristy Press, 1984, pp. 346-349.
[14] Heritage, J. (1984).Garfinkel and Ethnomethodology. Cambridge: Polity Press, 1984,
[15] Kärkkäinen, E.(2003). Epistemic Stance in English Conversation. A Description of its Interactional Functions, with a Focus on I Think. Philadelphia, PA, USA: John Benjamins Publishing Company, 2003,
[16] Sandlund, Erica (2004). Feeling by Doing the Social Organization of Everyday Emotions in Academic Talk-in-Interaction. Karlstad University, 2004,
[17] Davis M. H.(1983). "Measuring individual differences in empathy: Evidence for a multi-dimensional approach," Journal of Personality and Social Psychology, vol. 44, pp. 113-126, 1983.
[18] Kahnand BE and Isen AM (1993). "The influence of positive affect on variety seeking among safe, enjoyable products," J Consum Res, vol. 20, pp. 257-270, 1993.
[19] Schiffrin D.(1987)., Discourse Markers. Cambridge University Press, 1987,
[20] Tannen, D. (1984).Conversational Style Analyzing Talk among Friends. Ablex Publishing Corporation, 1984,
[21] Fredrickson, B. L.(1998). "What good are positive emotions?" Review of General Psychology: Special Issue: New Directions in Research on Emotion, vol. 2, pp. 300–319, 1998.
[22] Cacioppo, J. T. Klein, D. J., BerntsonG. C. and Hatfield, E. (1993). "The psychophysiology of emotion," in Handbook of Emotions M.Lewis and J.M. Haviland, Eds. New York: Guilford Press, 1993, pp. 119-142.
[23] Bois D. and Danae P.(1993)., "Outline of discourse transcription," in Talking Data: Transcription and Coding in Discourse Research J. A. Edwards and M. D. Lampert, Eds. Hillsdale, NJ: Erlbaum, 1993, pp. 45-89.
[24] Chafe W.(1994). Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing. Chicago: University of Chicago Press. 1994.
[25] Goodwin, M.H. and Goodwin, C. (2000). Emotion within situated activity. In N. Budwig, I. C. Uzgiris and J. V. Wertsch (Eds.). Communication: An arena of development. Stamford, CT: Ablex, 2000, pp. 33-54.
[26] Pomerantz A. (1984). Agreeing and disagreeing with assessments: some features of preferred/disperferred turn shapes. In J.M. Atkinson and J. Heritage (Eds.). Structures of Social Action. Studies in Conversation Analysis. Cambridge: Cambridge: Cambridge University Press, 1984, pp. 57-101.
[27] Schiffrin, Deborah. (1990). The management of a co-operative self during argument: The role of opinions and stories. In Grimshaw, Allen D.(ed.), Conflict Talk. Sociolinguistic investigations of arguments in conversations. Cambridge University Press. 1990.
[28] Vuchinch Samul. (1990). The sequential organization of closing in verbal family conflict. In Allen D.Grimshaw.Conflict Talk. Sociolinguistic investigations of arguments in conversations. Cambridge University Press.1990.
[29] Kipp Michael. (2003). "Anil 4.0 annotation of video and spoken language, user manual," University of the Saarland, German Research Center for Artificial Intelligence, Germany. 2003.

Affective Support in Narrative-Centered Learning Environments

Jennifer Robison
North Carolina State University
Raleigh, North Carolina, USA
jlrobiso@ncsu.edu

Abstract

The link between affect and student learning has been the subject of increasing attention in recent years. Affective states such as *flow* and *curiosity* tend to have positive correlations with learning while negative states such as *boredom* and *frustration* have the opposite effect. Consequently, it is a goal of many intelligent tutoring systems to guide students toward emotional states that are conducive to learning through affective interventions. While much work has gone into understanding the relation between student learning and affective experiences, it is not clear how these relationships manifest themselves in narrative-centered learning environments. These environments embed learning within the context of an engaging narrative that can benefit from "affective scaffolding." However, in order to provide an optimal level of support for students, the following research questions must be answered: 1) *What is the nature of affective experiences in interactive learning environments?* 2) *How is affect impacted by personal traits, beliefs and learning strategies, and what role does affect have in shaping traits, beliefs, and learning strategies?* 3) *What strategies can be used to successfully create an optimal affective learning experience?*

Keywords: Affective Interfaces, Intelligent Agents, Applications in Education

1 INTRODUCTION

Affect has begun to play an increasingly important role in intelligent tutoring systems. The intelligent tutoring system community has seen the emergence of work on affective student modeling [1], detecting frustration and stress [2, 3], modeling agents' emotional states [4, 5], devising affective-informed models of social interaction [6, 7], detecting student motivation [8], and diagnosing and adapting to student self-efficacy [9]. All of this work seeks to increase the fidelity with which affective and motivational processes are understood and utilized in intelligent tutoring systems in an effort to increase the effectiveness of tutorial interactions and, ultimately, learning.

This level of emphasis on affect is not surprising given the impact it has been shown to have on learning outcomes. Student affective states influence problem-solving strategies, the level of engagement exhibited by the student, and the degree to which he or she is motivated to continue with the learning process [10, 11, 12]. All of these factors have the potential to influence both how students learn immediately and their learning behaviors in the future. Consequently, developing techniques for keeping students in an affective state that is conducive to learning has been a focus of recent work [13, 14, 15].

Unfortunately, there is not yet a clear understanding of how emotions occur during learning and, in particular, how individual learning environments impact the emotional experience. It is also unclear which emotional states are optimal for individual students. This is likely to vary based on student needs and experience. Affective experiences may also have interesting immediate and long term effects on how students perceive learning and their levels of confidence and motivation moving forward. Finally, current research on how best to respond to student affect has yielded varying and often conflicting conclusions [9, 16, 17]. For these reasons, it is challenging to design affective support systems for learning environments.

The goal of this research is to examine these issues within narrative-centered learning environments. These environments embed the educational process within a story with the objective of leveraging narrative's motivating features such as compelling plots, engaging characters, and fantastical settings [18]. These environments also offer the potential for affective experiences that are supplementary to those experienced in more typical interactive learning environments [19]. The ability to understand and control the emotional experiences of students in narrative-centered learning environments could lead to significant gains for student learning and motivation. The proposed research aims to achieve this goal by answering the following three research questions: (1) What is the nature of affective experiences in interactive learning environments? (2) How is affect impacted by personal traits, beliefs and learning strategies and what role does affect have in shaping these qualities? (3) What strategies can be used to successfully create an optimal affective experience?

2 RESEARCH QUESTIONS

The overarching goal of the proposed research is to *characterize* and *facilitate* affective experiences in narrative-centered learning environments in order to achieve *optimal* immediate learning gains and future long-term outcomes. This goal can be achieved through exploration of three lines of investigation. The first relates to the *characterization* of affective experiences. In particular it is important to understand the types of emotions that are experienced in narrative-centered learning environments as well as the antecedents and consequences associated with each. The second research question relates to defining *optimal* experiences. Many user traits and beliefs about learning may impact affective responses and may in turn be altered by powerful learning experiences. Understanding the interrelation between these personal traits and student affect will contribute to a clearer understanding of optimal outcomes. Finally, the third research question builds directly on the second by attempting to *facilitate* affective experiences. These efforts are intended to yield a variety of strategies that can be shown to induce optimal affective experiences. Collectively, these three areas of investigation suggest a comprehensive approach to supporting student affect in interactive learning environments.

2.1 NATURE OF AFFECTIVE EXPERIENCE

While much effort has been invested by the affective community to characterize the emotional experiences associated with learning, there is still no comprehensive model of the role that affect plays in learning. Understanding the role that affect plays in interactive learning environments, which is the focus of this research, is especially challenging. The types of emotions students experience in one-on-one tutoring may vary significantly from those experienced in computational environments, and the characteristics of particular learning environments may further influence the range of emotions experienced. For instance, the presence of pedagogical agents may promote the occurrence of social emotions, such as *pride*, while learning environments with a narrative focus offer the opportunity to investigate a potentially different set of affective states.

While the precise cognitive and affective mechanisms underlying learning experiences is not yet well understood, there has been significant progress in attempting to identify the emotions that students are likely to experience and how these may affect the learning process. For instance Kort *et al.* present a model of learning emotions which can be represented as a cycle that occurs throughout the learning process [10]. Other studies have investigated how emotional experiences transpire in computational environments. Both D'Mello *et al.* and Baker *et al.* have shown that students are most likely to remain in the same state through time and that certain emotional transitions are more likely than others [20, 21]. The results of these studies, and their replication in subsequent work, suggest that there is an underlying model modulating the likelihood of students experiencing particular affective states. Understanding the details of this model is the major goal of this first research question.

2.2 PERSONAL TRAITS AND AFFECT

The next research question relates to how individual traits can impact and be impacted by the affective experiences of learning. There are many traits, such as goal orientation, self-efficacy and motivation, that significantly impact how students approach learning situations. Consequently, these differential traits likely result in varying emotional experiences. Similarly, our experiences shape our beliefs and approaches to learning. Therefore, strong or repeated affective experiences can greatly impact student traits in future learning. Understanding how these two notions of individual traits and affective experiences relate to one another is key to supporting student affect.

The impact of student traits on affective experiences can clearly be seen when examining the construct of goal orientation. *Goal orientation* refers to the focus of student behavior in learning situations [22]. Students are typically classified as having a mastery or performance orientation. *Mastery* orientation refers to the tendency of a student to focus on the academic material and particularly on the acquisition of knowledge and skill for their own intrinsic values. Alternatively, *performance* oriented students focus on how well they are achieving particular tasks and how their performance compares to others [22]. Interestingly, goal orientation has been shown to impact student affective experiences in response to learning tasks of varying difficulties [23].

Self-efficacy is also strongly related to emotional experiences. Defined as an individual's belief in his or her ability to successfully complete a given task [24], *self-efficacy* has very strong correlations with hope and optimism. These types of emotional experiences are likely to carry students through difficult tasks, and students with high self-efficacy are also less likely to experience negative emotions such as discouragement and despair which may cause them to avoid learning tasks. The experience of these negative states can also result in further decreases in confidence and self-efficacy while positive experiences improve student confidence in their ability to succeed [24]. Additionally, negative states experienced during learning may decrease student motivation to learn and pursue more difficult tasks, while positive affective states can improve student motivation [25].

While there are very clearly some strong examples of the interplay between personal traits and affective experiences, these are not understood in depth, nor is it clear how these finding carry over to interactive learning environments. Because these traits, and the experiences associated with them, have such strong impacts on learning, developing a clear understanding of their interaction is a critical next step in developing affective support strategies.

2.3 ACHIEVING OPTIMAL EXPERIENCE

While understanding the affective experiences learners encounter and the interplay of these experiences with personal traits is interesting in its own right, developing strategies that take advantage of these findings is the ultimate object of this research. Pedagogical interventions that focus on individual student needs and provide affective support throughout the learning process are likely to show increased learning immediately as well as have positive impacts on student attitudes toward learning in the future. Since a student's emotional state can strongly impact how the student learns [12] and interacts with learning environments [21], it is important to develop a clear understanding of what types of interventions create positive affective experiences. Additionally, with the insight gained from investigating the impacts of affect on long-term personal traits, these interventions have the potential to produce lasting results. Unfortunately, it is not yet clear how the variety of social and pedagogical strategies can be used to provide optimal affective support for individual students.

To date, a broad range of strategies has been suggested for improving student affect. Chaffar and Mclaren [13] propose an Emotional Intelligent Agent which utilizes guided imagery, music and presented images in an attempt to induce optimal emotional states for each student. D'Mello *et al.* [14] have also proposed methods for responding to student affect using empathetic and tutorial dialogue acts accompanied by visual facial expressions and emotionally synthesized speech. Alternatively, Murray and VanLehn provide affective support without directly encouraging or displaying emotional states [26]. Instead, they use a decision theoretic approach to determine when the delivery of hints may hinder or improve student morale and independence.

Though each of these strategies has the overarching goal of encouraging students to remain in emotional states that are conducive to learning, they utilize very different approaches. Understanding the effects of these types of approaches on individual students will support more precise affective intervention. This line of investigation constitutes the final component of the research agenda proposed for delivering optimal affective support. Together with a deep understanding of the affective tendencies of students and how these emotions affect lifelong learning, these intervention strategies will enable us to design interactive learning environments that directly address improved affect-informed learning experiences.

3 PROGRESS TO DATE

Work has begun on each of the research directions outlined above. These efforts have centered on empirical studies of the affective experiences of students interacting with the narrative-centered learning environment, CRYSTAL ISLAND [19, 27, 28]. This environment is being created in the domains of microbiology and genetics for middle school students. It features a science mystery set on a recently discovered volcanic island where a research station has been established to study the unique flora and fauna.

The user plays the protagonist, Alex, who is attempting to discover the source of an infectious disease at the research station. The story opens by introducing the student to the island and the members of the research team. As members of the research team fall ill, it is her task to discover the cause and the specific source of the outbreak. She is free to explore the world and interact with other characters while forming questions, generating hypotheses, collecting data, and testing her hypotheses. She can pick up and manipulate objects, and talk with characters to gather clues about the source of the disease. In the course of her adventure she must gather enough evidence to correctly identify the type and source of the disease that has infected the camp members.

To date, two studies have been completed to understand the affective experiences of students within the CRYSTAL ISLAND environment and examine the impacts of character-delivered affect support. The first of these studies (**Study 1**) focused on character empathetic behaviors in response to student affect. The subjects of the study consisted of 35 college students ranging in age from 21 to 60 (M = 24.4, SD = 6.41) including 9 females and 26 males. Among these students, 60% were Asian (n = 21), approximately 37% were Caucasian (n = 13) and one participant chose not to respond. The second study (**Study 2**) supplemented the empathetic feedback of agents with task-based support. The subjects of this study consisted of 41 college students ranging in age from 19 to 38 (*M* = 24.0, *SD* = 3.96) including 12 females and 29 males. Among these students, approximately 73% were Caucasian (*n* = 30), 17% were Asian (*n* = 7), and 10% were Other (*n = 4*).

Both studies followed a similar paradigm for measuring and responding to student affect. In each case, affective support was provided through text-based character dialogue. When subjects decided to interact with the agents, the agent would begin the conversation by asking the question, "Hi Alex, how are you feeling?" The subject would then self-report on their affective state by selecting one of the available emotions (*anger, anxiety, boredom, confusion, curiosity, delight, excitement, flow, frustration, sadness, fear*). The agent then responds to the subject's reported affective state with a randomized feedback response. In Study 1, feedback responses varied between parallel and reactive empathetic statements. In Study 2, these responses were varied between empathetic statements (both parallel and reactive) and task-based feedback. More details on these feedback strategies are presented in Section 3.3. The dialogue would then continue with the regular narrative content. At the conclusion of the interaction, the agent again asks the subject how she feels by asking "How are you feeling now?" to which the subject provides a second self-reported emotional state.

3.1 CHARACTERIZING STUDENT AFFECT (STUDY 1)

Initial work at characterizing the affective experiences of students in CRYSTAL ISLAND has shown some interesting similarities and differences from the results found in other learning environments [19]. For instance, the optimal learning emotion *flow* was reported most frequently (42% of self-reports) both in the CRYSTAL ISLAND environment and in different types of environments such as the 2-dimensional games and dialogue-based tutorial systems used by Baker *et al.* and D'Mello *et al,* respectively [20, 21]. Interestingly, *boredom* (3%) was reported less frequently and positive emotions such as *excitement* (14%) and *delight* (11%) were reported more often in the narrative-centered learning environment. Perhaps it is the interactive narrative nature of this environment that facilitates the occurrence of these emotions.

Analysis of the affective transitions experienced by students in CRYSTAL ISLAND also mirrored those of the other learning environments [19]. In general, students have a strong tendency to remain in the same affective state across time. However, when transitions to alternate affective states did occur, they followed interesting patterns. For instance, frustrated learners were very likely to transition to *confusion* or *fear* and were particularly unlikely to enter a positive state such as *flow* or *excitement*. Students experiencing the positive state of *flow* were likely to transition to *confusion*, which is still considered positive for learning and were unlikely to transition to the more negative state of *frustration*. Interestingly, confused learners were equally likely to transition to *flow* and *frustration*[19]. These findings suggest that the affective state of *confusion* and its antecedents and consequences are worth additional study to determine which factors contribute to a positive transition to *flow* or a negative transition into *frustration*.

3.2 IMPACT OF STUDENT TRAITS ON AFFECTIVE EXPERIENCE (STUDY 1)

Similar work has also been done to investigate how personal characteristics impact the affective experiences of students in CRYSTAL ISLAND [27]. The personal characteristics examined for this investigation include personality, goal orientation and gender. Personality was measured using the Big 5 Personality Questionnaire, which indexes student personality across five dimensions: openness, conscientiousness, extraversion, agreeableness and neuroticism [29]. Goal orientation measures students' objectives when engaged in learning activities and is measured in terms of mastery or performance approaches [22].

Interesting results were found across these dimensions both in the frequency of emotions experienced by students and their affective trajectories throughout the learning process. For example, extraverted students reported emotions such as *delight* ($t(34) = 1.82$, $p = .07$) and *anger* ($t(34) = 2.77$, $p = .009$) more frequently than introverted students who were more likely to report being in *flow* (t(34) = 2.14, p = .04). This suggests that extraverted students may focus more on the characters and narrative components of the environment while introverted students are more focused on the learning tasks. Interestingly, boredom was reported only by male students.

Students who were focused on their performance in the environment reported significantly more *anger* (*t*(34) = 2.28, *p* = .03) and *anxiety* (*t*(34) = 1.71, *p* = .09) than students who were focused on the educational content. These students reported significantly more *flow (t(34) = 2.25, p = .03)*. Goal orientation also had significant impacts on the type of emotional trajectories experienced. For instance, contrary to the typical trend, mastery oriented students were unlikely to remain in a state of *confusion* or *boredom*. Instead, they quickly resume engagement in learning activities and transition into *flow*. Performance oriented students in these cases are more likely to stay *bored* or *confused* or alternatively transition to *frustration* or *anxiety*. This distinction may provide additional insight into how *confusion* is experienced and explain the interesting findings reported in Section 3.1.

In order to consider how affective experiences may affect student learning and motivation, the notion of presence was also measured. Presence refers to students' active engagement within the environment and is hypothesized to impact motivation and willingness to utilize these types of environments for learning in the future [30]. Interestingly, *frustration* was reported more frequently (*t*(34) = 1.70, *p* = .09) by students who were not engaged in the environment and perhaps could be the source of this disengagement. Alternatively, students who reported high levels of presence reported more *anxiety* (t(34) = 2.23, p = .03*)*, perhaps because of their increased level of involvement. These students were also unlikely to remain *bored* and had a higher tendency to transition to *confusion, excitement* or *flow*. Students reporting low levels of involvement were more likely to remain *bored.*

3.3 Intervention Strategies (studies 1 and 2)

In addition to understanding the types of emotions experienced by students, several intervention strategies have been investigated in order to help maintain positive student affect. These strategies include affect-based empathy and task-based hints and suggestions. Empathy is defined as an awareness of another's affective state that generates emotions in the empathizer that reflect more than her own situation in attempt to foster a feeling of understanding or to motivate a more positive affective state [31]. Empathy can be further distinguished by two subtypes: parallel and reactive empathy. *Parallel* empathy typically involves an individual displaying an emotional state similar to that of the target of an empathetic response. In this case, the empathizer will demonstrate an understanding of the target's emotional state with a focus on the relevant situational factors. For instance, a parallel empathetic response to a frustrated student could be, "I'm frustrated too! This material is very difficult." Alternatively, in *reactive* empathetic responses, individuals exhibit a deeper level of understanding and focus on bringing about the optimal affective state of the target. Often this will manifest itself as a demonstration of a different affective state from the target. For example, "Don't worry, I'm sure you'll figure this out soon!" would be a reactive response to a frustrated student.

In Crystal Island, characters were given the capability of responding empathetically to students. Investigation of student responses indicated some interesting differences in how students react to exhibitions of parallel and reactive empathy. Students met with parallel empathy had a strong tendency to remain in the same or similar affective state [19]. This was true whether the student was in a positive or negative state. Alternatively, students who received reactive responses tended to transition to very different states. If the student was exhibiting a negative state, a reactive response would successfully encourage them to enter a more positive affective state. Unfortunately, reactive responses would bring students in a positive state down to a more negative state, often *confusion*. It appears that attempting to further motivate a student who is already feeling positive has unfortunate side effects.

These findings were used in a follow-up study, which sought to examine the instances in which students preferred affective feedback or task-related feedback to their emotional states [28]. In this study students were asked to rate the helpfulness of empathetic responses and task-based hints. In the case of empathetic responses, students received a parallel empathetic statement in response to positive emotion and reactive responses to negative affective states. For task-based feedback, students would receive a summary of current progress when they reported a positive emotional state and would be given additional hints to help them overcome a negative state if it was reported. In general, task-based feedback was rated more helpful by students than the empathetic responses. While in most cases the ratings for task-based feedback were significantly higher than empathetic responses ($M_1 = 2.88$, $M_2 = 2.33$, $p < 0.001$), this was not the case for the emotions that were not directly associated with learning. These emotions, which tie more closely with the narrative plot of the environment, were best met with empathetic responses. The emotional impact of these strategies is currently being examined to determine if they can effectively promote positive affect.

4 Research Agenda

While current results are promising and provide some insight into how to properly support students' affective experiences, there are many areas that are yet to be explored. For instance, initial work has examined *which* affective states students report while engaging in learning activities. Future work will examine *when* and *why* these states occur. If a student is experiencing frustration it is likely very important to understand the source of that frustration in order to properly respond to it. The student may be experiencing difficulties with the learning material, the controls of the environment or may simply be irritated by characters who are attempting to provide feedback. Understanding the sources of affective states will not only help identify the most appropriate interventions but will also contribute to better designs that will enable negative emotions to be effectively managed.

Current work has shown how goal orientation, personality and gender can affect emotional experiences and how these emotional experiences relate to student involvement within the environment. Work to date covers only a small subset of personal traits and beliefs that could impact learning and emotional experiences. Examining the impact of self-efficacy, confidence and beliefs about learning on emotional experiences represents a very promising direction for future work. Additionally, it will be important to determine how affective experiences in the environment affect students immediately following the interaction and in the future. Correlations of emotions with learning gains will provide valuable information regarding ideal affective states. Further, it will be important to determine if inducing positive affect increases students' motivation and willingness to learn in the present and in theun future.

While several strategies for affective intervention have been examined, there are still many more that may impact students' affective experience. For instance, perhaps the physical appearance or narrative role of characters impacts students' perceptions of their responses. It has been shown that these factors can greatly impact students' affective states and feelings about learning [32], and it would be useful to understand how these factors interact with the use of intervention strategies. Additionally, there are other areas of investigation outside of agent feedback for providing affective support. Introducing companion agents may promote a more enjoyable experience for students and foster social emotions in addition to those currently experienced. There may also be methods of intervention through modifications to the environment itself. Perhaps the best way to help a frustrated student is to simplify the task they are trying to accomplish.

Finally, the current strategy for detecting and analyzing student affect is through self-report with character dialogue. To increase the fidelity of analysis and reduce invasiveness, alternate methods should be investigated. This line of investigation will include using facial, posture, and physiological analyses to determine student emotional states when characterizing student experience. Another promising area of work will be creating models of affect detection that can be run in real time to inform intervention strategies.

To pursue the researchgoals noted above, the following studies will be conducted:

- **Study 3** will focus on developing informed models of affect detection and understanding. During this study, middle school students will interact with the CRYSTAL ISLAND environment without interruption or affective intervention. Students will be monitored using video and physiological and posture sensors. Students' interactions with the environment and their personal characteristic traits will also be logged. Following the interaction, students will be presented with videos of themselves and their experience with the environment and asked to retrospectively report on the emotions they have experienced and their sources. This information will be integrated with assessments by trained judges and used to develop a comprehensive model of student emotions within the environment. These models will then later be used at run-time for student affect detection to replace self-reports.
- **Study 4** will utilize the models produced from the results of Study 3 to inform agent affective interventions. In this study agents will respond to student affect using task-based, empathetic or alternative feedback strategies. The effectiveness of these strategies will then be measured by students' immediate affective experiences and learning gains and future impacts on motivation and self-efficacy. The objective of this study is to develop decision-theoretic models which can utilize a broad range of information sources including student characteristics, affective experience and prior knowledge to estimate and weigh the consequences of each affective intervention.
- **Study 5** will supplement Study 4 by focusing on the image and role of the agents delivering affective feedback. In this study, agent age, gender, race and narrative role will be varied, along with the type of affective intervention strategies utilized. The goal of this study is to determine if agent traits have an impact on the effectiveness of intervention strategies and how student traits may interact with these effects.
- **Study 6** will represent the capstone of this line of work and will utilize the results of all previous studies. This study will compare versions of CRYSTAL ISLAND with and without informed affective support models. These models will include the automatic detection of affect as learned by Study 3 as well as the decision-theoretic models developed from results of Studies 4 and 5. Students will be exposed to the CRYSTAL ISLAND environment with variations of these models and the utility of each model will be determined through measures of student learning gains, motivation and self-efficacy.

5 Conclusions

Affect permeates every aspect of human experience including learning. The types of emotions we associate with learning influence how likely we are to actively engage in learning activities and perceive learning as a positive experience. Because of this significant impact on learning, it is important to develop an understanding of students' emotional states during learning experiences. The objective of the research agenda presented in this paper is to provide affective support for students in narrative-centered learning environments. These environments offer a broad range of opportunities for motivating students and encouraging positive affective experiences. Providing appropriate affective support requires a clear understanding of how students experience emotions within narrative-centered learning environments, how these emotions impact (and are impacted by) student traits and beliefs, and discovering strategies for affective intervention that can promote effective learning experiences.

Acknowledgements

The authors wish to thank the members of the IntelliMedia research lab for their assistance in implementing Crystal Island, Omer Sturlovich and Pavel Turzo for use of their 3D model libraries, Valve Software for access to the Source™ engine and SDK and a special thanks to Scott McQuiggan for his generous support and guidance. This research was supported by the National Science Foundation under REC-0632450, IIS-0757535, DRL-0822200 and IIS-0812291. This material is based upon work supported under a National Science Foundation Graduate Research Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Conati, C., and Mclaren, H. (2005) Data-driven refinement of a probabilistic model of user affect. *Proceedings of the 10th Intl. Conf. on User Modeling*, Springer-Verlag, New York, NY, 40-49.

[2] Burleson, W. (2006) *Affective learning companions: Strategies for empathetic agents with real-time multimodal affective sensing to foster meta-cognitive and meta-affective approaches to learning, motivation, and perseverance.* PhD thesis, Massachusetts Institute of Technology.

[3] McQuiggan, S., Lee, S., and Lester, J. (2007) Early prediction of student frustration. In Proc. of the 2nd Intl. Conf. on Affective Computing and Intelligent Interaction, Portugal.

[4] André, E., and Mueller, M. (2003) Learning affective behavior. In *Proceedings of the 10th Intl. Conf. on Human-Computer Interaction*. Lawrence Erlbaum, Mahwah, NJ, 512-516.

[5] Graesser, A., Person, N., Magliano, J. (1995) Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Journal of Applied Cognitive Psychology* 9: 495-522.

[6] Johnson, L., and Rizzo, P. (2004) Politeness in tutoring dialogs: "run the factory, that's what I'd do". *Proceedings of the 7th Intl Conf. on Intelligent Tutoring Systems*. Springer-Verlag, New York, NY, 2004.

[7] Paiva, A., Dias, J., Sobral, D., Aylett, R., Woods, S., Hall, L., and Zoll, C. (2005) Learning by feeling: Evoking empathy with synthetic characters. *Applied Artificial Intelligence*, 19:235-266.

[8] de Vicente, A., and Pain, H. (2002) Informing the detection of the students' motivational state: an empirical study. *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, 933-943.

[9] Beal, C. and Lee, H. (2005) Creating a pedagogical model that uses student self reports of motivation and mood to adapt ITS instruction. *Workshop on Motivation and Affect in Educational Software, in conjunction with the 12th Intl. Conf. on Artificial Intelligence in Education*.

[10] Kort, B., Reilly, R., & Picard, R. (2001) An affective model of interplay between emotions and learning: Reengineering educational pedagogy—building a learning companion. *Proceedings IEEE Intl. Conf. on Advanced Learning Technology: Issues, Achievements and Challenges*. Madison, WI: IEEE Computer Society.

[11] Picard, R., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., Machover, T., Resnick, M., Roy, D., and Strohecker, C. (2004) Affective learning – a manifesto. *BT Technology Journal*, 22(4).

[12] Schwarz, N. (2000) Emotion, cognition, and decision making. *Journal of Cognition and Emotion*, 14(4):443-440.

[13] Chaffar, S. Frasson, C. (2004) Using an emotional intelligent agent to improve the learner's performance. *Proceedings of the Workshop on Social and Emotional Intelligence in Learning Environments in conjunction with Intelligent Tutoring Systems*, Maceio, Brazil.

[14] D'Mello, S., Jackson, T., Craig, S., Morgan, B., Chipman, P., White, H., Person, N., Kort, B., el Kaliouby, R., Picard, R., Graesser, A. (2008) AutoTutor detects and responds to learners affective and cognitive states. *Proceedings of the Workshop on Emotional and Cognitive issues in ITS in conjunction with the 9th International Conference on Intelligent Tutoring Systems*

[15] Forbes-Riley, K. and Litman, D. (2007) Investigating human tutor response to student uncertainty for adaptive system development. *Proceedings the 2nd International Conference on Affective Computing and Intelligent Interactions*, Lisbon, Portugal.

[16] Boyer K., Phillips R., Wallis M. Vouk, M. and Lester, J. (2008) Balancing cognitive and motivational scaffolding in tutorial dialog, in *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, 239-249.

[17] Shute V. (2007) *Focus on Formative Feedback,* ETS, Princeton, NJ.

[18] Malone T. and Lepper, M. (1987) Making learning fun: a taxonomy of intrinsic motivations for learning. *Aptitude, learning, and instruction: III. Cognitive and affective process analyses,* 223-253.

[19] McQuiggan, S., Robison, J., and Lester, J. (2008) Affective transitions in narrative-centered learning environments. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, 490-499.

[20] D'Mello, S., Taylor, R.S., Graesser, A. (2007) Monitoring affective trajectories during complex learning. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, Austin, TX, 203-208.

[21] Baker, R., Rodrigo, M., and Xolocotzin, U. (2007) The dynamics of affective transitions in simulation problem-solving environments. *Proceedings the 2nd International Conference on Affective Computing and Intelligent Interactions*, Lisbon, Portugal, 666-677.

[22] Elliot, A., and McGregor, H. (2001) A 2 x 2 achievement goal framework. *Journal of Personality and Social Psychology*, 80(3):501-519

[23] Steele-Johnson, D., Beauregard, P., Hoover, P., and Schmidt, A. (2002) Goal orientation and task demand effects on motivation, affect and performance. *Journal of Applied Psychology,* 85(5):724-738.

[24] Bandura, A. (1977) Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review*, 84(2):191-215

[25] Pekrun, R. (1992) The impact of emotions on learning and achievement: towards a theory of cognitive/motivational mediators. *Journal of Applied Psychology,* 41(4):359-376.

[26] Murray, C. and VanLehn, K. (200) DT-Tutor: a decision-theoretic, dynamic approach for optimal selection of tutorial actions. *In Proceedings of the 5th International Conference on Intelligent Tutoring Systems.*

[27] Robison, J., McQuiggan, S., and Lester, J. (2008) Differential affective experiences in narrative-centered learning environments. *Proceedings of the Workshop on Emotional and Cognitive issues in ITS in conjunction with the 9th International Conference on Intelligent Tutoring Systems.*

[28] Robison, J., McQuiggan, S. Lester, C. (2009) Modeling task-based vs. affect-based feedback behavior in pedagogical agents: an inductive approach. *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED)* Brighton, England.

[29] McCrae, R. and P. Costa (2003). *Personality in adulthood: A five-factor theory perspective.* New York: Guilford Press.

[30] Rowe, J., McQuiggan, S. and Lester, J. (2007) Narrative presence in intelligent learning environments *Working Notes of the 2007 AAAI Fall Symposium on Intelligent Narrative Technologies*, 126-133.

[31] Davis, M. (1994) *Empathy: A Social Psychological Approach*. Brown & Benchmark Publishers, Madison, WI.

[32] Baylor, A. (2005) The impact of pedagogical agent image on affective outcomes. *Proceedings of the Workshop on Affective Interactions: Computers in the Affective Loop in conjunction with the International Conference on Intelligent User Interfaces*

Social Network Analysis in Multimedia Indexing: Making Sense of People in Multiparty Recordings

Sarah Favre
Idiap Research Institute
CP 592, 1920 Martigny, Switzerland
Ecole Polytechnique Federale de Lausanne
1015 Lausanne, Switzerland
sfavre@idiap.ch

Abstract

This paper presents an automatic approach to analyze the human interactions appearing in multiparty data, aiming at understanding the data content and at extracting social information such as *Which role do people play?*, *What is their attitude?*, or *Can people be split into meaningful groups?*. To extract such information, we use a set of mathematical techniques, namely *Social Networks Analysis* (SNA), developed by sociologists to analyze social interactions. This paper shows that a strong connection can be established between the content of broadcast data and the social interactions of the individuals involved in the recordings. Experiments aiming at assigning each individual to a social group corresponding to a specific topic in broadcast news, and experiments aiming at recognizing the role played by each individual in multiparty data are presented in this paper. The results achieved are satisfactory, which suggests on one side that the application of SNA to similar problems could lead to useful contributions in the domain of multimedia content analysis, and on the other side, that the presented analysis of social interactions could be a significant breakthrough for affective computing.

Keywords: Social Network Analysis, Role Recognition, Story Segmentation, Broadcast data, Meeting Recordings.

1 Introduction

The amount of audio and video material available in digital form is increasing rapidly with the progress of capture and storage technologies, but without effective techniques for structuring its content, it is not possible to make an asset of it. Many research efforts have addressed the problem of extracting information from data for indexing purposes. For example, existing systems extract the information from: automatic speech transcriptions, shot transitions in video, faces and objects in images. However, when I started working on my PhD thesis three years ago, no major efforts were made to automatically analyze social interactions, even if they represent a common subject of multimedia recordings. Human interactions are present everywhere: in movies, television shows, broadcast news, radio programs, meeting recordings, call center conversations, etc. Moreover, psychologists showed that social interactions are one of the main channels through which we understand reality (Kunda, 1999). For these reasons, we decided to investigate automatic approaches for the audio content analysis, based on the extraction of social interactions in multiparty recordings.

We started investigating the audio content analysis on broadcast news, this type of data accounts for large collections of real unconstrained conversations. Our idea was to establish a link between the content of the news, i.e. its structure intended as a sequence of topics, and the social interactions between the individuals who presented the different topics. The rationale of our approach was that individuals involved in the same topic interact more with each other than individuals involved in different topics. To this end, we aimed at identifying *social groups*, i.e. individuals characterized by a high degree of mutual interactions. Preliminary experiments showed that the groups corresponding to the most important (i.e. dominant) topics could be detected effectively (Vinciarelli, 2007). These results suggest that social interactions lead the structure of broadcast data and can facilitate the content analysis of such recordings.

Moreover, we know that individuals involved in news recordings play a specific role or function, and that those roles are governing the structure of the conversations (e.g. an anchorman conducting an interview). We thus extend our original idea that broadcast news are subdivided into social groups, assuming that the interactions between the individuals involved in the social groups are governing the structure of the recordings. We investigated a social-interactions-based approach for the automatic recognition of the roles played by broadcast-news participants. Numerical experiments revealed that around 80 percent of the data-time was correctly labeled in terms of role (Favre, 2008, 2009). This seems to suggest that there is a strong connection between role interactions and content of news data.

Another interesting challenge was to apply our approach to less structured data. In fact, broadcast news follow a predefined structure by assigning specific roles or functions to every person (such as anchorman, guest, interviewer). We hypothesize the existence of a connection between social interactions and content of small-group meetings, where individuals have a position in a given social system and do not follow stable behavioral patterns. Therefore, we investigated an approach for the automatic recognition of the roles played by the participants in the AMI meeting corpus (McCowan, 2005). The results showed that the best recognized role was the *Project Manager*, which acts as a *chairman* and thus follows predictable behavioral patterns. This suggests that the features extracted from social interactions in such data (containing spontaneous interactions) are not sufficient to analyze the structure, and that lexical content is necessary to obtain an effective content analysis (Garg, 2008).

The experiments proposed in this paper aimed at analyzing the audio content, to improve tasks performed in the multimedia content analysis community. However, the results of my work reveal that the approach we propose for the analysis of social interactions could be relevant for the affective community as well. In fact, as soon as people interact, they adopt a specific behavior depending on the social context, depending on how they percieve their interlocutor, and according to their emotions. The analysis of social interactions could thus facilitate the recognition and interpretation of human emotions.

In this paper, we describe the approach we proposed for extracting the social interactions through a set of mathematical techniques used by sociologists, namely *Social Network Analysis* (SNA) (Wasserman, 1994). We also describe how we applied machine learning techniques to the social patterns extracted from the data, aiming at classifying individuals into social groups or into roles.

The rest of this paper is organized as follows: Section 2 describes the approach we have applied to extract social groups and roles, Section 3 outlines the experimental setup and the automatic story segmentation results as well as the automatic role recognition results, and finally Section 4 draws some conclusions summarizing the main contributions of this work.

2 Our Approach

The approach we propose includes three main steps: the first is *Speaker Diarization*, which splits the audio into speaker turns, i.e. segments corresponding to single speaker intervals (see Section 2.1). The aim of this first step is to detect the persons involved in the recordings and the sequence of their interventions. The second step is *Feature Extraction*, which applies Social Affiliation Networks (see Section 2.2) (Wasserman, 1994) to represent each person in terms of their relationships with the others. The third step is the actual *Classification* step (see Section 2.3), where each person is assigned to a social group or to a role.

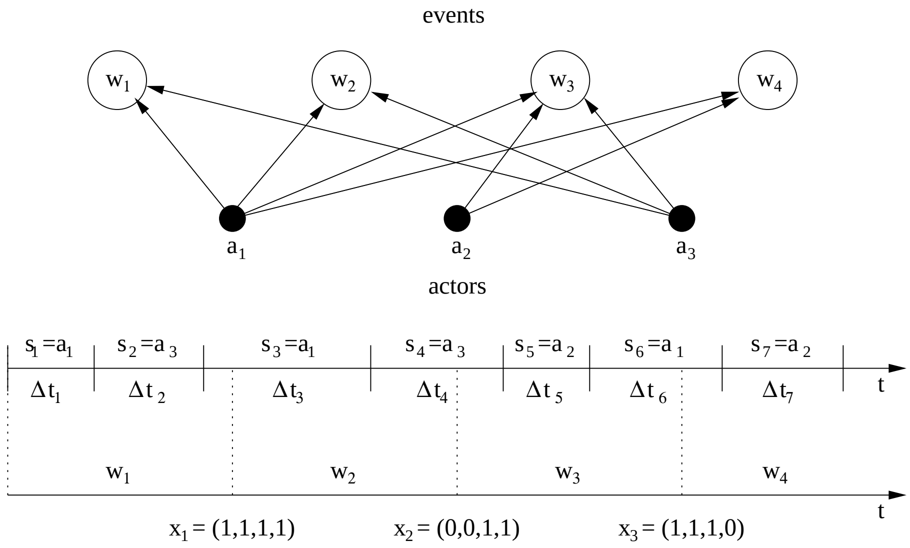

Figure 1: Social Affiliation Network extraction. The events correspond to the segments w_j and the actors are linked to the events when they talk during the corresponding segments. The actors are represented using vectors $\vec{x}_i$ where the components account for the links between actors and events.

2.1 Speaker Diarization

The aim of a speaker diarization system is to segment an audio recording into time intervals during which there is only one speaker talking. The experiments presented in this paper were performed over broadcast data and meeting recordings which requires two different approaches to cope with the characteristics of each type of data. The diarization technique for broadcast data is fully described in Ajmera (2003). A speech/non-speech detection system is used for meeting recordings and is fully described in Dines (2006). The techniques are not described here for space reasons as it is not the main element of interest of this work.

For both kinds of data, the speaker diarization process converts each recording into a sequence of turns $S = \{(s_k, \Delta t_k)\}$, where $k = 1, \ldots, N$, s_k is the speaker label corresponding to the voice detected in the k^{th} turn, and Δt_k is the duration of the k^{th} turn. The label s_k belongs to the set $A = \{a_1, \ldots, a_G\}$ of G unique speaker labels as provided by the speaker diarization process.

2.2 Feature Extraction

The turn sequence S generated by the speaker diarization system is used to build a Social Affiliation Network (SAN), a graph with two types of nodes (*actors* and *events*) where links are not allowed between nodes of the same type (see Figure 1) (Wasserman, 1994). In our experiments, the actors correspond to the speakers as detected in the diarization process, and the events correspond to D uniform non-overlapping segments spanning the whole length of the recording (see lower part of Figure 1). The rationale is that actors participating in the same events (i.e. participants speaking during the same time intervals) are likely to interact with each other. Therefore, the SAN extracts the evidence of interactions in terms of: *who talks to whom and when.*

One of the main advantages of this representation is that each actor a_i can be represented with a D-dimensional vector $\vec{x}_i$, where the component x_j accounts for the presence or the absence of each actor a_i in the different events j. The component x_j is set to 1 if the actor a_i talks during the j^{th} segment and 0 otherwise (the corresponding vectors are shown at the bottom of Figure 1). The persons that interact more with each other tend to talk during the same segments and are represented by similar vectors.

The number D of events was set through crossvalidation during the experiments.

2.3 Classification

We applied our approach to two different tasks: the first is *the story segmentation*, i.e. the segmentation of radio and television news into the different topics presented one after the other, corresponding to social groups (more details can be found in Vinciarelli (2007)). The second is *the role recognition*, i.e. the automatic recognition of the roles played by the individuals participating in broadcast news or multiparty meetings (see (Favre, 2008, 2009; Garg, 2008) for more details). The next two sections outline the machine learning techniques used to perform these two tasks.

2.3.1 The Story Segmentation Approach

This section presents an approach to make the content of a long recording more accessible: an automatic segmentation in terms of topics, i.e. *stories.*

The main idea behind the approach presented here is that people involved in the same story interact more with each other than people involved in different stories. This means that the stories can be identified by grouping the people having a high degree of mutual interaction or, in sociological terms, by detecting *social groups.*

The goal of the story segmentation is to assign the sequence of speakers talking during a conversation (represented by vectors $\vec{x}_i$ as explained in Section 2.2), a sequence of labels h_i corresponding to the stories presented one after the other during the recordings.

This corresponds to finding the sequence $H^* = (h_1, \ldots, h_M)$ which maximizes the *a-posteriori* probability:

$$H^* = \arg \max_{H \in \mathcal{H}} p(X|H)p(H) \tag{1}$$

where $\mathcal{H}$ is the set of all possible H sequences. The term $p(X|H)$ can be estimated by using a fully connected Hidden Markov Models (HMMs) (Rabiner, 1989) with S +1 states, where S is the maximum number of stories that can be observed. In fact, S states account for stories and one state accounts for the anchorman role. The emission probability function for each state is a mixture of Gaussians. The term $p(H)$ can be estimated using a tri-gram statistical language model (SLMs) (Rosenfeld, 2000).

2.3.2 The Role Recognition Approach

This section presents an approach for content analysis using the roles of the person involved in broadcast data and group meetings.

The idea of the approach is that the interactions between the roles played by the persons involved in the recordings, are governing the structure of the data.

We have considered two different approaches for the role classification: the first assigns a specific role to each speaker voice involved in the recordings using *Bayesian classifiers.* The second approach considers the sequence of speakers talking during a conversation, taking into account the dynamics of the conversation, and aligns the sequence of speakers with a sequence of roles applying *probabilistic sequential models.*

The Role Recognition Approach based on Bayesian Classifiers Section 2.2 has shown that the interaction patterns of every speaker i can be represented by a vector $\vec{x}_i$. Furthermore, every speaker i talks during a fraction τ_i of the total time of a recording. We can thus represent every speaker by a vector $\vec{y}_i = (\vec{x}_i, \tau_i)$. Consider the vector $\vec{r} = (r_1, \ldots, r_G)$, where r_i is the role of speaker i, and the vector of observation $Y = \{\vec{y}_1, \ldots, \vec{y}_G\}$, where $\vec{y}_i$ is the vector representing speaker i. The problem of assigning the role to all speakers can be thought of as the maximization of the *a-posteriori* probability $\mathrm{p}(\vec{r}\,|\,Y)$. By applying Bayes Theorem and by taking into account that $p(Y)$ is constant during recognition this problem is equivalent to finding $\vec{\hat{r}}$ such that:

$$\vec{\hat{r}} = \arg \max_{\vec{r} \in \mathcal{R}^G} \mathrm{p}(Y \,|\, \vec{r})\, \mathrm{p}(\vec{r}), \tag{2}$$

where $\mathcal{R}$ is the set of the predefined roles.

In order to simplify the problem, we make the assumption that the observations are mutually conditionally independent given the roles. In the case we are considering, it seems also reasonable to assume that the observation $\vec{y}_i$ of speaker i only depends on their role r_i and not on the roles of the other speakers. To further simplify the problem, we assume that the interaction vectors $\vec{x}_i$ and the speaking time τ_i are statistically independent given the role, and thus Equation (2) can be rewritten as:

$$\hat{\vec{r}} = \arg \max_{\vec{r} \in \mathcal{R}^G} \mathrm{p}(\vec{r}) \prod_{k=1}^{G} \mathrm{p}(\vec{x}_k \,|\, r_k)\, \mathrm{p}(\tau_k \,|\, r_k). \tag{3}$$

We estimated $\mathrm{p}(\vec{x} \,|\, r)$ using mixtures of Bernoulli distributions (Bishop, 2006). Probability $\mathrm{p}(\tau \,|\, r)$ was estimated using Gaussian distributions. The *a-priori* probability of the roles $\mathrm{p}(r)$ was estimated making the assumption that the roles are independent. We modeled $\mathrm{p}(r)$ as the fraction of speakers in the training set labeled with the role r, and thus we do not take into account the constraints that the role distribution across different participants in a given recording must respect, e.g. there is only one *Anchorman* in a talk-show, or there is only one *Project Manager* in a meeting, etc. (see Favre (2008) for more details).

Role Recognition Approach based on Probabilistic Sequential Models In this approach (see Favre (2009) for more details), we consider the sequence of speakers, taking into account the dynamics of the conversation. We have seen in Section 2.2 that the interaction patterns of every speaker i can be represented by a vector $\vec{x}_i$. Furthermore, every speaker i talks during a fraction τ_i of the total time of a recording. We can thus represent every speaker by a vector $\vec{y}_i = (\vec{x}_i, \tau_i)$. Therefore, each recording can be represented by a sequence $Y = (\vec{y}_1, \ldots, \vec{y}_N)$, where N is the number of turns detected at the speaker diarization step.

The role recognition can be thought of as finding the role sequence R^* satisfying the following equation:

$$R^* = \arg \max_{R \in \mathcal{R}^N} p(Y|R)p(R), \tag{4}$$

where $R = (r_1, \ldots, r_N)$ is a sequence of roles of length N, $r \in \mathcal{R}$ ($\mathcal{R}$ is a predefined set of roles), and $\mathcal{R}^N$ is the set of all role sequences of length N.

In our experiments, the likelihood $p(Y|R)$ is estimated with a fully connected, ergodic, Hidden Markov Models (HMMs) (Rabiner, 1989) where each state corresponds to a role $r \in \mathcal{R}$. The *a-priori* probability $p(R)$ is estimated using a 3-gram statistical language model (Rosenfeld, 2000):

$$p(R) = \prod_{k=3}^{N} p(r_k|r_{k-1}, r_{k-2}). \tag{5}$$

3 Data and Results

This section outlines the experimental setup and the results.

3.1 Story Segmentation Results

The story segmentation experiments were performed over two corpora: the first, referred to as C2 in the following, contains 27 one hour long talk-shows broadcasted by *Radio Suisse Romande* (the French speaking Swiss National broadcasting service) during February 2005. Each bulletin is managed by two anchormen that start and stop the stories by giving the floor to different participants. The average number of participants is 25. The second corpus is the largest existing database of news video, namely TRECVID (TRECVID, 2003), which consists of 229 news video of 30 minutes provided by ABC and CNN.

We used a leave-one-out approach to train HMMs models and achieved a performance in terms of purity (Ajmera, 2003) of 0.80 and 0.64 over broadcast data (C2) and the television news (TRECVID) respectively, showing that the groups corresponding to the most important stories are detected effectively (by important stories we mean the most dominant stories). These results show

Table 1: Role recognition performance for C1 and C2. The table reports both the overall accuracy and the accuracy for each role.

	overall (σ)	AM	SA	GT	IP	HR	WM
Results over C1							
C1 Bayes	82.5 (6.9)	98.0	3.6	91.8	8.0	64.6	79.9
C1 HMMs + 3-gram	79.7 (9.3)	97.8	10.3	81.4	25.7	59.5	78.0
Results over C2							
C2 Bayes	82.6 (6.9)	75.0	88.3	91.6	N/A	18.3	6.7
C2 HMMs + 3-gram	86.1 (6.8)	74.4	91.9	92.0	N/A	72.8	30.5

Table 2: Role recognition performance for C3. The table reports both the overall accuracy and the accuracy for each role.

	overall (σ)	PM	ID	ME	UI
C3 Bayes	43.5 (23.9)	75.3	15.1	15.1	40.0
C3 HMMs + 3-gram	46.9 (24.9)	61.8	25.0	39.4	33.8

that there is a link between the social interactions and the content of news data. The proposed approach enables to extract social groups that can be used to analyze the content of the data.

3.2 Role Recognition Results

The experiments on the role recognition task are performed over three different corpora. The first, referred to as C1 in the following, contains 96 news bulletins broadcasted by *Radio Suisse Romande* (the French speaking Swiss National broadcasting service) during February 2005. The average length of C1 recordings is 11 minutes and 50 seconds, and the average number of participants is 12. The second corpus is C2 (see Section 3.1). The third corpus, referred to as C3 in the following, is the AMI meeting corpus (McCowan, 2005), a collection of 138 meeting recordings involving 4 persons each and with an average length of 19 minutes and 50 seconds.

The roles of C1 are *Anchorman* (AM), *Second Anchorman* (SA), *Guest* (GT), *Interview Participant* (IP), *Headline Person* (HP), and *Weather Man* (WM). Roles with the same name are played in C2 (with the exception of IP that appears only in C1), but they correspond to different functions (e.g., AM are not expected to deliver the news but to entertain in talk-shows). The roles of C3 are *Project Manager* (PM), *Marketing Expert* (ME), *User Interface Expert* (UI), and *Industrial Designer* (ID).

Table 1 reports the results achieved on C1 and C2, and in Table 2, those obtained on C3. The performance is measured in terms of *accuracy*, intended as the percentage of time correctly labeled in terms of roles in the test set. We used a leave-one-out approach to train our models and to select the number D of segments used to split the recordings (see Section 2.2). For each corpus, the first line reports the results achieved with the approach based on a Bayesian classifier (Bayes) (see Section 2.3.2), and the last row reports the results when using probabilistic sequential models (HMMs + 3-gram) (see Section 2.3.2). Each accuracy value is accompanied by the standard deviation of the accuracies achieved over the different recordings of each corpus.

The results show that the overall role recognition accuracy is above 80 percent for both C1 and C2. This highlights the strong connection between the interactions of the speakers playing roles and the content of the recordings in broadcast data. However, in meeting recordings, the roles are recognized with a lower accuracy. The explanation is that the roles in meetings are *informal*, i.e. they correspond to a position in a given social system and do not correspond to stable behavioral patterns like in the case of the *formal* roles in broadcast data. In meetings, the only role recognized

C1	HMM C	HMM W
Bayes C	77.1	2.7
Bayes W	5.4	14.8
C2	HMM C	HMM W
Bayes C	81.1	5.0
Bayes W	1.5	12.4
C3	HMM C	HMM W
Bayes C	27.1	14.0
Bayes W	11.1	47.7

Table 3: Diversity assessment. The table reports the percentage of data-time where the Bayes and HMM based role recognition approach are both correct (C), both wrong (W), or one wrong and the other correct.

approach	overall	PM	ID	ME	UI
SNA	43.1	75.7	13.4	16.4	41.2
lexical	67.1	78.3	53.0	71.9	38.1
SNA+lexical	67.9	84.0	50.1	69.8	38.1

Table 4: Role recognition results for C3. The table reports both the overall accuracy and the accuracy for each role.

with a high accuracy is the *Project Manager* (PM). The reason is because the PM also acts as a *chairman*, playing thus a more formal role than the domain experts ID, ME, and UI.

The comparison between the approach based on HMMs and the one based on Bayesian classifiers results in a significant degree of *diversity* (see Table 3). The probabilistic sequential approach results in an improved recognition of less frequent roles, which are typically penalized by Bayesian classifiers because of their low *a-priori* probability. The combination of the two approaches could thus lead into significant performance improvements, and will be the subject of future work.

The limitation of our approach is represented by the low values of accuracy obtained on the meeting recordings. This suggests that the social interaction based role recognition approach is not well suited for unconstrained data characterized by spontaneous interactions (such as multiparty meetings). Table 4 shows the role recognition accuracy for the C3 corpus combining a SNA and lexical based role recognizer. The first line reports the accuracies obtained by using only SNA, the second line those obtained using only the lexical approach, and the last line those obtained using the combination of the two. The lexical approach appears to be a more reliable cue for the recognition of the roles in such meeting recordings (more details can be found in (Garg, 2008)).

4 Conclusion and Contributions

This paper has presented automatic approaches aiming at analyzing the human interactions appearing in multiparty data in order to indexing multimedia content. The idea developed in this paper is that a strong connection can be established between the content of multiparty material and social interactions.

We demonstrated that this assumption is relevant in the case of data characterised by structured human interactions, such as broadcast news. We investigated a story segmentation task where the social interactions allow to index the content into social groups corresponding to the topics presented in the news. Moreover, we performed a role recognition task where the social interactions allow to assign a role to each individual, and thus to structure the content of the recordings. In the case of more spontaneous interactions (like meetings), the experiments show that our proposed approach is not sufficient and that lexical content analysis is necessary.

This paper suggests that the analysis of social interactions in combination with other behavioral cues extracted from audio (e.g. prosodic and lexical features) and video (e.g. gestures, visual focus of attention) could lead to useful contributions in the domains of multimedia content analysis and affective computing. In fact, the analysis of social groups could be further developed into discriminating between *positive* vs *negative* stories, or *sad* vs *happy* stories.

Moreover, following Kelly (2001), we can hypothesize that the social groups are characterized by a composition of the moods, emotions, and sentiments brought by group members. The analysis of social groups, combined with lexical and prosodic features, could help in understanding group affect, e.g. *Can we define group mood?*, or *Do groups share emotions as do individuals?*.

Acknowledgements

This work is supported by the Swiss National Science Foundation (under the National Centre of Competence in Research on Interactive Multimodal Information Management), and by the European Community's Seventh Framework Programme (FP7/2007-2013), under grant agreement no. 231287 (SSPNet). The author wishes to thank Alessandro Vinciarelli and Alfred Dielmann.

References

Ajmera, J. and Wooters, C. (2003). A robust speaker clustering algorithm. In *Proceedings of IEEE Workshop on Automatic Speech Recognition Understanding*, pages 411–416.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning.* Springer Verlag.

Dines, J. and Vepa, J. and Hain, T. (2006). The segmentation of multi-channel meeting recordings for automatic speech recognition. In *Proceedings of Interspeech*, pages 1213–1216.

Favre, S. and Salamin, H. and Dines, J. and Vinciarelli, A. (2008). Role Recognition in Multiparty Recordings using Social Affiliation Networks and Discrete Distributions. In *Proceedings of the ACM International Conference on Multimodal Interfaces*, pages 29–36.

Favre, S. and Dielmann, A. and Vinciarelli, A. (2009). Automatic Role Recognition in Multiparty Recordings Using Social Networks and Probabilistic Sequential Models. To Appear in *Proceedings of the 2009 ACM International Conference on Multimodal Interfaces.*

Garg, N. and Favre, S and Salamin H. and Hakkani-Tur, D. and Vinciarelli, A. (2008). Role Recognition for Meeting Participants: an Approach Based on Lexical Information and Social Network Analysis. In *Proceedings of the ACM International Conference on Multimedia*, pages 693–696.

Kelly, Janice R and Barsade, Sigal G. (2001). Mood and Emotions in Small Groups and Work Teams. In *Organizational Behavior and Human Decision Processes*, vol.86, no.1, pages 99–130.

Kunda, Z. (1999). Social Cognition. MIT Press.

McCowan, I. and al. (2005). The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, page 4.

Rabiner, L.R. (1989). A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE*, vol.77, pages 257–286.

Rosenfeld, R. (2000). Two decades of statistical language modeling: Where do we go from here? In *Proceedings of the IEEE*, vol.88, no.8, pages 1270–1278.

Vinciarelli, A. and Favre, S. (2007). Broadcast News Story Segmentation Using Social Network Analysis and Hidden Markov Models. In *Proceedings of ACM International Conference on Multimedia*, pages 261–264.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis.* Cambridge University Press.

TREC Video retrieval Evaluation (2003). http://www-nlpir.nist.gov/projects/tv2003/tv2003.html.

Emotive and Personality Parameters in Multimedia Recommender Systems

Marko Tkalčič, Jurij Tasič and Andrej Košir
University of Ljubljana Faculty of Electrical Engineering
Tržaška 25, Ljubljana, Slovenia
{marko.tkalcic|jurij.tasic|andrej.kosir}@fe.uni-lj.si

Abstract

The PhD research presented in this paper discusses the inclusion of affective computing in multimedia recommender systems. Our objective is to fill the missing gaps of the scenario where the system detects the emotive response of a user watching a multimedia item. The emotive state is then used to model items and users in the recommender system which filters only relevant items for each specific user. The thesis is composed of four original scientific contributions: (i) a module for emotion detection from video sequences of user faces into the valence-arousal-dominance (VAD) space, (ii) a content-based recommender system with affective user and item modelling, (iii) a collaborative filtering recommender algorithm with a personality-based user similarity measure and (iv) a database that forms the basis for the first three contributions. The paper gives results for contribution (ii) and preliminary results for contribution (iii). Contribution (iv) is also presented while a tentative future plan for contribution (i) is given.

Keywords: content-based recommender system, collaborative recommender system, emotive response, emotion detection, big five personality model

1 Introduction

The presented PhD research addresses affective computing in multimedia recommender systems (MRS). Recommender systems are getting more and more attention as the amount of available multimedia content is growing. The goal of MRS is to make a narrow selection of multimedia content that is relevant for each specific user based on her/his preferences. The user is thus more satisfied with the system. Although this is a task that should intuitively follow the end user's affective responses and personality, state of the art MRS mostly ignore the affective approach. Most of today's MRS create the user profiles (a data structure that contains the knowledge about a user) based on past user's actions. This user feedback can be explicit (e.g. ratings) or implicit (observing the user's behaviour in a non-intrusive manner). In laboratory experiments we usually rely on explicit feedback which is more accurate but in real life applications it is undesired because it is too intrusive and tends to turn users away. Thus an implicit feedback collection technique is preferable.

1.1 Problem Statement and Proposed Solution

Except for few research contributions (González et al., 2004; Lekakos and Giaglis, 2006; Shan et al., 2009), MRS do not exploit the affective aspect in human computer interaction (HCI). The two most popular MRS approaches, content-based recommenders (CBR) and collaborative filtering recommenders (CF), both use very technical information to perform the prediction of how a user would rate (and like) an item.

We believe that bringing emotions and personality in the field of user modelling for multimedia recommender systems would improve their performance. We propose a user scenario (inspired by video-on-demand like services) where end users consume multimedia content, their affective responses are detected in a non intrusive fashion and used to build the user and item profiles for the CBR system. As an alternative to the CBR we propose to have a CF recommender where the user similarity measure is calculated based on end users' personalities. The user is then offered a narrow selection of relevant items instead of browsing the whole database. Figure 1 shows the proposed scenario. We want to get closer to the system predicted by Picard (2000): *...that when the machine presents you with something you like, it sees that you like it. And when it does something you do not like, it sees that too [page 101].*

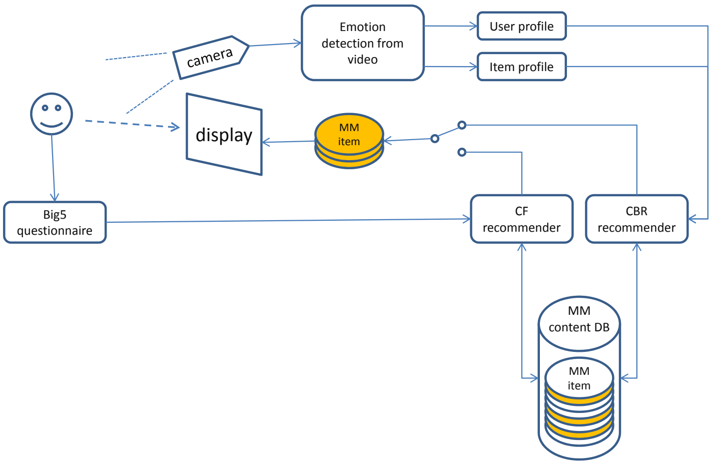

Figure 1: The recommender system scenario: while the user is consuming an item (movie, picture etc.) the system detects her/his emotive response in a non-intrusive way. This information is used to build the profiles for the CBR system which retrieves only the items relevant for that user from the database. As an alternative to the CBR, a CF recommender that exploits the users' personalities can be used to perform the filtering of relevant items.

1.2 Objectives and Scientific Contributions

Some elements of such a system already do exist but are being used for other purposes and need to be properly adapted while some others are missing. Our objective is to fill these gaps. Thus the original scientific contributions of this PhD research will be the following

Emotion detection from video A non intrusive emotion detection technique that maps video sequences of users' faces into emotive states in the valence-arousal-dominance (VAD) emotive space.

Affective item and user profiles for CBR systems A CBR that uses information about the user's emotive response during the content consumption as metadata for modelling both items and users (instead of using genre based metadata).

Personality based similarity measure for CF systems A CF recommender that calculates the user similarity using a personality based measure (instead of using rating based measures).

Dataset of user interaction with affective metadata and explicit ratings A database of user interaction with a multimedia consumption system that contains explicit ratings, affective responses in the VAD space, users' big five personality values and video sequences of the users while consuming the items.

By fulfilling our goals we wish to bring the affective computing approach in the area of multimedia recommender systems where we believe it should be present because the entertainment business is primarily about consumer's emotions. This research work addresses several standards, in particular the W3C EmotionML (see Schröder et al. (2008)) and MPEG7 Usage History (see Manjunath et al. (2002)).

1.3 Related Work

A lot of work has been done in the field of detecting emotions from various modalities, see Donato et al. (1999); Chibelushi and Bourel (2003); Fasel and Luettin (2003); Picard and Daily (2005); Zeng et al. (2009) for excellent overviews. All of these methods yield one of the basic emotions as output.

Overviews of recommender systems are given by Adomavicius and Tuzhilin (2005); Burke (2002); Lew et al. (2006). The basic taxonomy is given (content-based, collaborative and hybrid) and the pros and cons are described. An example of a CBR system is the work done by Pogačnik et al. (2005) while Kunaver et al. (2007) have been working on CF and hybrid methods.

There have been attempts at introducing affective modelling of end users in recommender systems. Lekakos and Giaglis (2006) took the user's lifestyle as the central parameter of their collaborative recommender. González et al. (2004) introduced the Smart User Model (SUM) which is a data structure composed of objective attributes, subjective attributes and psychological traits.

Several authors tried to annotate multimedia content with information about the emotive state an item induces in end users. Hanjalic (2006) extracted mood from video sequences. Lang et al. (2005) performed a large scale experiment with a big image database and several subjects who manually annotated each image with their emotive response which yielded the IAPS database.

There are several taxonomies for the description of emotions (Cowie et al., 2001; Posner et al., 2005; Villon and Lisetti, 2006). One line follows the work started by Charles Darwin and is called the *basic emotions theory* (Ekman and Friesen, 2003). The other popular description of emotions in HCI is the three dimensional space valence - arousal - dominance (VAD), where each emotive state is represented by a triple of numerical values that describe a certain qualitative aspect of the emotion. The *circumplex model of emotions* maps the basic emotions into the VAD space (Posner et al., 2005; Villon and Lisetti, 2006).

2 Methodology

In this section we present the methodology for each of the four scientific contributions. We provide the hypotheses, describe the plan of work and/or work already done, give results where available and provide a list of open issues for each contribution.

The basis of all the contributions is the database of users consuming digital items (images) which is described in more details in Sec. 2.4. We collected the big five personality values for all users, tracked their explicit ratings for each image and recorded the emotive responses of their faces with a video camera. Each image was annotated with a genre attribute and a six-tuple of the induced emotive state.

2.1 Emotion Detection from Video Sequences

Our hypothesis is that it is possible to detect the emotive responses of end users from video sequences of their faces with a certain success rate. The target space of emotion description is the VAD space.

At the time of writing we have not started yet with the experimental part of the emotion detection from video. However we performed a preliminary literature review (Wang and Guan, 2008; Kim and André, 2008; Zeng et al., 2009; Donato et al., 1999; Chibelushi and Bourel, 2003; Fasel and Luettin, 2003) which led us to the following methodological plan:

1. Face detection in all video sequences using the Viola-Jones algorithm (Viola and Jones, 2004) implemented in OpenCV

2. Registration of faces to improve the Viola-Jones detection results. We will start with the template matching technique (Matlab)
3. Feature extraction: based on literature review we will first try with Gabor filtering on each frame and use the Hidden Markov model HMM to produce length invariant features (Matlab)
4. Merge the features of video sequences and emotion labels (VAD values) into the dataset
5. Evaluate various machine learning algorithms with the dataset. We will use the 10 fold cross validation scheme to build the training and test sets. (Matlab, Weka)
6. Perform statistical tests to assess the results yielded

2.1.1 Open Issues

As we have not started yet with emotion detection there are several open issues and many more that we are currently not aware of: (i) choice of features: which are the best features taking into account accuracy and speed of emotion detection, (ii) choice of the machine learning algorithm: which is the most suitable classifier, (iii) determination of the acceptable success rate and (iv) determination of the classifier's output: nominal or numerical classes.

2.2 Content Based Recommender with Affective User Modelling

The hypothesis of this contribution is that CBR algorithms with affective based metadata perform better than CBR with standard genre based metadata.

In order to accept (or reject) the above hypothesis we constructed two metadata sets for the item profiles: an affective and a standard (non affective) and compared the performance of a CBR using both metadata sets for item/user profiles.

Each multimedia content item induces an emotive response in end users. This response is described with the values v, a and d in the VAD space. The first two statistical moments of the responses of a set of users that have watched the same item h formed the affective metadata set $\mathcal{V}$ of the item h:

$$\mathcal{V} = (\bar{v}, \sigma_v, \bar{a}, \sigma_a, \bar{d}, \sigma_d) \tag{1}$$

which is a six-tuple. The standard metadata set A was composed of the genre and the average watching time for the item h. The genre was a value from a set of ten genres and the average watching time was calculated from different users for the observed item h.

2.2.1 Experiment and Results

We performed the prediction of item relevancy with machine learning algorithms. We evaluated the AdaBoost, NaiveBayes, C4.5 and Support Vector Machine (SVM) classifiers. We trained the classifiers with two metadata sets: A (affective) and $A \times \mathcal{V}$ (standard and affective) . We used the ten fold cross validation for splitting the dataset into the training and test sets. From the confusion matrices we calculated the precision P, recall R and F-measure as defined by Herlocker et al. (2004). We also performed the Pearson χ^2 significance test comparing the confusion matrices yielded by the metadata sets A and $A \times \mathcal{V}$.

The results of the CBR recommender with affective user modelling have been submitted for publication to a journal (Tkalčič et al., 2009). In table 1 we present the performance results of the different metadata sets and classifiers. The Pearson χ^2 significance test showed that the difference was significant in all four classifiers.

2.2.2 Open Issues

The results showed that modelling users based on their preferences for emotive responses gives better results than modelling with genre based parameters. But we are still unsure whether the approach taken could be improved by using the emotive response parameters in a different way.

Thus the open issues regarding this topic are: (i) is the *VAD* space better than the *basic emotions* space for MRS user modelling and (ii) in which way could the described approach be improved (multivariate analysis).

metadata set	classifier	P	R	F
A	AdaBoost	0,57	0,42	0,48
	C4.5	0,60	0,46	0,52
	NaiveBayes	0,58	0,58	0,58
	SVM	0,61	0,55	0,58
$A \times \mathcal{V}$	AdaBoost	0,63	0,56	0,59
	C4.5	0,64	0,57	0,60
	NaiveBayes	0,57	0,64	0,61
	SVM	0,65	0,61	0,63

Table 1: Precision, recall and F measure for the two metadata sets and four classifiers.

2.3 Collaborative Recommender with Personality-based User Similarity Measures

The hypothesis is that the usage of personality based user similarity measures yields statistically equivalent or significantly better results than the usage of rating based similarity measures in terms of the recommender system's performance (precision, recall, F-measure).

CF recommenders are based on the presumption that when the similarity between two users $sim(u_1, u_2)$ is high both users will give similar ratings to the item h. The user similarity measure is thus a crucial part of any CF system. The similarity measures used in state-of-the art CF systems are rating based. The main drawback is that it needs to be calculated on a regularly basis as new ratings are added to the system and the computational complexity is high.

We propose to use similarity measures whose calculations are fast and do not need to be recalculated on a regular basis. We evaluated the usage of a set of new similarity measures for a CF system based on the big five personality model, which describes the personality of a single user u by giving numerical values to five dimensions: openness, conscientiousness, extraversion, agreeableness, and neuroticism (Goldberg et al., 2006). We denote these with the vector $\vec{b} = (b_1, \ldots, b_5)$. The assessment of these values for a specific person is usually done through a questionnaire like the one proposed by Goldberg et al. (2006).

We evaluated two approaches for the calculation of distances between users u_1 and u_2 (which are inversely related to the similarity $sim(u_1, u_2)$) based on the users' personality vectors $\vec{b}_1$ and $\vec{b}_2$:

1. Euclidian distance: $d_E(\vec{b}_1, \vec{b}_2)^2 = \sum_i |b_{1i} - b_{2i}|^2$
2. Weighted Euclidian distance: $d_{WE}(\vec{b}_1, \vec{b}_2)^2 = \sum_i w_i^2 |b_{1i} - b_{2i}|^2$ where the weights w_i were the coefficients of the first vector yielded by the PCA

2.3.1 Experiment and Preliminary Results

We applied the acquired dataset (see Sec. 2.4) in the CF recommender system developed by Kunaver et al. (2007). Each similarity measure was evaluated two times: once with a higher weight on close neighbours and once with a higher weight on all users (public). The reference rating-based similarity measure (implemented by Kunaver et al. (2007); with which we compared the personality based measures) and the two big five measures thus yielded six different combinations of similarity measures to calculate and evaluate.

We used the precision P, recall R and F-measure as performance measures. We also performed a one-way analysis of variance to determine whether the differences of mean values F of all six measures were statistically significant.

In terms of mean values of P, R and F the big five based approaches performed better than the ratings based approaches. The mean values of P, R and F are reported in Tab. 2. The ANOVA analysis (with the significance level $\alpha = 0.05$) of the F measure further showed that all the big five based approaches performed significantly better than the distance based measure with

	P	R	F
distances - neighbours	0.6666	0.5895	0.6268
distances - public	0.7042	0.7401	0.7232
big5 neighbours	0.6309	0.8533	0.7062
big5 public	0.7093	0.8068	0.7442
weighted big5 neighbours	0.6455	0.8398	0.7165
weighted big5 public	0.7104	0.8064	0.7450

Table 2: Mean values of P, R and F for the different combinations of measures and weighting.

higher weight on the neighbours and were statistically equivalent to the distance based measure with higher weight on the public. The box plot of the F-measure values is shown in Fig. 2.

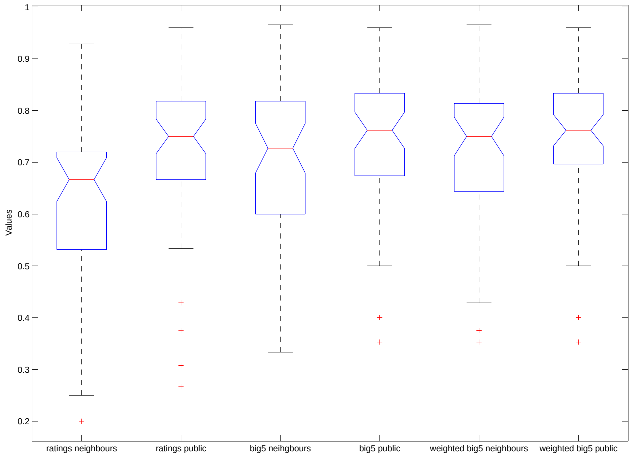

Figure 2: Box plots of the F measure results for six variants of measures.

2.3.2 Open Issues

The personality based measures seem to produce quality neighbours, which lead the CF recommender to results equivalent to the rating-based measures. However we identified three important issues pending: (i) in order to get the big five values end users must first fill in a questionnaire which is annoying. A better approach would be a solution similar to the one proposed by Berger et al. (2007) who found significant dependencies between photos and tourist types. Based on this they developed a game where end users drag a set of photos into a *like-it* bin and a *don't like-it* bin and the system classifies the user in a category of the kick-pack plane used in tourist profiling. We plan to investigate the relation between the ratings of specific images from our database and the personality values in order to develop a funny automatic personality detector. The second issue is (ii) the absence of context awareness (a user can have different neighbours in different situations, e.g. watching a movie with her/his friends vs. watching a movie with her/his kids) in the calculation of neighbours and the third is (iii) that we need to evaluate more similarity measures to give the results heavier foundations.

2.4 Database

Although the database has been the first methodological step that was performed we put it as the last subsection because it is easier to understand the requirements for the database at this point.

We needed a dataset in the form of a history log of interactions of users with a device displaying multimedia content. The following attributes were needed in the dataset: user's big five personality values, explicit ratings and induced emotive states.

We performed an emotion induction experiment. We chose 70 images from the IAPS database (Lang et al., 2005) with apriori known VAD values of the induced emotions. We had 52 users (15 males and 37 females) aged between 17 and 20 taking part in the experiment. Each user was shown a sequence of images and was requested to give an explicit rating from a five point Likert scale to each image. The users were recorded during the consumption of the images with a web camera. Furthermore each user filled in a questionnaire from the IPIP pool (http://ipip.ori.org/, 2009) in order to asses her/his big five personality values.

2.4.1 Open Issues

A further statistical analysis of the dataset is needed. In the future we plan to extend the database with: (i) more users, (ii) more items and (iii) contextual information. We plan to perform the experiment with each user repeating it in different social contexts (alone, single gender group, mixed gender group etc.). Furthermore, (iv) the ground truth quality of the dataset could be questionable.

3 Conclusion

We presented four scientific contributions for the application of affective recommender systems for multimedia items. We provided the underlying database acquisition and basic statistics, experimental results for the CBR and CF recommenders and a workplan for the emotion detection part.

Acknowledgement

We would like to thank the members of the LDOS group for their assistance in the dataset acquisition procedure as well as the students who participated in the experiment. A special thank goes to our colleagues Tomaž Požrl and Matevž Kunaver.

References

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749.

Berger, H., Denk, M., Dittenbach, M., Pesenhofer, A., and Merkl, D. (2007). Photo-based user profiling for tourism recommender systems. In Psaila, G. and RolandWagner, editors, *E-Commerce and Web Technologies*, volume 4655/2007, pages 46–55. Springer Berlin / Heidelberg.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370.

Chibelushi, C. and Bourel, F. (2003). Facial expression recognition: A brief tutorial overview. *CVonline: On-Line Compendium of Computer Vision*, 9.

Cowie, R., Cowie, E. D., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80.

Donato, G., Bartlett, M. S., Hager, J. C., and Sejnowski, P. E. T. J. (1999). Classifying facial actions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(10):974–989.

Ekman, P. and Friesen, W. V. (2003). *Unmasking the face.* Malor Books, Cambridge MA 02238.

Fasel, B. and Luettin, J. (2003). Automatic facial expression analysis: A survey. *Pattern Recognition*, 35(2):259–275.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., and Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40:84–96.

González, G., López, B., and J. LL, d. l. R. (2004). Managing emotions in smart user models for recommender systems. In *Proceedings of 6th International Conference on Enterprise Information Systems ICEIS 2004*, volume 5, pages 187–194.

Hanjalic, A. (2006). Extracting moods from pictures and sounds. *IEEE Signal Processing Magazine*.

Herlocker, J. L., Konnstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactionson Information Systems*, 22(1).

http://ipip.ori.org/ (last accessed: April 2009). International personality item pool: A scientific collaboratory for the development of advanced measures of personality and other individual differences. web site.

Kim, J. and André, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2067–2083.

Kunaver, M., Požrl, T., Pogačnik, M., and Tasič, J. (2007). Optimisation of combined collaborative recommended systems. *International Journal of Electronic Communications*, 61:433–443.

Lang, P., Bradley., M., and Cuthbert, B. (2005). International affective picture system (iaps): Affective ratings of pictures and instruction manual. technical report a-6. Technical report, University of Florida, Gainesville, FL.

Lekakos, G. and Giaglis, G. M. (2006). Improving the prediction accuracy of recommendation algorithms: Approaches anchored on human factors. *Interacting with computers*, 18:410–431.

Lew, M. S., Sebe, N., Djeraba, C., and Jain, R. (2006). Content-based multimedia information rtrieval: State of the art and challenges. *ACM Transactions on multimedia computing*, 2(1):1–19.

Manjunath, B., Salembier, P., and Sikora, T., editors (2002). *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley.

Picard, R. and Daily, S. B. (2005). Evaluating affective interactions: Alternatives to asking what users feel. In *CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches*, Portland, OR.

Picard, R. W. (2000). *Affective Computing*. MIT Press.

Pogačnik, M., Tasič, J., Meža, M., and Košir, A. (2005). Personal content recommender based on a hierarchical user model for the selection of tv programmes. *User Modeling and User Adapted Interaction*, 15:425–457.

Posner, J., Russell, J. A., and Peterson, B. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17:715–734.

Schröder, M., Baggia, P., Burkhardt, F., Martin, J.-C., Pelachaud, C., Peter, C., Schuller, B., Wilson, I., and Zovato, E. (2008). Elements of an emotionml 1.0. W3C Incubator Group Report 20 November 2008.

Shan, M.-K., Kuo, F.-F., Chiang, M.-F., and Lee, S.-Y. (2009). Emotion-based music recommendation by affinity discovery from film music. *Expert Syst. Appl.*, 36(4):7666–7674.

Tkalčič, M., Tasič, J., and Košir, A. (2009). Usage of affective parameters in a content-based multimedia recommender system. *Submitted to the Elsevier International Journal of Human-Computer Studies*.

Villon, O. and Lisetti, C. (2006). A user-modeling approach to build user's psycho-physiological maps of emotions using bio-sensors. In *The 15th IEEE International Symposium on Robot and Human Interactive Communication, 2006. ROMAN 2006.*, pages 269–276.

Viola, P. and Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

Wang, Y. and Guan, L. (2008). Recognizing human emotional state from audiovisual signals. *IEEE Transactionson multimedia*, 10(5):936–946.

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58.

A Unified Features Approach to Human Face Image Analysis and Interpretation

Zahid Riaz, Suat Gedikli, Micheal Beetz and Bernd Radig
Department of Informatics,
Technische Universität München
85748 Garching, Germany
{riaz|gedikli|beetz|radig}@in.tum.de

Abstract

Human face image analysis has been one of the challenging fields over the last few years. Currently, many commercially available systems can interpret face images in an efficient way but are generally limited to only one specific application domain. On the other hand, cameras are becoming a useful tool in human life and are the vital constituent of most of the interactive systems. In this regard, we present a technique to develop a unified set of features extracted from a 3D face model. These features are successfully used for higher level facial image interpretation in different application domains. We extract the feature once and use the same set of features for three different applications: face recognition, facial expressions recognition and gender recognition and obtain a good accuracy level while preserving the generality and efficiency of our feature extraction technique. The proposed technique is easy to implement and real time.

Keywords: Features extraction, Model based face image analysis, Facial behaviors analysis, Human robot interaction (HRI)

1 Introduction

By the recent advancement in the field of robotic technology, future systems would be mainly relying on intelligent machines interacting with the humans in daily life environment. A good example is the assistive robots (1) where robots serve as an attendee or nurse to the elderly people or handicaps. On the other side, cameras are the integral part of the most of the daily life technical systems, for example mobiles, notebooks, ATM machines and access control applications. This role of cameras and quality can further improve future systems by embedding higher level image interpretation. Currently available cameras have face detection capabilities, automatic smile capture and laptops with face recognition systems.

In human robot interaction (HRI) faces play an important role and are the natural way to interact. Human can predict identity, gender, facial expression, behavior, ethnical background and can estimate age almost on very first glance. Whereas the current assistive robots are struggling not only to be structurally humanoid but also to exhibit humanly behavior. For instance, in face recognition applications currently available systems work well under constrained environments however still require improvements. The efficiency and performance is still a trade off in unconstrained systems. Further extracting all possible information from the face images at the same time is quite difficult and yet not fully described on a common platform by the research community. In this paper we focus on the development of such a system which can perform these capabilities at the same time and hence give sufficient time to the robot to continue its job for other interaction activities. We focus on a model based approach to extract multi-features from human face images. These features can be used for face recognition, facial expressions, facial behaviors and gender classification. Further this system is trained to work under different variations like poses and illumination.

In this regard, an intuitive approach is to borrow the concepts of human-human interaction and train the machines for comparable performance in real world situations. Human faces play an important role in daily life interaction. Therefore, learning identity, expressions, gender, age and facial behavior assures better interaction and utilizes improved context information. For example, a system aware of person identity information can better employ user's habits and store current interaction knowledge for improving future interactions. In the recent decade model based image analysis of human faces has become a challenging field due to its capability to deal with the real world scenarios. Further it outperforms the previous techniques which were constrained to user's intervention with the system either to manually interact with system or to be frontal to the camera. Currently available model based techniques are trying to deal with some of the future challenges like developing state-of-the-art algorithms, improving efficiency, fully automated system development and versatility in different applications. In this paper we deal especially with versatility and diversity of the problem. Our aim is to extract features which are fully automatic and versatile enough for different applications. These capabilities of the system suggest to apply it in interactive scenarios like human machine interaction, token less access control, facial analysis for person behavior and person security. Models take benefit of the prior knowledge of the object shape and hence try to match themselves only with the object in an image for which they are designed. Face models impose knowledge about human faces and reduce high dimensional image data to a small number of expressive model parameters. We integrate the three-dimensional Candide-III face model (10) that has been specifically designed for observing facial features variations defined by facial action coding system (FACS) (16). The model parameters together with extracted texture and motion information are utilized to train classifiers that determine person-specific information. Our feature vector for each image consists of structural, textural and temporal variations of the faces in the image sequence. Shape and textural parameters define active appearance models (AAM) in partial 3D space with shape parameters extracted from 3D landmarks and texture from 2D image. Temporal features are extracted using optical flow. These extracted features are more informative than conventional AAM parameters since we consider local motion patterns in the image sequences in the form temporal parameters.

The remainder of this paper is divided in four sections. Section 2 explains the related research work for model based approaches, including 2D and 3D models. Section 3 describes the problem statement in detail followed by the proposed solution in section 4. Section 5 describes feature extraction technique in our case along with the future extensions. These features are used in section 6 for experimentation. In the last section, we describes the analysis of our technique along with the future extensions of this research work.

2 Related work

The proposed system consists of different modules which work adjacently to each others. We describe related work regarding to each module independently and explain if there exists any dependency among these modules. The overall system consists of face detection and localization for model fitting and structural features extraction, textural features extraction, pose compensation, illuminations compensations, temporal features extraction and finally synthesizing the image for final features extraction. The feature set is then applied different types of applications. Our approach consists of different parallel and sequentially working processes.

Human face image analysis for higher level information extraction is generally performed using appearance based, templates based, model based or hybrid methods based approaches (13). Appearance based analysis comprise of holistic approaches which exploit subspace projection. The extracted features are analyzed at a global level and local facial features are not considered in this regard. These methods mainly involve principal component analysis (PCA), linear discriminant analysis (LDA) and independent component analysis (ICA). The performance of these methods in general, degrades as more variations are considered in the face images like misalignment of local features, pose and lighting variations. In order to overcome these discrepancies, methods of image alignment have been devised by the researchers (13). Several face models and fitting approaches have been presented in the recent years. Cootes et al. (8) introduced modeling face shapes with Active Contours. Further enhancements included the idea of expanding shape models with texture information (9). In contrast, three-dimensional shape models such as the Candide-3 face model consider the real-world face structure rather than the appearance in the image. Blanz et al. propose a face model that considers both, the three-dimensional structure as well as its texture (2). However, model parameters that describe the current image content need to be determined in order to extract high-level information, a process known as model fitting. In order to fit a model to an image. Van Ginneken et al. learned local objective functions from annotated training images (3). In this work, image features are obtained by approximating the pixel values in a region around a pixel of interest. The learning algorithm which uses to map images features to objective values is a k-Nearest-Neighbor classifier (kNN) learned from the data. We use similar methodology developed by Wimmer et al. (7) which combines multitude of qualitatively different features (22), determines the most relevant features using machine learning and learns objective functions from annotated images (3). To extract descriptive features from the image, Michel et al. (17) extracted the location of 22 feature points within the face and determine their motion between an image that shows the neutral state of the face and an image that represents a facial expression. The very similar approach of Cohn et al. (18) uses hierarchical optical flow in order to determine the motion of 30 feature points. A set of training data formed from the extracted features is utilized to learn on a classifier. For facial expressions, some approaches infer the expressions from rules stated by Ekman and Friesen (16). This approach is applied by Kotsia et al. (19) to design Support Vector Machines (SVM) for classification. Michel et al. (17) train a Support Vector Machine (SVM) that determines the visible facial expression within the video sequences of the Cohn-Kanade Facial Expression Database by comparing the first frame with the neutral expression to the last frame with the peak expression. In order to perform face recognition applications many researchers have applied model based approaches. Edwards et al (5) use weighted distance classifier called Mahalanobis distance measure for AAM parameters. However, they isolate the sources of variation by maximizing the inter class variations using Linear Discriminant Analysis (LDA), a holistic approach which was used for Fisherfaces representation (6). Riaz et al (20) apply similar features for explaining face recognition using Bayesian networks. However results are limited to face recognition application only. They used expression invariant technique for face recognition, which is also used in 3D scenarios by Bronstein et al (11) without 3D reconstruction of the faces and using geodesic distance. Park et. al. (12) apply 3D model for face recognition on videos from CMU Face in Action (FIA) database. They reconstruct a 3D model acquiring views from 2D model fitting to the images.

3 Problem Statement

In the recent decade, model based approaches have attained a huge attention of the research community owing to their compactness and detailed description over the other techniques. Models described a large size image in small set of parameters. These few descriptors are called model parameters. Further they narrow the search domain in an image and precisely look for the object of interest for which they are designed. This reduces false alarms in finding an object. The models used in face image analysis are active shape models (ASM), active appearance models (AAM), deformable models, wireframe models, and 3D morphable models (21). We address the problem in which a robotic system is able to extract a common feature set automatically from face images and capable to classify gender, person identity and facial behavior. In such applications an automatic and efficient feature extraction technique is necessary to be developed which can interpret every possible face information. Currently available systems lack this property. A major reason is that researchers focus on isolating the sources of variations while focusing on a particular application. For example, in face recognition application, many researchers normalize face in order to remove facial expressions variations to improve face recognition results. So the extracted features do not contain facial expressions information. We address an idea to develop a unified feature set which is used for different applications like face recognition, facial expressions and behavior and gender classification.

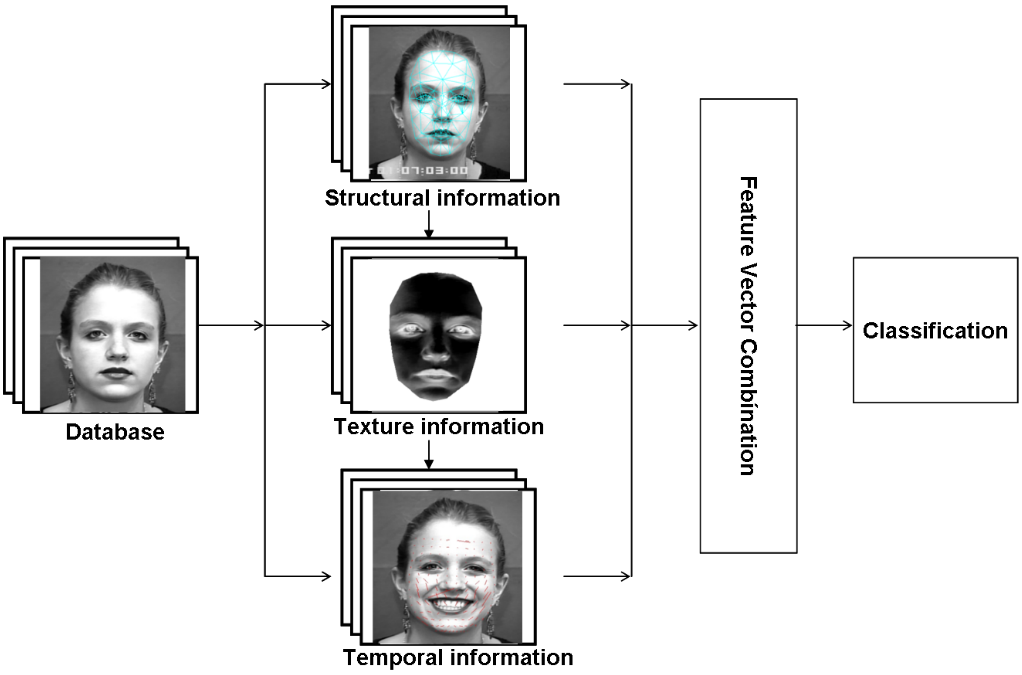

Figure 1: Our Approach: Sequential flow for feature extraction

4 Our approach: proposed solution

We propose a model based image analysis solution to this problem. Model parameters are obtained in an optimal way to maximize information from face region under various factors like facial pose, expressions, illuminations and self occlusions. In this section we explain our approach in different modules including shape model fitting, image warping, pose, illumination and expressions normalizations and finally synthesizing the image to extract model parameters.

We use a wireframe 3D face model known as Candide-III (10). The model is fitted to the face image by learning objective functions. We find correspondences between 2D and 3D points for model fitting and texture mapping. Texture information is mapped from the example image to a reference shape which is the mean shape of all the shapes available in database. However the choice of mean shape is arbitrary. Image texture is extracted using planar subdivisions of the reference and the example shapes. We use delaunay triangulations of the distribution of our model points. Texture warping between the triangulations is performed using affine transformation. Principal Component Analysis (PCA) is used to obtain the texture and shape parameters of the example image. This approach is similar to extracting AAM parameters. In addition to AAM parameters, temporal features of the facial changes are also calculated. Local motion of the feature points is observed using optical flow. We use reduced descriptors by trading off between accuracy and run time performance. These features are then used for classification. Our approach achieves real-time performance and provides robustness against facial expressions in real-world scenarios. Currently the system finds the pose information implicitly in structural parameters whereas illuminations are dealt in appearance parameters. This computer vision task comprises of various phases shown in Figure 1 for which it exploits model-based techniques that accurately localize facial features, seamlessly track them through image sequences, and finally infer facial features. We specifically adapt state-of-the-art techniques to each of these challenging phases.

5 Determining High-Level Information

In order to initialize, we apply the algorithm of Viola et al. (23) to roughly detect the face position within the image. Then, model parameters are estimated by applying the approach of Wimmer et al. (7) because it is able to robustly determine model parameters in real-time.

To extract descriptive features, the model parameters are exploited. The model configuration represents information about various facial features, such as lips, eye brows or eyes and therefore contributes to the extracted features. These structural features include both, information about the person's face structure that helps to determine person-specific information such as gender or identity. Furthermore, changes in these features indicate shape changes and therefore contribute to the recognition of facial expressions.

The shape $\boldsymbol{x}$ is parameterized by using mean shape $\boldsymbol{x_m}$ and matrix of eigenvectors $\boldsymbol{P_s}$ to obtain the parameter vector $\boldsymbol{b_s}$ (14).

$$\boldsymbol{x} = \boldsymbol{x_m} + \boldsymbol{P_s b_s} \tag{1}$$

Once we have shape information of the image, we extract texture from the face region by mapping it to a reference shape. A reference shape is extracted by finding the mean shape over the dataset. Image texture is extracted using planar subdivisions of the reference and the example shapes. Texture warping between the subdivisions is performed using affine transformation. This extracted texture is parameterized in the same manner as shape information using PCA.

The extracted texture is parameterized after illumination normalization using mean texture $\boldsymbol{g_m}$and matrix of eigenvectors $\boldsymbol{P_g}$to obtain the parameter vector $\boldsymbol{b_g}$ (14). Figure 2 shows shape model fitting and texture extracted from face image.

$$\boldsymbol{g} = \boldsymbol{g_m} + \boldsymbol{P_g b_g} \tag{2}$$

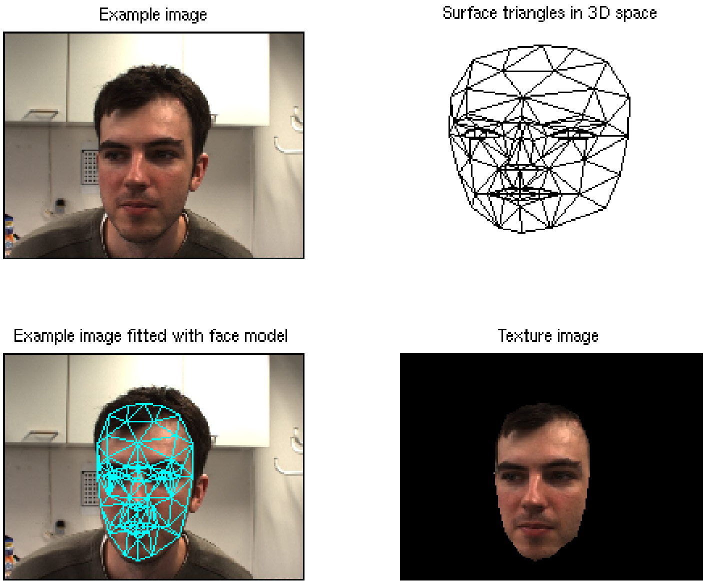
Example image
Surface triangles in 3D space
Example image fitted with face model
Texture image

Figure 2: Texture information is represented by an appearance model. Parameters of the fitted model are extracted to represent single image information.

Further, temporal features of the facial changes are also calculated that take movement over time into consideration. Local motion of feature points is observed using optical flow. We do not specify the location of these feature points manually but distribute equally in the whole face region. The number of feature points is chosen in a way that the system is still capable of performing in real time and therefore inherits a tradeoff between accuracy and runtime performance. Figure 3 shows motion patterns for some of the images from database.

	BDT	BN
Face Recognition	98.49%	90.66%
Facial Expressions Recognition	85.70%	80.57%
Gender Classification	99.08%	89.70%

Table 1: Unified features tested with two classifiers Bayesian Networks (BN) and Binary Decision Tree (BDT)

We combine all extracted features into a single feature vector. Single image information is considered by the structural and textural features whereas image sequence information is considered by the temporal features. The overall feature vector becomes:

$$\boldsymbol{u} = (\boldsymbol{b_{s1}}, ..., \boldsymbol{b_{sm}}, \boldsymbol{b_{g1}}, ..., \boldsymbol{b_{gn}}, \boldsymbol{b_{t1}}, ..., \boldsymbol{b_{tp}},) \tag{3}$$

Where $\boldsymbol{b_s}$, $\boldsymbol{b_g}$ and $\boldsymbol{b_t}$ are shape, textural and temporal parameters respectively.

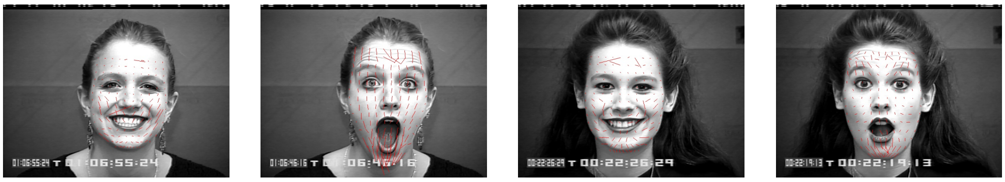

Figure 3: Motion patterns within the image are extracted and the temporal features are calculated from them. These features are descriptive for a sequence of images rather than single images.

6 EXPERIMENTS

For experimentation purposes, we benchmark our results on Cohn Kanade Facial Expression Database (CKFED). The database contains 488 short image sequences of 97 different persons performing six universal facial expressions (15). Furthermore, a set of action units (AUs) has been manually specified by licensed Facial Expressions Coding System (FACS) (16) experts for each sequence.

In order to experiment feature versatility we use two different classifiers with same feature set on three different applications: face recognition, facial expressions recognition and gender classification. The results are evaluated using classifiers from weka (24) with 10-fold cross validation. Table 1 shows different recognition rates achieved during experimentations using Bayesian Networks (BN) and Binary Decision Tree (BDT). In all three cases BDT outperforms BN. This can be analyzed in the figures. Figure 4 shows true positive and false positive rates for all the subjects in the database for face recognition. Figure 5 shows receiver operating characteristic (ROC) curves for six different facial expressions. Since laugh and fear are mostly confused facial expressions, it

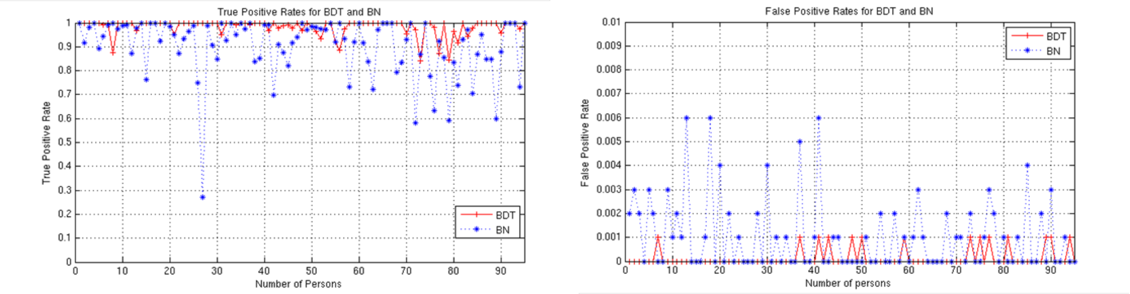

Figure 4: True positive and false psitive rates for face recognition

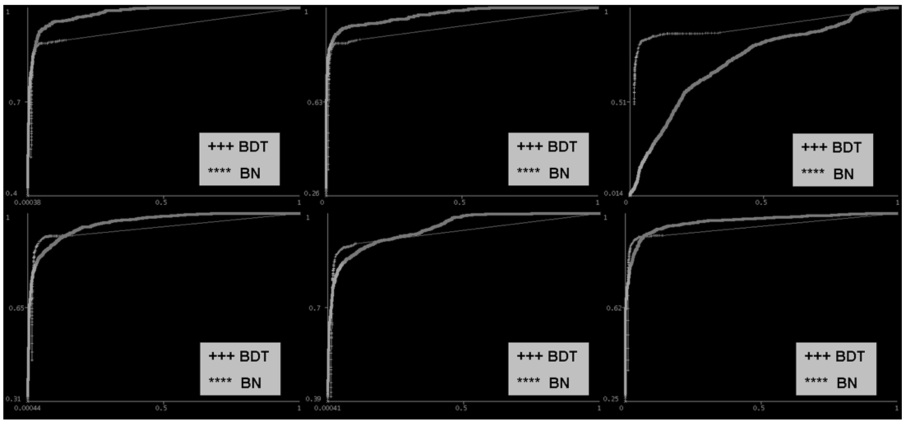

Figure 5: ROC curves for facial expressions Top (left to right): anger, disgust, fear, Bottom (left to right): laugh, sadness, surprise

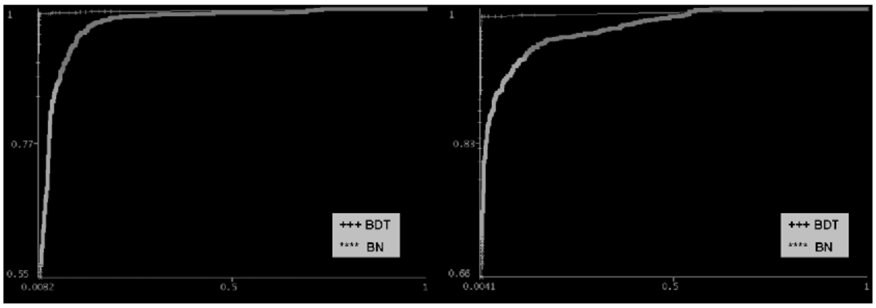

Figure 6: ROC curves for gender classification (left) female, (right) male

can be analyzed from the curves that there exists some confusion between these two expressions for BN classifier. Figure 6 shows gender classification results.

7 Conclusions

In this paper we introduced an idea to solve the face image analysis problem in human robot joint interaction scenarios using a common set of features. The approach is intuitive to human-human interaction where humans can extract these information on a very first look. Another advantage of this approach is to find a robust features set only once during the joint interaction and leaving much of the time for other joint activities. We addressed a feature extraction technique for a face recognition in the presence of facial expressions, recognizing facial expressions and gender. We deal majorly with facial expressions variations on a benchmark database and pose and lighting conditions are considered implicitly in the shape and texture parameters respectively. However, future work includes extension of these features under the presence of other variations. We aim a robust system to apply in real time human robot interaction.

References

[1] Beetz M. et. al. The Assistive Kitchen — A Demonstration Scenario for Cognitive Technical Systems In *Proceedings of the 4th COE Workshop on Human Adaptive Mechatronics (HAM)*, 2007

[2] Blanz V., Vetter T. Face Recognition Based on Fitting a 3D Morphable Model. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol.25 no. 9, pp 1063 - 1074, 2003.

[3] B. Ginneken, A. Frangi, J. Staal, B. Haar, and R. Viergever. Active shape model segmentation with optimal features. *IEEE Transactions on Medical Imaging*, 21(8):924–933, 2002.

[4] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[5] G. J. Edwards, T. F. Cootes and C. J. Taylor, Face Recognition using Active Appearance Models *in Proceeding of European Conference on Computer Vision* 1998 vol. 2, pp- 581-695, Springer 1998.

[6] P. N. Belheumeur, J. P. Hespanha and D. J. Kreigman, Eigenfaces vs Fisherfaces: Recognition using Class Specific Linear Projection *IEEE Transaction on Pattern Analysis and Machine Intelligence* Vol 19, No. 7, July 1997.

[7] Wimmer M, Stulp F, Tschechne S, and Radig B, Learning Robust Objective Functions for Model Fitting in Image Understanding Applications. In *Proceedings of the 17th British Machine Vision Conference* pp1159–1168, BMVA, Edinburgh, UK, 2006.

[8] Tim F. Cootes and Chris J. Taylor. Active shape models – smart snakes. In *Proceedings of the 3rd British Machine Vision Conference* pages 266 - 275. Springer Verlag, 1992.

[9] Cootes T. F., Edwards G. J., Taylor C. J. Active Appearance Models. In *Proceedings of European Conference on Computer Vision* Vol. 2, pp. 484-498, Springer, 1998.

[10] J. Ahlberg. An Experiment on 3D Face Model Adaptation using the Active Appearance Algorithm. *Image Coding Group, Deptt of Electric Engineering* Linköping University.

[11] Bronstein, A. Bronstein, M. Kimmel, R. Spira, A. 3D face recognition without facial surface reconstruction, In Proceedings of European Conference of Computer Vision Prague, Czech Republic, May 11-14, 2004

[12] Unsang Park and Anil K. Jain 3D Model-Based Face Recognition in Video *2nd International Conference on Biometrics* Seoul, Korea, 2007

[13] W. Zhao, R. Chellapa, A. Rosenfeld and P.J. Philips, Face Recognition: A Literature Survey In *ACM Computing Surveys* Vol. 35, No. 4, December 2003, pp. 399-458.

[14] Stan Z. Li and A. K. Jain. Handbook of Face recognition *Springer* 2005

[15] Kanade, T., Cohn, J. F., Tian, Y. (2000). Comprehensive database for facial expression analysis In *Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FGR00)*, Grenoble, France, 46-53.

[16] P. Ekman and W. Friesen. *The Facial Action Coding System: A Technique for The Measurement of Facial Movement.* Consulting Psychologists Press, San Francisco, 1978.

[17] P. Michel and R. E. Kaliouby. Real time facial expression recognition in video using support vector machines. In *Fifth International Conference on Multimodal Interfaces*, pages 258–264, Vancouver, 2003.

[18] J. Cohn, A. Zlochower, J.-J. J. Lien, and T. Kanade. Feature-point tracking by optical flow discriminates subtle differences in facial expression. In *Proceedings of the 3^{rd} IEEE International Conference on Automatic Face and Gesture Recognition*, pages 396 – 401, April 1998.

[19] I. Kotsia and I. Pitaa. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transaction On Image Processing*, 16(1), 2007.

[20] Riaz Z. et al. A Model Based Approach for Expression Invariant Face Recognition In *3^{rd} International Conference on Biometrics*, Italy, June 2009

[21] A. Scheenstra, A. Ruifrok and R. C. Veltkamp, A Survey of 3D Face Recognition Methods In *AVBPA 2005, LNCS 3546* pp. 891-899, 2005

[22] S. Romdhani. *Face Image Analysis using a Multiple Feature Fitting Strategy.* PhD thesis, University of Basel, Computer Science Department, Basel, CH, January 2005.

[23] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[24] Ian H. Witten and Eibe Frank Data Mining: Practical machine learning tools and techniques *Morgan Kaufmann*, 2nd Edition, San Francisco, 2005.

Gaining Rapport by Voicing Appropriate Emotional Responses Based on User State

Jaime C. Acosta
University of Texas at El Paso
El Paso, Texas 79968
jcacosta@miners.utep.edu

Abstract

This paper describes research that will be used to build a spoken dialog system capable of engaging in the rich emotional interplay associated with gaining rapport. For this work, dialogs between a persuasive graduate coordinator and undergraduate students were collected and labeled using the three dimensions of emotion. Analysis of the voice-only conversations reveals that emotions are present and perceivable. Moreover, there is evidence that the graduate coordinator voices emotion in her responses based on the students emotion in the previous utterance. These findings will be used to implement a spoken dialog system. This system will be evaluated according to its ability to gain rapport with users.

Keywords: rapport, emotion detection, emotion synthesis, immediate response patterns, adjacency pair analysis

1 Introduction

As information sources become richer and technology advances, the use of computers to deliver information is increasing. In particular, interactive voice technology for information delivery is becoming more common due to improvements in technologies such as automatic speech recognition, and speech synthesis.

Several problems exist in these voice technologies including speech recognition accuracy and lack of common sense and basic knowledge. Among these problems is the inability to achieve rapport.

Gratch et al. (2007) defines rapport as *a feeling of connectedness that seems to arise from rapid and contingent positive feedback between partners and is often associated with socio-emotional processes.* In the field of neuro-linguistics, O'Connor and Seymour (1990) stated that matching or complimenting voice features such as volume, speed, and intonation, is important to gain rapport. Communication Accommodation Theory (Shepard et al., 2001) states that humans use prosody and backchannels in order to adjust social distance with an interlocutor. These features of voice can also be associated with emotions.

Previous work has shown that automated systems can gain rapport by reacting to user gestural nonverbal behavior (Chartrand and Bargh, 1999; Gratch et al., 2007; Cassell and Bickmore, 2003). In contrast, this research looks at how rapport can be gained through voice-only interaction.

Preliminary analysis of human-human dialog provides evidence that prosodic features, associated with emotion by two judges, are used by an interlocutor during persuasive dialog. Figure 1 shows the pitch of a sound snippet from the corpus and how it differs from neutral, computer synthesized voice (produced using MaryTTS). This illustrates the more general fact that when humans speak to each other, we display a variety of nonverbal behaviors in voice, especially when trying to build rapport. The main hypothesis of this research is that a spoken dialog system with emotional intelligence will be effective for gaining rapport with human users.

The rest of this paper is structured as follows: first, related work is reviewed and current limitations for building automated rapport are described. Afterwards, the hypotheses and expected contributions of this work are described along with the research approach. Lastly, broader significance of this work is discussed.

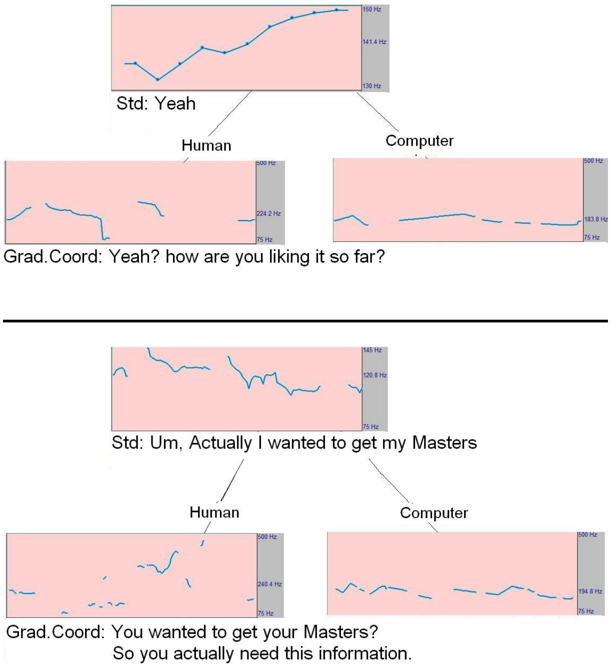

Figure 1: Pitch levels of a conversation taken from the persuasive dialog corpus includes a student (Std) and a graduate coordinator (Grad.Coord). Pitch was analyzed using the Praat software. It can be seen that the student displays rich prosody in voice (tree parents) and that the human response (left branch) contains more varied prosody than the computer synthesized voice (right branch).

2 Related Work

Communication Accommodation Theory states that people use nonverbal feedback to establish social distance during conversation. In order to gain rapport, people would most likely want to decrease social distance to achieve the connectedness and smoothness in conversation that is seen in human social interaction. Research in human-computer interaction has pursued these nonverbal behaviors through appropriate backchanneling, head nods, and gaze techniques, but still missing is attention to user emotional state, which can be detected through some of these nonverbal behaviors in voice.

Two methods for describing emotions are discrete and dimensional. Discrete emotions include anger, disgust, fear, joy, sadness, and surprise. Dimensional emotions use two or more components to describe affective state. More commonly used dimensions are evaluation (a.k.a. valence), activity, and potency (a.k.a. power) as described in Osgood (1957). Automatic emotion recognition has had limited success with discrete emotions, *e.g.* D'Mello et al. (2008). In the tutoring domain, some have looked at appropriately responding to students based on their prosody in voice (Forbes-Riley and Litman, 2007; Hollingsed and Ward, 2007). The difficulty of recognizing discrete emotions exists because humans typically show more subtle emotions in most real human-human interactions (Batliner et al., 2000). Forbes-Riley and Litman (2004) had promising results by looking at a three-class set of emotions (positive, negative, neutral).

The intent of this research is to develop a method for detecting three dimensions of emotion from voice in order to build rapport. There is a possibility that using a dimensional approach will enable more accurate modeling of subtle emotions that exist in spontaneous human-human dialogs.

3 Hypotheses and Expected Contributions

The main hypothesis of this work is that a spoken dialog system with emotional intelligence will be more effective for gaining rapport than a spoken dialog system without emotional intelligence. In order to test this hypothesis, I will implement and evaluate a spoken dialog system. This system will choose topics and content depending on user emotional state. The resulting system will advance the state of the art in three technologies: recognizing appropriate emotion, planning accordingly, and synthesizing appropriate emotion. The system will also demonstrate how to integrate these components.

In addition to choosing the correct content based on user emotional state, this research will investigate the effect of adding emotion to voice for rapport. The second hypothesis of the research is that expressing emotion in voice and choosing words, compared to expressing emotion only by choosing words, will be more effective for building rapport with users.

4 Approach

This section outlines the steps that have been completed and those that are still pending to accomplish the goals of the research.

4.1 Corpus Analysis and Baseline System

This work is based on a persuasive dialog corpus consisting of audio recordings of 10 interactions averaging 16 minutes in length. The corpus consists of rougly 1000 turns between a graduate coordinator and individual students. The graduate coordinator was a personable female staff member who was hired by the University to raise the graduate student count. The students were enrolled in an introductory Computer Science course and participated in the study as part of a research credit required for course completion. The students had little knowledge of the nature or value of graduate school and of the application process. Preliminary analysis of the corpus showed evidence of a graduate coordinator building rapport with students by using emotion.

A baseline system built using commercial state-of-the-art software was implemented based on the corpus (mainly the topics covered). Informal user comments about the baseline system helped determine missing features for automated rapport building technology. One salient feature that is missing is attention to emotion in voice. This confirmed the direction of this research.

This corpus was transcribed and annotated with dimensional emotions (activation, valence, and power) by two judges. Activation is defined as sounding ready to take action, valence is the amount of positive or negative sound in voice, and power is measured by the amount of dominance in voice. The dimensions are annotated numerically on scales from -100 to +100.

The following are examples taken from the corpus with annotated acoustic features.

- Example 1

  Grad.Coord(GC1): *So you're in the 1401 class?* [rising pitch]

  Subject(S1): *Yeah.* [higher pitch]

  GC2: *Yeah? How are you liking it so far?* [falling pitch]

  S2: *Um, it's alright, it's just the labs are kind of difficult sometimes, they can, they give like long stuff.* [slower speed]

  GC3: *Mm. Are the TAs helping you?* [lower pitch and slower speed]

  S3: *Yeah.* [rising pitch]

GC4: *Yeah.* [rising pitch]

S4: *They're doing a good job.* [normal pitch and normal speed]

GC5: *Good, that's good, that's good.* [normal pitch and normal speed]

- Example 2

  GC6: *You're taking your first CS class huh.* [slightly faster voice]

  S5: *Yeah, I barely started.* [faster voice]

  GC7: *How are you liking it?* [faster voice, higher pitch]

  S6: *Uh, I like it a lot, actually, it's probably my favorite class.* [faster, louder]

  GC8: *Oh good.* [slower, softer]

  S7: *That I'm taking right now yeah.* [slightly faster, softer]

  GC9: *Oh that's good. That's exciting.* [slow and soft then fast and loud]

  GC10: *Then you picked the right major you're not gonna change it three times like I did.* [faster, louder]

In the first example, the coordinator noticeably raises her pitch at the end of her utterance. This is probably so that she can sound polite or interested. On line S2, the subject displays a falling pitch (which sounds negative) and the coordinator responds with a lower fundamental frequency and a slower speed. The subject sounds unsure by displaying a rising pitch in his answer (S3). The coordinator mirrors his response (GC4) and finally both interlocutors end with normal pitch and normal speed.

In the second example, the subject speaks faster than usual (S5). The coordinator compensates by adjusting her speed as well. From S6 through GC8, when the subject's voice gets louder, the coordinator's voice gets softer, almost as though she is backing off and letting the subject have some space. In GC9 the coordinator responds to the student's positive response (liking the class) and becomes immediately faster and louder.

A next step for the analysis is to determine the most expressive acoustic correlates for emotions. Informal auditory comparisons show some possible correlations (see Table 1). These correlations seem promising because many correspond with previous work (Schroder, 2004).

Table 1: Informal analysis reveals acoustic correlates possibly associated with the dimensions of emotion

Dimension	High	Low
Activation	Faster, more varied pitch, louder	Slower, less varied pitch, softer
Valence	Higher pitch throughout, laughter, speed up	Falling ending pitch, articulation of words, increasing loudness
Power	Faster, louder, falling ending pitch, articulation of word beginnings, longer vowels	Softer, higher pitch throughout, quick rise in pitch, smoother word connection

The emotion annotations of the two judges show that strategies for adaptive emotion responses can be extracted from the corpus. Communication Accomodation Theory states that interlocutors mirror nonverbal behaviors during interaction when attempting to decrease social distance. The coordinator's emotional responses were correlated with the student's emotional utterances to determine if emotional mirroring (matching student emotion and coordinator response) was present in the persuasive dialog corpus. This was the case in the valence dimension, which showed a correlation coefficient of 0.34. However, regarding power, there was an inverse relationship; if the student showed more power, the coordinator showed less (–0.30 correlation coefficient). Activation showed a small correlation coefficient (–0.14).

To realize a spoken dialog system that could model this responsive behavior, machine learning was used. The students' three emotion dimensions were taken as attributes and were used to predict the coordinators emotional responses using Bagging with REPTrees. Measuring the correlations between the predictions of the model and the actual values in the corpus revealed correlation coefficients of 0.347, 0.344, and 0.187 when predicting the coordinator's valence, power, and activation levels, respectively. One possible reason for the especially low activation scores may be attributed to the recording quality. The student and coordinator sometimes adjusted the microphones leading to varying loudness and introduction of noise in some utterances.

4.2 Full System

The full system will provide a means to evaluate whether emotion contributes to automated rapport building. This system will be based on several available technologies and previous research in spoken dialog systems.

Figure 2 shows the different components anticipated for the full system. The components that will be implemented for this research include emotion recognition, user modeling components, and text and emotion strategy databases. The other components will be based on available open source software packages. The implementation effort also includes the integration of all components.

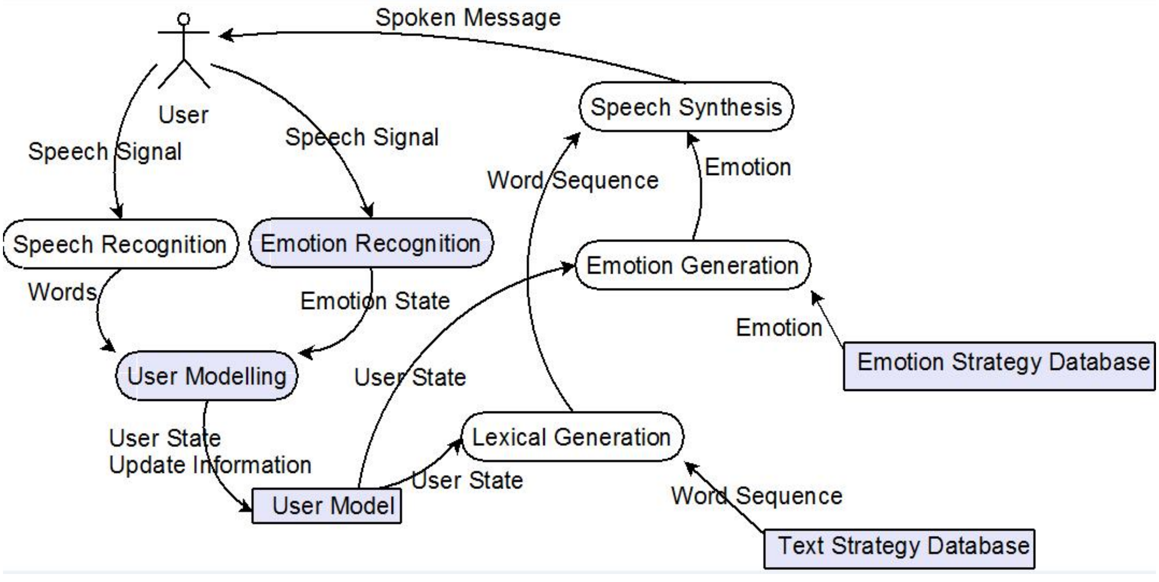

Figure 2: Full System Dataflow Diagram

The following is a scenario that depicts how the full system will operate.

1. The system begins by saying "How are you doing today?"
2. The user says "I'm doing good" with a negative sounding voice.
3. The voice signal is then processed through the speech recognizer and emotion recognizer in parallel. The speech recognizer extracts words from the voice signal while the emotion recognizer extracts emotion.

4. This data is sent to the user modeling component which determines the immediate user state based only on the current emotion and the words spoken. In this scenario, the user's state will be negative even though the user stated otherwise.

5. This user state update information is then passed to the user model which updates the current user state. This component contains knowledge, beliefs and feelings of the user. Since there was no previous user state, the current emotion is set to negative. Stored in user knowledge will be the fact that the user was asked "How are you doing today?". Some information about the user's contradictory state is stored as user beliefs: stated good, but sounds negative.

6. Next, this information is used to select some predefined text from the lexical generation along with an associated emotion from the emotion strategy database (these two are done in parallel). Since the user's state is negative, the system may choose to ask another question such as "ok, do you have any concerns?" with a negative sounding voice (to mirror the valence dimension). In contrast, if the user was positive, the system may have chosen something similar to "great, let's get going then" with a highly positive voice.

7. Lastly, the text with corresponding emotion coloring is rendered to speech and played to the user by the speech synthesis component.

4.3 Evaluation

To achieve the final goal of determining whether emotion helps gain rapport, the final system described herein will be evaluated.

The final system will be configurable; it will allow for enabling emotion in voice (*voiced*) or disabling the emotions in voice (*not voiced*). In addition, there will be a control configuration, perhaps one that will display a random emotion (*random*). A user study (hopefully within subjects) will be conducted that will ask users to interact with four versions of the system (baseline, *voiced*, *not voiced*, and *random*). A post-test questionnaire consisting of Likert scales will ask users how much rapport they felt with each version of the system. In addition, some objective metrics such as disfluency count and interaction time will be collected. This will help test the two hypotheses of this research. First, it is expected that subjects will have more rapport with the *not voiced* configuration than with the baseline system. The second hypothesis will be verified by determining if subjects have more rapport with the *voiced* than with the *not voiced* system. The *random* configuration will be used to determine whether the system's adaptive responses are better than random responses.

5 Broader Significance

This research addresses methods for gaining rapport as an important dimension of successful human-computer interaction, and one likely to be useful even for business-like dialogs. For example, building rapport with customers can decrease the number of disfluencies, which are currently a problem for speech recognizers. In addition, customer support systems will have the ability to tailor responses to decrease negative emotion.

Similarly, the learned rules for detecting emotion and responding appropriately could be used to train people how to more effectively gain rapport. Lastly, this work can supplement other rapport research that uses other forms of nonverbal behavior such as gaze and gestures seen especially in embodied conversational agents.

6 Acknowledgements

I would like to thank my advisor, Nigel Ward for his help. Also, I would like to thank Anais Rivera and Sue Walker for the collection of the persuasive dialog corpus and Jun Zheng for his help in fine tuning the baseline system. This work is supported in part by NSF grant IIS-0415150.

References

Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E. (2000). Desperately Seeking Emotions or: Actors, Wizards, and Human Beings. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. ISCA.

Cassell, J. and Bickmore, T. (2003). Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents. *User Modeling and User-Adapted Interaction*, 13(1):89–132.

Chartrand, T. and Bargh, J. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910.

D'Mello, S., Craig, S., Witherspoon, A., McDaniel, B., and Graesser, A. (2008). Automatic detection of learners affect from conversational cues. *User Modeling and User-Adapted Interaction*, 18(1):45–80.

Forbes-Riley, K. and Litman, D. (2004). Predicting emotion in spoken dialogue from multiple knowledge sources. *Proc. Human Language Technology Conf. of the North American Chap. of the Assoc. for Computational Linguistics (HLT/NAACL).*

Forbes-Riley, K. and Litman, D. (2007). Investigating Human Tutor Responses to Student Uncertainty for Adaptive System Development. In *Affective Computing and Intelligent Interaction Second International Conference, ACII 2007.* Lisbon, Portugal, September 12-14, 2007: Proceedings. Springer.

Gratch, J., Wang, N., Okhmatovskaia, A., Lamothe, F., Morales, M., van der Werf, R., and Morency, L. (2007). Can Virtual Humans Be More Engaging Than Real Ones? *12th International Conference on Human-Computer Interaction.*

Hollingsed, T. K. and Ward, N. G. (2007). A combined method for discovering short-term affect-based response rules for spoken tutorial dialog. *Workshop on Speech and Language Technology in Education (SLaTE).*

O'Connor, J. and Seymour, J. (1990). *Introducing neuro-linguistic programming.* Mandala.

Osgood, C. (1957). *The Measurement of Meaning.* University of Illinois Press.

Schroder, M. (2004). Dimensional Emotion Representation as a Basis for Speech Synthesis with Non-extreme Emotions. *In Proceedings Workshop Affective Dialogue Systems*, 3068:209–220.

Shepard, C., Giles, H., and Le Poire, B. (2001). Communication accommodation theory. *The new handbook of language and social psychology*, pages 33–56.

Synthesis of Nonverbal Listener Vocalizations

Sathish Pammi
DFKI GmbH, Saarbrücken, Germany
Sathish.Pammi@dfki.de

Abstract

Listener vocalizations play an important role in communicating listener intentions while the interlocutor is talking. Synthesis of listener vocalizations is one of the focused research areas to improve emotionally colored conversational speech synthesis. The major objective of the work presented in this paper is providing a new functionality to text-to-speech synthesis system that can synthesize nonverbal listener vocalizations. As synthesis of listener vocalizations is a new topic in conversational speech synthesis, many research questions are raised. A methodology is proposed to conduct research on those questions which can provide solutions to build a system to generate nonverbal listener vocalizations. We discuss the work done so far according to proposed working strategy and tentative plans for future work.

Keywords: nonverbal listener vocalizations, back-channel, multi-modal interaction, speech synthesis

1 Introduction

In multimodal human-computer interaction, the ability of systems to generate listener vocalizations (Gardner, 2002) is an important requirement for generating affective interaction.

Listener vocalizations include back-channel utterances (Yngve, 1970; Ward and Tsukahara, 2000) related to the flow of the conversation as well as affect vocalizations (Schröder et al., 2006) based on the listener's affective state (Scherer, 2003). For example, nonverbal listener vocalizations like *mm-hm* or *uh-huh* can be used as back-channel utterances to keep the floor open for the current speaker to continue speaking. Listener vocalizations can also transmit affective states like excited, bored, confused, surprised, etc. For example, *wow* can be used for both back-channel and to communicate affective meaning. Listener vocalizations also include non-linguistic vocalizations like laughter or sigh as well as some response tokens like *yes*, *right*, *really* or *absolutely*.

Nowadays, speech synthesis systems are providing high quality synthetic reading speech. Synthesis of nonverbal listener vocalizations, a new functionality to text-to-speech synthesis systems, provides an opportunity to build interactive synthesis systems suitable to multi-modal interaction systems. Database collection, annotation and realization of speech waveform are crucial steps in building speech synthesis systems. Above three major steps need more investigation in case of the new functionality. For example, traditional speech synthesis databases including expressive speech material were recorded in a studio environment with a single speaker using predefined recording scripts, but this traditional recording setup is not suitable to capture listener vocalizations as they are natural only in a conversation. Success in generation of listener vocalizations depends on the answers to the following questions:

- How to collect a database of listener vocalizations?
- What kinds of meanings are expressed through listener vocalizations?
- What form is suitable for a given meaning?
- How to annotate meaning and behavior (form) of a listener vocalization?
- How to realize the form using a technological framework?

Many listener vocalizations are short and nonverbal in nature. As synthesis of nonverbal vocalizations is a new topic in synthesis, we are not aware of any technological framework to synthesize these vocalizations. In the level of realization, some technological research questions should be answered like:

- What kind of technology is suitable to synthesize nonverbal vocalizations? Unit-selection, HMM-based or other.

- If it is Unit-selection, what strategy would be better to select a unit?
- If it is HMM-based, how to model and realize nonverbal vocalizations?
- How to get advantage from signal modification algorithms?

The major objective of this work is not only providing answers to the above research questions, but also building a system, which will be integrated into SEMAINE (SEMAINE, 2008) multi-modal interaction system, to synthesize nonverbal listener vocalizations. The system has to be robust and it has to use standard representation like eXtensible Markup Language (XML) formats in the view of future inter-module communication. A possibility is there to raise more research questions when we try to evaluate our final system as part of a real-time SEMAINE demonstration system.

A methodology is proposed in Section 2 to conduct research on synthesis of nonverbal vocalizations. We describe the results of the data collection and annotation in Section 3 and this section also explains our baseline system. In Section 4, we discuss our tentative plans and proposals to build a system for realization of listener vocalizations in a speech synthesis framework.

2 Methodology

The SEMAINE system, a demonstration of audiovisual Sensitive Artificial Listener(SAL) (Douglas-Cowie et al., 2008), aims to build a virtual dialog partner who intends to engage the user in a conversation by paying attention to the user's emotions and nonverbal expressions. Different 'action proposers' in the system produces different 'action commands' to synthesize a meaningful agent behavior. Simulation of a convincing audiovisual listener behavior is one major part of the system. According to the project plans, an action proposer, with the help of multi-modal inputs, will be planning the intention of the listener as well as the timing information to trigger the behavior. The description of listener intention uses standard XML representation ('Multi-modal XML input' in the Figure 2). Our part of the work is mainly focusing on modules for synthesis of appropriate listener vocalizations when the intended meaning behind the listener intention is given.

This section describes conceptual model of our proposed methodology to build a framework for synthesis of nonverbal vocalizations. The proposed work consists of three different levels (as shown in Figure 1): Data collection, Annotation and Realization.

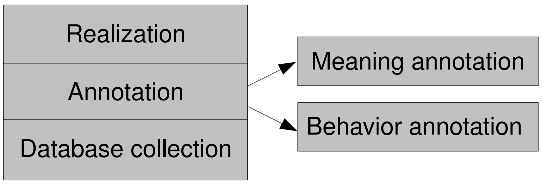

Figure 1: Major aspects of proposed work

2.1 Database collection

As the traditional way of recording setup is not useful to capture nonverbal listener vocalizations, we propose to record a natural dialog speech between an actor and his dialog partner in an anechoic studio because listener vocalizations seem to be natural only in a conversation. According to the new proposed recording setup, the actor and his dialog partner will sit in different rooms and hear each other using headphones, so that we can record each speaker's voice on a different channel without interference of the other speaker's speech. As we are aiming to capture listener vocalizations, the actor will be instructed to participate in a free dialog, but to take predominantly a listener role.

2.2 Annotation

To know different kinds of meanings expressed through listener vocalizations, the intended meaning behind each vocalization should be annotated. Similarly, the annotation of behavioral properties will be useful to know suitable behavior for a given meaning. Initially, we do not know how many meaning or behavior categories can be used to annotate all listener vocalizations, so we propose to annotate all nonverbals using informal descriptions to make sure that we are not guided by any pre-existing set of categories. Pre-existing sets of categories may or may not be suitable to represent all listener vocalizations available in our data. So informal descriptions will be helpful to understand better the structure of both behavior and meaning. Subsequent grouping of these descriptions will help to understand the types of behavior and meaning of listener vocalizations, at least for the speaker we studied. In the later stages, a suitable limited set of categories that capture the essence of meaning as recorded in informal descriptions will be identified.

The sequence of steps involved in the proposed annotation scheme is: Firstly, start-end time labels will be annotated for all listener vocalizations made by the actor. Secondly, informal descriptions will be provided for each labeled segment in three different levels: content, behavior, sub-texts. In latter stages, suitable meaning category will be identified for each vocalization with the help of informal descriptions. Finally, annotation for behavioral properties like intonation, voice quality etc.. will be provided.

2.3 Realization

The conceptual model for the realization system, as shown in Figure 2, contains off-line and runtime processing modules. Data analysis on annotated speech samples is a crucial step in off-line processing which provides relations between behavior and meaning. The experience from this analysis will let us know whether the relation between meaning and behavior is one-one mapping or a single behavior can be usable to simulate multiple intended meanings. A thorough research is expected in the level of technological framework to realize a nonverbal listener vocalization. For example, we have to find a way to model and generate nonverbal vocalizations if we choose Hidden Markov Model (HMM) based synthesis as a technological framework.

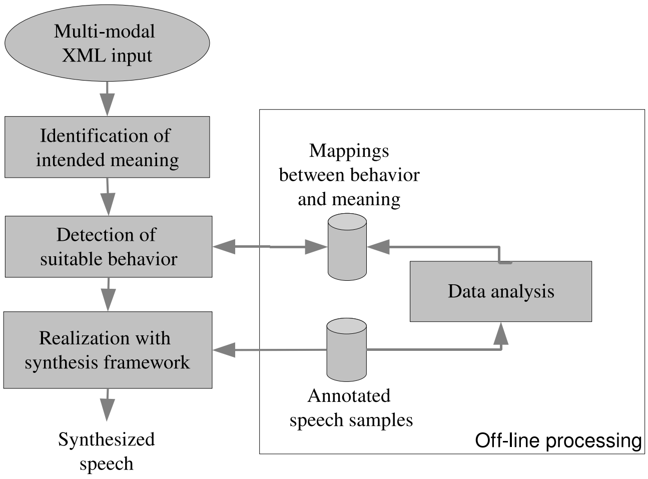
Multi-modal XML input
Identification of intended meaning
Detection of suitable behavior
Realization with synthesis framework
Synthesized speech
Mappings between behavior and meaning
Data analysis
Annotated speech samples
Off-line processing

Figure 2: Conceptual model of proposed realization system

The proposed runtime system will work as follows: Initially, an XML front-end processing module will identify the intended meaning behind requested nonverbal vocalizations. The next module will be finding suitable behavior to the requested meaning category with the knowledge of relations between behavior and meaning. Finally, another module will realize appropriate behavior with a synthesis technology like Unit-selection or HMM-based.

3 Results so far

The work has been progressing on all three levels described in Section 2. This section explains the results of the work done so far.

3.1 Database collection

We recorded dialog speech in a studio environment as described in Section 2.1. Our speaker is a professional male German actor with whom we had already recorded expressive speech synthesis databases in the past. Using this speaker was essential for being able to use the recorded vocalizations with our synthesis voices in the future. Recordings were made in several stages and in sessions of about 20 minutes each. In the initial stage, we instructed the actor to “be himself” (not to act) and in the later stages, he was instructed to act like one of three characters representing different emotionally colored personalities (Douglas-Cowie et al., 2008): Spike is always aggressive, Obadiah is always gloomy, Poppy is always happy. Two female student assistants took turns as the dialog partner, and tried to keep the actor in listening mode for a maximum amount of time while they were talking to the actor about a topic of their choice. The dialogue partners were sitting in separate rooms and hearing each other using headphones. Each speaker’s voice was recorded on a separate channel.

As a result of the database collection exercise, we obtained around six hours of German dialog speech. Table 1 provides statistics of dialog speech material.

The actor status	Corpus duration (in minutes)	Number of listener vocalizations
Natural	190	568
Obadiah	45	181
Poppy	45	93
Spike	70	238
Total	350	1080

Table 1: Corpus duration in minutes when the actor is being himself (natural) or acted like an emotional character.

3.2 Annotation

So far in this project, we have worked on meaning annotation only. A detailed version of results in meaning annotation were reported in (Pammi and Schröder, 2009), but an overview of those results were shortly discussed in this section. As outlined in Section 2.2, so far in this work, we have worked on informal description and meaning annotation only.

3.2.1 Informal descriptions

In order to get a fuller picture of the data, we use a detailed informal description of each vocalization before trying to find suitable categories to represent the meaning and behavior observed. An informal description in this work contains an annotator’s description of the form, content and subtext of each listener vocalization using his/her own vocabulary. The form provides information about phonetic segments, voice quality, duration and/or intonation. Similarly, the content and subtext tiers describe the meaning and, optionally, a suitable text substitution.

3.2.2 Instructions given to annotate meaning

We used the Baron-Cohen (Baron-Cohen et al., 2004) set of 33 categories describing epistemic-affective states as a starting point for our tag set. Annotators were instructed to use only those categories from the set that seemed appropriate, and to add categories that seemed necessary to describe the data but were not contained in the Baron-Cohen set. They could use categories from the Geneva Emotion Wheel (Scherer, 2005) or propose their own category labels as they felt appropriate.

According to informal descriptions provided from annotators, listener vocalizations seem to differ with respect to their reference: self expression, stance towards the other, attitude towards the topic. Bühler’s (Bühler, 1934) Organon model provides a structure that distinguishes the above three types. So, we instructed annotators that they could optionally indicate the reference according to the Organon model: (S)elf reference, (O)ther reference, or (T)opic reference.

3.2.3 Results of meaning annotation

Annotators used 24 out of the 33 Baron-Cohen categories to annotate meaning. They added nine out of the 40 categories of the emotion wheel (Geneva, 2005), as well as four custom categories. The 37 categories used are shown in Table 2. The number of frequently used categories is much smaller, though.

Baron-Cohen categories	**anticipating**, cautious, concerned, confident, contemplative, decisive, defiant, **despondent**, **doubtful**, **friendly**, hostile, insisting, **interested**, nervous, playful, preoccupied, regretful, serious, suspicious, **tentative**, **thoughtful**, uneasy, upset, worried
Emotion wheel categories	**amused**, angry, compassionate, disgusted, happy, **irritated**, relieved, **scornful**, **surprised**
Custom categories	depressed, excited, ironic, outraged

Table 2: The list of categories used for annotation. Frequently used categories ($> 5\%$) are highlighted in bold, and most frequent categories ($> 10\%$) are underlined. (Pammi and Schröder, 2009)

The full descriptions of meaning are summarized in terms of meaning categories associated with types of functional reference. The results show that Baron-Cohen's affective-epistemic categories are not sufficient to describe our data – it is necessary to add a number of categories from the Geneva Emotion Wheel as well as some custom categories. The results from reference annotation according to Bühler's Organon model suggest that distinguishing the reference in addition to affective-epistemic meaning categories is a useful means to gain insights regarding a character's mood or personality (Self reference), interpersonal stance (Other reference) and attitude towards a topic (Topic reference).

A subset of 102 listener vocalizations from the non-acted part of the dialog corpus was annotated by both annotators with meaning and reference categories for inter-rater agreement. As described in (Pammi and Schröder, 2009), we computed Kappa for each meaning category and each reference type. The Kappa values for the most frequently used meaning categories friendly, interested and amused were 0.02, 0.41 and 0.82 respectively. Among the less frequent categories, Kappa values for decisive, confident, tentative, doubtful and surprised scores range between 0.22 and 0.43, whereas anticipating, thoughtful, ironic, irritated, outraged, angry show nearly no agreement between two annotators. For reference categories, there is no consistent agreement between the two annotators. It remains to be seen whether this is due to an intrinsic ambiguity or due to insufficient instructions.

3.3 Realization

A base-line system was implemented in MARY (Schröder and Trouvain, 2003; Schröder et al., 2008) Text-To-Speech(TTS) framework for synthesis of nonverbal listener vocalizations. This simple system can generate nonverbal listener vocalizations based on an XML request. It stores all nonverbal listener vocalizations in the form of datagrams in a single time-line waveform file and a corresponding unit file containing index numbers and start-end timestamps of each vocalization to retrieve efficiently. We can request a nonverbal vocalization with or without index number. When the XML request does not have an index number then the system will select any one among the vocalizations existing in the database. The baseline system was integrated to the first version of the open source SEMAINE (Schröder and et al., 2008) demonstration system for generating back-channel vocalizations when requested.

An example XML request:

```
<?xml version="1.0" encoding="UTF-8"?>
<maryxml xmlns="http://mary.dfki.de/2002/MaryXML" version="0.4" xml:lang="de">
        <voice name="spike">
                <nvv variant="6"/>
        </voice>
</maryxml>
```

4 Future Work

So far the results of the work related to database collection and meaning annotation were described. This section proposes our plans for behavior annotation, realization strategies and evaluation.

4.1 Annotation of behavior

Behavior annotation is one of the crucial tasks as this part of the work directs the way to surface level realization of nonverbal listener vocalization.

The following elements are expected in the behavior annotation:

1. A representation of intonation
2. A suitable phonetic segmental form in alignment with the waveform
3. Aspects of volume, para-language and voice quality

The intonation of a nonverbal vocalization can be extracted automatically from any pitch tracking algorithm available in computer programs like Praat (Boersma and Weenink, 2005) and can be stored as a set of points of the pitch contour or a set of polynomial coefficients which can represent the pitch contour of the nonverbal vocalization. A suitable phonetic segmental form of a nonverbal vocalization in alignment with the waveform should be annotated manually as we do not have any immediate procedure to do that automatically. The phonetic segmental form is useful for lip synchronization of the visual synthesis system, when we integrate with audiovisual synthesis system like GRETA (Poggi et al., 2005). A suitable set of descriptors should be identified to annotate aspects of volume, para-language and voice quality. A pilot study (Douglas-Cowie et al., 2003) was conducted on the Belfast naturalistic database (Douglas-Cowie et al., 2003) for the description of naturally occurring emotional speech. The descriptors, as shown in Table 3, provided from the study will be a starting point to annotate aspects of volume, para-language and voice quality.

Para-language Descriptors	Laughter, Sobbing, Break in Voice, Tremulous Voice, Gasp, Sigh, Exhalation and Scream
Voice Quality Descriptors	Creak, Whisper, Breathy, Tension and Laxness
Volume Descriptors	Raised Volume, Lowered Volume and Excessive Stressing

Table 3: A set of descriptors which are considered to be strongly indicative of emotion (Douglas-Cowie et al., 2003)

4.2 Relation between meaning and behavior

The system has to identify a suitable behavior for surface-level realization whenever the multi-modal interaction system requests a nonverbal vocalization with an intended meaning. In order to provide this functionality, we must carry out research on the relation between the meaning and the behavior of nonverbal vocalizations. The data analysis on annotated samples might provide an answer to the question whether the relation between meaning and behavior is one-one mapping pattern or any other. If the relation is having one-one mapping pattern, a simple lookup table will be able to find an appropriate behavior.

4.3 Realization with different technologies

This section outlines our plans regarding the technological realization of nonverbal vocalizations. Nowadays Unit-selection (Hunt and Black, 1996) and HMM-based (Tokuda et al., 2000; Black et al., 2007) speech synthesis technologies are the most popular. The corpus-based unit selection approach can produce near-natural high quality speech; it simply relies on runtime selection and concatenation of units from a speech database using explicit matching criteria. HMM-based speech synthesis provides an efficient model-based parametric method for speech synthesis that is based on a statistical framework of HMMs. In the scope of this work, we propose to perform experiments with both technologies to identify the pros and cons of the different technologies for the task at hand.

4.3.1 Unit-selection speech synthesis

In general, the selection of a unit at runtime is a crucial task in unit-selection synthesis framework. In MARY TTS, the unit can be a diphone or a half-phone. But here the unit is a nonverbal listener vocalization. One challenge in this framework is to find a way to choose a nonverbal vocalization with the help of behavioral properties identified from the mappings between meaning and behavior.

We can propose two possible solutions regarding the selection of a unit: One possibility could be finding a suitable nonverbal vocalization with appropriate behavior descriptors using explicit matching criteria. Another possibility could be training a classification tree to find the index of a nonverbal vocalization with a given set of behavioral properties. In the latter case, it is possible to choose a vocalization with closest but not exact behavior. Signal modification algorithms may be useful to realize exact behavior.

4.3.2 HMM-based speech synthesis

We do not yet have a clear view regarding the realization of nonverbal vocalizations in the HMM-based synthesis framework. A simple starting point would be a copy-synthesis mechanism using the MLSA (Mel Log Spectrum Approximation) filter (Tokuda et al., 2002), which would have to support external prosody specification. The sequence of steps involved in the simple proposal system is: 1. Extract Mel Frequency Cepstral Coefficients (MFCCs) of each vocalization and store them as one of the behavior properties. 2. Re-synthesize the vocalization using the MLSA filter with external prosody specification according to requested behavior.

5 Evaluation

The system will be implemented with based on the annotations of form and meaning described above and it will use all nonverbal listener vocalizations available from the dialog speech corpus. The evaluation of the system is perhaps the most significant challenge. One major objective of the system is the generation of nonverbal listener vocalizations that support effective human-computer interaction. Therefore, a subjective evaluation of the dialog system with and without support for generation of nonverbal listener vocalizations would be a promising strategy.

6 Discussion and Conclusion

The term ’nonverbal vocalizations’ does not quite describe the types of vocalizations that this work aims to cover. Not all of them are nonverbal. The listener responses like *yes*, *absolutely*, *really*, etc. are actual words that can be found in a dictionary. Several terms are considered to describe the types of vocalizations, but we did not find such single and appropriate one. For example, if the term ’epistemic vocalizations’ is taken, the term does not describe continuers’ like *mhm* or *uh-huh* since no epistemic stance seems to be involved. So the term ’nonverbal’ is a place holder for the moment. However, finding a proper term that describe the types of vocalizations in the scope of this work is an open issue.

An emotionally colored conversational synthesis system is required to synthesize not only listener nonverbal vocalizations, but also speaker’s nonverbal vocalizations with it’s context speech. For example, sentences like ’*Oh! My dear daddy*’ and ’*Wow! It is wonderful*’. Though the topic of speaker’s nonverbal vocalizations is not relevant to the discussion so far, the annotation and technological realization strategies are expected to be same as we discussed in this paper. But this topic raises another interesting question, namely how to realize behavior of nonverbal vocalization which matches the context speech. For example, do we see any similar patterns of behavior in a nonverbal vocalization (*ex: Wow!*) and it’s context speech (*ex: It is wonderful*)?. We have not yet confirmed whether the dialog speech recorded with the strategy used for data collection provides sufficient coverage of speaker’s nonverbal vocalizations as we do not have annotation for them. However, we will be able to extend this work to synthesize all kinds of non verbal vocalizations if there is no data coverage problem regarding speaker’s nonverbal vocalizations.

To conclude, the solutions identified from the proposed research work will lead us towards expressive conversational speech synthesis. The main contribution of this research work is not only providing technological solutions to generate nonverbal listener vocalizations, but also building a real-time system that can be integrated with the SEMAINE project demonstration system which is aiming to build an audiovisual SAL system.

Acknowledgments

The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 211486 (SEMAINE). Thanks to Dr. Marc Schröder for useful discussions and suggestions.

References

Baron-Cohen, S., Golan, O., Wheelwright, S., and Hill, J. (2004). *Mind Reading: The Interactive Guide to Emotions*. Jessica Kingsley Publishers, London.

Black, A. W., Zen, H., and Tokuda, K. (2007). Statistical parametric speech synthesis. In *in Proc. ICASSP, 2007*, pages 1229–1232.

Boersma, P. and Weenink, D. (2005). Praat: doing phonetics by computer. Computer program, available: *http://www.praat.org/*.

Bühler, K. (1934). *Sprachtheorie*. Gustav Fischer Verlag, Stuttgart, Germany.

Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication*, 40:33–60.

Douglas-Cowie, E., Cowie, R., Cox, C., Amir, N., and Heylen, D. (2008). The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In *Proceedings of LREC*, pages 1–4, Marrakech, Morocco.

Douglas-Cowie, E., Cowie, R., and Schröder, M. (2003). The description of naturally occurring emotional speech. In *In Proceedings of the 15th International Conference on Phonetic Sciences*, Barcelona, Spain.

Gardner, R. (2002). *When Listeners Talk: Response Tokens and Listener Stance*. John Benjamins Publishing Co.

Geneva (2005). Geneva emotion wheel. *http://www.unige.ch/fapse/emotion/resmaterial/gew.zip*, Accessed 6 April 2009.

Hunt, A. J. and Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *ICASSP '96: Proceedings of the Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings., 1996 IEEE International Conference*, pages 373–376, Washington, DC, USA.

Pammi, S. and Schröder, M. (2009). Annotating meaning of listener vocalizations for speech synthesis. In *Proc. Affective Computing and Intelligent Interaction (ACII) 2009*, Amsterdam, The Netherlands.

Poggi, I., Pelachaud, C., de Rosis, F., Carofiglio, V., and de Carolis, B. (2005). Greta. a believable embodied conversational agent. pages 27–45.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227–256.

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729.

Schröder, M., Charfuelan, M., Pammi, S., and Türk, O. (2008). The MARY TTS entry in the Blizzard Challenge 2008. In *Proc. Blizzard Challenge 2008*, Brisbane, Australia.

Schröder, M. and et al. (2008). Semaine deliverable d1b : First integrated system. *http://semaine.sourceforge.net/SEMAINE-1.0/D1b20system.pdf*.

Schröder, M., Heylen, D., and Poggi, I. (2006). Perception of non-verbal emotional listener feedback. In *Proc. Speech Prosody 2006*, Dresden, Germany.

Schröder, M. and Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech technology*, 6:365–377.

SEMAINE (2008). Semaine project page, *http://www.semaine-project.eu*.

Tokuda, K., Kobayashi, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speech parameter generation algorithms for hmm-based speech synthesis. In *Proc. ICASSP*, pages 1315–1318.

Tokuda, K., Zen, H., and Black, A. (2002). An HMM-based speech synthesis system applied to English. In *Proc. of 2002 IEEE SSW*, Santa Monica, CA, USA.

Ward, N. and Tsukahara, W. (2000). Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, 32(8):1177–1207.

Yngve, V. H. (1970). On getting a word in edgewise. In *Chicago Linguistic Society. Papers from the 6th regional meeting*, volume 6, page 567.

Toward Natural Human-Robot Interaction: Exploring Facial Expression Synthesis on an Android Robot

Laurel D. Riek [*]
Computer Laboratory, University of Cambridge
15 JJ Thompson Avenue, Cambridge, CB3 0FD, UK
Laurel.Riek@cl.cam.ac.uk

Abstract

Robots are entering domestic environments in increasing number. However, the present means for interacting with them is quite limited and difficult for people who are not technically inclined or able-bodied. Thus, many in the field of robotics are moving toward natural human robot interaction, which means allowing people the ability to communicate with robots in ways similar to how they communicate with other people, i.e., via verbal and nonverbal channels. My work focuses on a part of this problem: how to accurately model and synthesize natural human facial expressions on a realistic, human-like robot head. The main goal of this research is to see if by providing such expressions on a robot lead to people feeling more at ease while interacting with it, and thus be more likely to accept such technology in domestic settings.

Keywords: Affective Computing, Emotion Synthesis, Empathy, Human-Robot Interaction

1 Introduction

Each year, robots are entering domestic environments in greater and greater numbers. According to a 2007 report by the International Federation of Robotics, 3.4 million personal service robots are in use worldwide. The report forecasts that this number is expected to increase by 4.6 million robots by 2012 (IFR, 2008). These domestic robots are being used to serve as health aids and companions, help with household chores, and provide education and entertainment to their users.

The domestic robot user presents a unique challenge to robot designers. Elderly users are likely to be uncomfortable with domestic robots due to a lack of exposure to technology, disabled users might have difficulty using robots that do not provide interaction modalities that accommodate their needs, and people using robots for household chore assistance are unlikely to have much time to devote to learning to use complexly designed systems. One way to address some of these problems is to design robots that allow people the ability to interact with robots naturally.

Natural interaction means allowing people the ability to communicate with robots in ways similar to how they communicate with other people. This includes both verbal communication (speech and non-speech vocalization) and nonverbal communication (body gesture, gaze, movement, and facial expression). Most people are able to express themselves in this way and easily interpret such expressions in others. While people generally do not expect such ease of interaction with machines, evidence suggests having it would help improve user engagement with the robot (Sidner et al., 2005). Indeed, by taking advantage of these interactive modalities, robot designers can go a long way toward ensuring their robots are accepted.

The task of enabling natural human-robot interaction is by no means trivial; it is a complex problem that requires drawing from many areas of computer science (computer vision, human-computer interaction, natural language processing, machine learning, robotics), cognitive science, and the social sciences. It requires being able to accurately characterize how humans interact

[*] Parts of this paper have previously appeared in the following publications: (Riek and Robinson, 2008), (Riek and Robinson, 2009), (Riek et al., 2009a), (Riek et al., 2009b), (Riek et al., 2009c)

with one another so that robots can be programmed to appropriately recognize or synthesize such behaviors themselves. It also requires robots to have at least a rudimentary understanding of the social context in which they are placed.

My work tackles a small piece of this larger problem; namely, how to accurately model and synthesize natural human facial expressions on a realistic, human-like android robot head. This is an unexplored area of research; to date most emotion synthesis work on facially-expressive robots has used zoomorphic or mechanical-looking robots with limited expressivity. Furthermore, most of these robots display exaggerated, non-naturalistic, repetitive facial expressions that are generated by an actor or animator. My approach differs in that it is based on a large collection of human facial expressions made in naturalistic conversational settings, which are then normalized to create a great repertoire that the robot can use to generate novel expressions.

The main goal of this research is to see if by providing such natural, human-like expressions on a robot people feel more at ease while interacting with it. And, further, that people may be more willing to accept the presence of such a robot in domestic settings.

2 Background

The problem of natural interaction with robots is something that spans across a variety of different disciplines. It is not practical to give a detailed survey of each one; instead I will briefly introduce the two most relevant to my work: Affective Computing and Human-Robot Interaction.

2.1 Affective Computing

Affective computing is a discipline dedicated to the idea of giving machines the ability to recognize and generate affect (Picard, 1997). In some ways, the field exists to address the failings of traditional human-computer interactive systems, which typically neglect affective state changes in users. In fact, some argue that such neglect is a reason many users view interactions with computers as “cold, incompetent and socially inept.” To address this, several leaders in the field have said that it is critical that interfaces of the future are able to “detect subtleties of and changes in the user’s behavior, especially his/her affective behavior, and to initiate interactions based on this information rather than simply responding to the user’s commands” (Zeng et al., 2009).

Until recently, most of the approaches to affect recognition centered around posed data with exaggerated affective expressions, were limited to a small set of emotions (such as anger, fear, and happiness), and were restricted to single modes of expression (just face or just speech). However, the field is now shifting toward looking at recognizing multi-modal, less-constrained naturalistic expressions (Zeng et al., 2009). For example, el Kaliouby (2005) worked on the generalization of facial affect inference for complex mental states while Sobol Shikler (2007) worked on inferring affect from naturally-evoked speech. Bernhardt and Robinson (2007) worked on inferring affect from body posture and gesture.

2.2 Human-Robot Interaction

In the field of human-robot interaction (HRI), quite a number of interactive robots have been designed to try to facilitate natural interaction by recognizing and generating affect. Breazeal et al. (2008) and Fong et al. (2003) present thorough surveys of many such robots and their theoretical emotional underpinings. I will present a subset of these robots and also introduce a few others using role categories commonly used in the literature. For each category I will list the names of some representative robots, and highlight one robot in particular as an example.

2.2.1 Epigenetic Robots (Cog, HOAP-3, iCub, *Kismet*, Leonardo)

A number of interactive robots have been created with some degree of affective understanding and generation capability using an epigenetic approach. This approach uses ideas from developmental psychology to help robots learn sophisticated social behaviors (Scassellati, 1998). Many of these developmentally-based robots inherently take social context into account in order to learn to adapt to the humans interacting with them. One of the first of these robots is Kismet, an

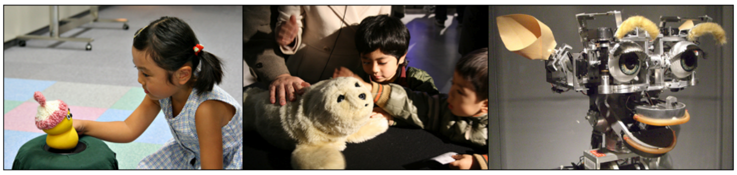

Figure 1: A few exemplar robots that recognize and generate affect. From left to right: KeepOn (*Photo: Janne Moren*), PARO (*Photo: Shoko*), Kismet (*Photo: Carol Nichols*).

anthropomorphic, expressive robot designed entirely for emotional interaction with humans. By understanding the social cues of humans in the environment, Kismet is able to respond in an emotionally appropriate way to people (Breazeal, 2002). Its thoughtful design has lead to it being a very well accepted and regarded robot.

2.2.2 Entertainment Robots (AIBO, ASIMO, AUR, *Keepon*)

Kozima and Michalowski were interested in building a robot that could interact with children in a pleasant and natural way. Their first attempt was the Infanoid robot, which was a highly mechanical-looking, very expressive robot. From observational studies the researchers found that the appearance and behavior of this robot was overwhelming children. This insight led them to the successful design of the robot Keepon, which is a minimally-designed interactive dancing robot. The robot only has 4 degrees of freedom, but is easily able to express attention via head direction and emotion via rocking motions. Its design was well informed by observing hundreds of children interacting with the robot for over 400 hours in total (Kozima et al., 2008).

2.2.3 Therapeutic Robots (Huggable, iCat, KASPAR, *PARO*, Shybot)

Shibata et al. (1997) describe their desire to build an affect robotic pet that was capable of sensing the emotions of the people it was interacting with and alter its affect accordingly. From the outset they concerned themselves with how their robot would interact emotionally with users, and tailored the robot's design accordingly. This mindset led the researchers to later create the very successful implementation of PARO the robotic seal, which has been used effectively to reduce stress and depression among the elderly (Wada and Shibata, 2007).

2.2.4 Peer Robots (Vikia, Robonaut, *Valerie*, GRACE, Mel)

Kirby et al. designed Valerie, a robot receptionist designed to facilitate long-term social interaction with people. The robot was thoughtfully designed to facilitate natural interaction - the robot's physical appearance, its station, and its behaviors were carefully considered to create an engaging experience with users (Kirby and et al., 2005).

2.2.5 Mentor Robots (*Basketball Coach*, Chips, RoCo)

Liu et al. (2006) describe a robotic basketball coach that monitored the physiological signals (heartrate and galvanic skin response) of people while they shot baskets. Depending on how anxious people seemed to be, the robot altered the game's level of difficulty. The researchers found through this style of interactive teaching people's performance improved.

3 Work To Date

3.1 Affective Centered Design Paradigm

As I began studying previous work in HRI, HCI, and Affective Computing, I came to realize that the fields had very little overlap when it came to robot design. In particular, one of the primary design methodologies espoused by interactive robot engineers, human-centered design, completely lacked an explicit affective element. And, unsurprisingly, most commercially available robots lack any sort of ability to recognize or generate affect.

Thus, I attempted to address this problem by introducing a new paradigm for the design of interactive robots called affective-centered design (Riek and Robinson, 2009). The idea behind this contribution was to provide practical guidelines to interactive robot designers to help them improve the affect aspects of their robots' designs. Furthermore, for my own research, it was helpful to think about ways in which one can measure successful interactions.

3.2 Empathizing with Robots that Mimic Expressions in Real-Time

When evaluating the affective quality of natural human-robot interaction, it is very important to consider the expression of empathy. Empathy expression is a key aspect of human-human social communication that allows people to experience and understand what others are emotionally conveying (Ross et al., 2008). One of the most basic forms of expressive empathy is known as emotional contagion, where an observer mimics the behavior of a target, and by virtue of that mimicry, comes to experience an emotional state similar to that of the target (Davis, 2006).

Facial expression and head gesture mirroring are common forms of empathic conveyance that typically include head nodding, laughing, smiling, etc. This mirroring is so vital to emotional communication that if an individual's ability to mirror others is physically blocked, that individual will actually be impaired in their ability to identify emotions (Oberman et al., 2007).

Given how important facial mimicry is in human-human communication, I wondered if it might also be important in human-machine communication. In particular, might a conversational robot that mimics a few low-fidelity expressions and head gestures in real-time create a more satisfying interactive experience for people? To address this question, we built a real-time, autonomous, head gesture mimicking robot and performed two pilot studies (Riek and Robinson, 2008; Riek et al., 2009a). In the studies, subjects sat in front of the robot and were asked to perform two conversational tasks. The first task served as an acclimation task, "Please tell the robot the route that you took to the lab today", and the second was intended to be emotionally salient, "Please tell the robot your first memories of arriving in Cambridge." Following the study subjects completed a 7-point (strongly agree - disagree), 13-question survey intended to probe emotional interaction satisfaction.

The quantitative results from these two studies were not statistically significant, however, our qualitative analyses revealed a number of valuable insights. The first surprising insight is that one ought to be cautious when making assumptions about where the robot should "look" when mimicking. We assumed subjects would primarily move their head while talking, and positioned our camera accordingly. However, some subjects never moved their head at all, but moved their head, neck, shoulders, hands, or feet quite a lot. As far as our software was concerned, those subjects were sitting still, and thus they were not mimicked.

The second insight was with regards to response appropriateness. One subject really helped to illustrate this problem by saying that she only ever told the robot positive things; thus, it was less critical for the robot to make the appropriate responses back to her. Whereas if she was saying something negative, sympathetic expressions would be far more important. This was a profound insight, because not only does an empathetic robot need to be able to make appropriate expressions, it also needs to make them at the appropriate time.

The third surprising insight is that people often behave in completely unexpected (albeit interesting) ways during an affective interaction. One subject spoke deliberately slowly to the robot. Some subjects "leaned in" to talk to the robot, or otherwise adopted odd postures. And by far the most interesting unexpected behavior was co-nodding. Two participants co-nodded with the robot; meaning, the participant nodded, the robot nodded in response, and then the participant nodded to acknowledge the robot's nod. This is something worthy of further exploration, as it seems to indicate a high level of engagement with the robot.

3.3 Empathizing with Robots of Varying Degrees of Human-likeness

In the previous studies, the kind of empathy we tested might be thought of as a first-order level of empathy - how does a subject feel after directly interacting with a robot. However, another aspect of empathy is the ability to put one's self into the shoes of another, in other words, 'simulating'

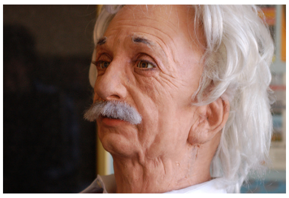

Figure 2: The Einstein robot, developed by Hanson Robotics.

their situation (Goldman, 2006). In the psychology literature, this is a well-established theory called “Simulation Theory”, and might be thought of as a kind of second-order empathy.

In robotics, a debate has been ongoing for many years regarding the degree of human-likeness a robot ought to have. Some researchers argue that robots “should look like robots”, and we ought not build androids. Others argue that we should push the envelope and try to make robots as human-like as possible. The debate is usually framed within the context of the Uncanny Valley, which is a theory proposed by Mori (1970) that posits as robots become more humanlike they become more familiar (and thus more likeable) until the mismatch between their form, interactivity, and motion quality makes people feel uncomfortable. Despite this theory remaining unproven, it is still widely cited and used as justification for robot design decisions (Draude, 2009).

Nonetheless, if Simulation Theory is correct and humans have an innate simulative system, it follows that it should be easier to empathize with the emotions and mental states of a robot that appears similar to us than with one that does not. Certainly this is true for how we interact with other humans - in-group bias and consequent referential treatment can be triggered by markers of physical similarity (e.g. skin color) (Turner, 1978); therefore, it would be unsurprising to find similar effects for how humans interact with robots.

Thus, to test second-order empathy with robots, we conducted a web-based experiment that measured how people empathized with four robots shown to be experiencing mistreatment by humans. The robots varied in appearance from not anthropomorphic (i.e., machine-like) to highly anthropomorphic (i.e., human-like). Subjects viewed both emotionally evocative and emotionally neutral film clips for each protagonist (10 in total). Following each clip, subjects were asked to rate on a scale from 1-6 how sorry they felt for the protagonist. Our quantitative results showed that indeed people empathize more strongly with human-like robots and less-strongly with mechanical-looking ones. Our detailed experimental methodology is described in Riek et al. (2009b).

4 Future Work

Thus far, I have explored the affective aspects of natural interaction with robots, from both an appearance and an empathetic interaction perspective. Next I will focus specifically on synthesizing natural, human-like expressions on an android robot. The main goal of this research is to see if by providing such expressions people feel more at ease during their interactions with the robot, and, further, if they are more willing to accept the presence of such technology in domestic settings.

The robot I will use for my work will look similar to the one depicted in Figure 2. The robot is developed by Hanson Robotics. It has 19 degrees-of-freedom in its face, and each degree of freedom is intended to represent the human facial musculature. Its movements are incredibly natural and life-like; in fact this robot is undoubtedly the most realistic facial humanoid in the world.

The robot resembles an older man, and our choice to acquire such a robot was deliberate for a few reasons. First, an older person has many more wrinkles on their face; thus on a robot it is far easier to create believable facial expressions with little effort. Second, should our future research include elder care, it may be useful to have a robot that looks elderly. Third, while we were ambivalent toward the robot’s apparent gender, it turned out that most of Hanson Robotics’ baseline robots are either male or androgynous, and we wanted to purchase one that had a well-established, believable appearance. Thus, we chose a less iconic version of the Einstein robot.

4.1 Natural Data Collection from Human Conversational Dyads

As mentioned previously, posted data is viewed by many people in the affective computing community as less useful than natural data, because it is rarely generalizable to the real world, and in daily life most people make far more subtle expressions. However, to date, most facial robots make exaggerated expressions programmed by an animator, or are otherwise based on acted data. My work aims to fill this gap, by using natural data as the basis for expression synthesis.

I will collect data in a way similar to Morency et al. (2008), in which subjects will be asked to participate in face-to-face, quasi-monologic storytelling dyads. One of the pair will be assigned to be a speaker, and the other a listener. The faces of both pairs of the dyad will be videotaped, though I will primarily be interested in data from subjects who are listeners. This is because I will be programming our robot to behave as an active listener (Bavelas et al., 2000), primarily to avoid issues with speech synchronization, well-known to cause cognitive dissonance in subjects.

4.2 Data Transformation and Normalization

Once the video data is collected, I will run each subject's video through the Neven Vision facial feature tracker. The tracker produces 31 facial feature points per video frame. Thus, for a 10-minute long video, one would expect to see about 18,000 sets of 31 points.

After the data has been transformed into feature points, I will attempt to normalize it by performing some rudimentary multivariate visualizations on it to determine what sorts of similarities exist between subjects. Once the data is adequately normalized, I will use it to train an HMM.

4.3 Novel Facial Expression Synthesis

In order to generate novel, natural-looking expressions on the robot, appropriate "paths" need to be chosen from the HMM. In computer animation, some researchers have had success with novel facial generation using Active Appearance Models (Bettinger and Cootes, 2004). It is presently unclear whether this approach will work for physical robots, since subtle oddities in behavior can be extremely noticeable, simply due to the physicality of the robot. But this approach will be a very reasonable starting point.

4.4 Evaluation

As stated previously, the main goal of this research is to see if by providing a robot with natural, human-like expressions people are more at ease during their interaction with it, and are thus more likely to accept such technology in domestic settings. This evaluation will center around the robot's expression generation capability, measured quantitatively and qualitatively using the techniques described below.

4.4.1 Round-trip synthesis

In the field of machine translation (MT) there is a technique known as "Round-trip translation". The idea is that one takes a sentence, uses an MT engine to translate it into a foreign language, and then uses the same system to translate the sentence back into the original language. For experts in the MT community this is not a well-regarded technique, but for non-experts it is a useful way to get a rough idea of what an MT engine can do.

For robot expressions, "round-trip synthesis" may actually prove to be a fruitful first evaluative step. For robotic expressions, the requirements for precision are less stringent than in the case of MT. And further, the circumstances are different: the robot is trained on human expressions, but is then making its own kind of expressions on a non-human face.

The evaluation would work as follows: after the robot has been trained to generate its expressions, it will be filmed making them. Then, the films of the robot will be run through the face tracker. Next, the HMM used to train the robot for synthesis will be instead used for recognition. Finally, the recognition rates would be examined for significance.

4.4.2 Human Labeling

In affective computing, a common evaluative technique involves humans labeling emotional corpora, and then inter-rater reliability is assessed as the primary metric for label accuracy. Afzal and Robinson (2008) describe an interface and methods to perform such annotations on natural data sets; I will likely use this interface to perform human evaluations of the robot's expressions.

4.4.3 HRI Approaches

Another possible avenue of evaluation is to employ evaluative techniques commonly used in HRI, such as common ground analysis, embodiment analysis, questionnaires, interviews, and other techniques. These techniques can be used to both survey the affective states of users and evaluate the affect generation capabilities of the robot. This type of evaluation can be performed "live", i.e., the user is placed in front of the robot, or else via video playback (the user watches pre-recorded videos of the robot).

4.4.4 Physiological and Neurological Approaches

It may be possible to use physiological and neurological approaches to assess how a user responds to (and possibly mimics) the robot's facial displays. There is a precedent for this approach in the cognitive psychology and affective neuroscience literature for measuring how humans perceive and respond to the emotional expressions of others. Researchers have successfully measured people's cardiovascular activity, skin conductance, fMRI, and electromyography in response to faces.

5 Conclusion

To date I have explored the affective aspects of natural interaction with robots, from both an appearance and an empathetic interaction perspective. Soon I will be working on the synthesis of natural, human-like expressions on an android robot. The goal is to see if by providing such naturalistic expressions people feel more at ease during their interactions with the robot, and, further, if they are more willing to accept the presence of similar technology in domestic settings. The results of this work will hopefully be useful to the affective computing community.

6 Acknowledgments

I would like to thank my supervisor, Professor Peter Robinson, for his support, and the Qualcomm Research Studentship for generously funding my research.

References

Afzal, S. and Robinson, P. (2008). An interface to simplify annotation of emotional behaviour. In *Proceedings of HUMAINE Workshop on Corpora in Emotion Research (LREC).*

Bavelas, J. B., Coates, L., and Johnson, T. (2000). Listeners as co-narrators. *J Pers Soc Psychol*, 79(6):941–952.

Bernhardt, D. and Robinson, P. (2007). Detecting affect from non-stylised body motions. In *Proc. of 2nd Int'l Conf. on Affective Computing and Intelligent Interaction (ACII).* Springer-Verlag.

Bettinger, F. and Cootes, T. F. (2004). A model of facial behaviour. In *Proc. of IEEE Int'l Conf. on Automatic Face and Gesture Recognition.*

Breazeal, C. (2002). *Designing Sociable Robots.* MIT Press, Cambridge, MA, USA.

Breazeal, C., Takanishi, A., and Kobayashi, T. (2008). Social robots that interact with people. In Siciliano, B. and Khatib, O., editors, *Springer Handbook of Robotics*, pages 1349–1369. Springer.

Davis, M. H. (2006). Empathy. In Stets, J. E. and Turner, J. H., editors, *Handbook of the Sociology of Emotions.* Springer Press, New York.

Draude, C. (2009). Who's afraid of virtual humans? In *Proc. of the AISB Symposium on The Social Understanding of Artificial Intelligence (SSoAI)*, Edinburgh.

el Kaliouby, R. (2005). *Mind-reading Machines: the automated inference of complex mental states from video.* PhD thesis, Computer Laboratory, University of Cambridge.

Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42:143–166.

Goldman, A. I. (2006). *Simulating Minds: The philosophy, psychology, and neuroscience of mindreading.* Oxford University Press, New York.

IFR (2008). World robotics 2008: Statistics, market analysis, forecasts, case studies, and profitability of robot investment. Technical report, International Federation of Robotics.

Kirby, R. and et al. (2005). Designing robots for long-term social interaction. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).*

Kozima, H., Michalowski, M., and Nakagawa, C. (2008). Keepon: A playful robot for research, therapy, and entertainment. *International Journal of Social Robotics.*

Liu, C., Rani, P., and Sarkar, N. (2006). Human-robot interaction using affective cues. In *Proc. of the IEEE Int'l Symposium on Robot and Human Interactive Communication (RO-MAN).*

Morency, L.-P., de Kok, I., and Gratch, J. (2008). Context-based recognition during human interactions: automatic feature selection and encoding dictionary. In *Proc. Int'l ACM Conf. on Multimodal Interfaces (ICMI).*

Mori, M. (1970). Bukimi no tani [the uncanny valley]. *Energy*, 7(4):33–35.

Oberman, L. M., Winkielman, P., and Ramachandran, V. S. (2007). Face to face: Blocking facial mimicry can selectively impair recognition of emotional expressions. *Soc Neurosci*, 2(3):167–178.

Picard, R. (1997). *Affective Computing.* MIT Press.

Riek, L., Paul, P., and Robinson, P. (2009a). When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. *Under Review.*

Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., and Robinson, P. (2009b). Empathizing with robots: Fellow feeling along the anthropomorphic spectrum. In *Proc. of 3nd Int'l Conf. on Affective Computing and Intelligent Interaction (ACII).* Springer-Verlag.

Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., and Robinson, P. (2009c). How anthropomorphism affects empathy toward robots. In *Proc. of the ACM/IEEE Int'l Conf. on Human-Robot Interaction (HRI).*

Riek, L. D. and Robinson, P. (2008). Real-time empathy: Facial mimicry on a robot. In *Workshop on Affective Interaction in Natural Environments (AFFINE) at the Int'l ACM Conf. on Multimodal Interfaces (ICMI).*

Riek, L. D. and Robinson, P. (2009). Affective-centered design for interactive robots. In *Proc. of the AISB Symposium on Workshop on New Frontiers in Human-Robot Interaction*, Edinburgh.

Ross, D. M., Menzler, S., and Zimmermann, E. (2008). Rapid facial mimicry in orangutan play. *Biology Letters*, 4(1):27–30.

Scassellati, B. (1998). Building behaviors developmentally: A new formalism. In *1998 AAAI Spring Symposium "Integrating Robotics Research"*. AAAI.

Shibata, T., Yoshida, M., and Yamato, J. (1997). Artificial emotional creature for human-machine interaction. In *IEEE Conference on Systems, Man, and Cybernetics.*

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., and Rich, C. (2005). Explorations in engagement for humans and robots. *Artif. Intell.*, 166(1-2).

Sobol Shikler, T. (2007). *Le ton fait la musique - Analysis of expressions in speech.* PhD thesis, University of Cambridge.

Turner, J. C. (1978). Social comparison, similarity and ingroup favouritism. In Tajfel, H., editor, *Differentiation between social groups: Studies in the social psychology of intergroup relations*, pages 235–250. Academic Press, New York.

Wada, K. and Shibata, T. (2007). Living with seal robotsits sociopsychological and physiological influences on the elderly at a care house. *IEEE T Robot*, 23(5):972–980.

Zeng, Z., Pantic, M., Roisman, G., and Huang, T. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE T Pattern Anal*, 31(1):39–58.

Endowing artificial agents with emotional autobiographical memories

Davi D'Andréa Baccan, Luís Macedo

CISUC, Department of Informatics Engineering, University of Coimbra,
Pólo II, Pinhal de Marrocos, 3030, Coimbra, Portugal
{baccan|macedo}@dei.uc.pt

Abstract

The interest in designing and developing computer-based systems to deal with human memories has been steadily increasing. These systems are likely to pervade almost all aspects of the human life and can be very useful in a variety of domains. In this context, autobiographical memories, the memories which human beings develop throughout life, play a very important role. The majority of the computational approaches to deal with autobiographical memories rely either on computational models for artificial agents or on the technologically-driven lifelogging approach. However, recent studies have been highlighting the need to move away from the latter approach. In addition, the advances in wireless networks and spatial location systems, along with the significant accomplishment of mobile, pervasive, and ubiquitous computing built the basis for the rise of the so-called pervasive user-generated content. The purpose of our work is to bring together the need for designing and developing computer-based systems for dealing with autobiographical memories in the pervasive user-generated content scenario.

1 Introduction

The interest in designing and developing computer-based systems to capture, store, analyse, retrieve, and share human memories has been steadily increasing [18], [11], [3]. These systems are likely to pervade almost all aspects of the human life and can be very useful in a variety of domains, for example, in healthcare and well being as well as in personal life. In the former, they could be used to improve the quality of life of elders as well as to serve as potential memory prostheses helping people with different kinds of brain disorders or injuries to make sense of the vast amount of information recorded during a lifetime. In the latter, they make it possible to help people to organize their multimedia collection, as well as to capture, store, and share memories, stories, and knowledge for the future generations, offering the possibility to recollect a past experience. However, these systems need to deal with a variety of challenging issues such as human memory representation, memory content, organization and retrieval, and selective attention. In this context, autobiographical memories plays a very important role. Autobiographical memories are those memories which human beings develop throughout life [31]. It combines memories about general autobiographical knowledge about ones own's life as well as a very large set of detailed mental representations of past experiences, which are known as "episodic memory". Episodic memory consists of episodes stored in temporally dated episodes, and temporally spatial relations among these episodes. It contains a vast amount of contextual (time and space), emotional, and sensory stored information associated with the past episode.

In parallel, the advances in wireless sensor networks and spatial location systems, associated with the significant accomplishment of mobile, pervasive, and ubiquitous computing have built the platform for the development of the so-called pervasive user-generated content [12], [15]. Pervasive user-generated content consists basically in novel forms of spatiotemporal annotation not restricted to the desktop in itself. In particular, it takes advantage of the increasing ubiquity of mobile devices and its services, which goes far beyond its original functionality. While for its users these devices provide interaction, connectivity, and entertainment, for researchers these devices are powerful tools to design and develop novel approaches to understand the social and physical components of the human dynamics when interacting not only with each other but also with the surrounding physical environment.

The purpose of this work is to bring together the need for designing and developing computer-based systems capable of dealing with autobiographical memories in the pervasive user-generated content scenario.

2 Background

We consider that our work brings together three major research areas namely human memory, especially autobiographical memories, emotions, and pervasive computing, especially the so-called pervasive user-generated content.

2.1 Human Memory

Basically human being possess three main types of memories namely sensory memory, short-term memory (also known as working memory) and long-term memory [2], [3].

Long-term memory is a complex mechanism and it is divided into two types: procedural memory, and declarative memory. Declarative memory is the memory of storing knowledge/facts which are consciously available. It is divided into two types: semantic memory, and episodic memory. Semantic memory refers to generalized knowledge about the external world such as facts, meanings, and understandings. The knowledge available in semantic memory derives from the episodic memory. Indeed, the interaction both with the environment and others might result in similar episodes. As those episodes accumulate, specific features (micro-details) of the episodic memory are lost, resulting in a fragment of generalized, semantic memory.

Finally, autobiographical memories are the memories which human beings develop throughout life. It combines memories about general autobiographical knowledge about one's life as well as a very large set of detailed mental representations of past experiences, which are known as "episodic memory" [31], [32]. Episodes are stored in temporally dated episodes, and temporally spatial relations among these episodes [19]. It contains a vast amount of emotional, sensory, and spatial (context and time) stored information associated with the past episode. In other words, episodic memory attempts to store information about an event in several dimensions such as "what", "when", and "where". Autobiographical memories can be considered in terms of three levels of specificity namely lifetime periods, general events, and event-specific knowledge (ESK) [4]. Lifetime periods, such as "When I was working at X", "When I was dating Y", consist of general knowledge of significant activities, locations, plans, goals, people, and so on, characteristic of a period. A lifetime period refers to distinct periods of time with identifiable beginnings and endings, although some of these may be vague rather than discrete. In addition, several lifetime periods may themselves be grouped to form a higher order theme such as "work", and "relationships". General events are more specific and heterogeneous than lifetime periods. It covers both repeated events, such as "At company X, I used to drink a coffee with my friends after lunch", and single events, such as "Our trip to A", which lasted for days up to months. General events may also represent sets of associated events and then encompass memories linked together by a theme. Event-specific knowledge (ESK), such as "Once when I was working at X, my boss...", "When me and my wife were at A, we had dinner at a wonderful place", concerns detailed information unique to single events which lasted for seconds up to hours. In general, an ESK refers to a very remarkable (sometimes with an emotional component) "situation" so that detailed information of the whole context in which it took place is stored. ESK may also be accompanied by "images that pop into the mind".

2.2 Emotions

Recent studies have been providing a better comprehension about the importance of emotions both for encoding, storing and retrieving memories [16], [6], [9]. Emotions significantly contribute to the consolidation of long-term memories, maintaining emotional episodes, whether positive or negative emotions are associated, vivid in memory.

A fundamental issue related to the studies of emotions is to make a clear distinction between the terms emotional state, emotional expression, and emotional experience [25], [1], [27]. Emotional state refers to the mental and physical state of the organism who has the emotion. The word organism is used because emotions are not a human being peculiarity, but they also exist even in animals whose physiological structure is very primitive [6]. So, once emotions occur within the organism, it cannot be observed directly. However, they can be expressed, voluntarily or involuntarily, to others by the organism through external mechanisms such as body language, facial expression, and voice. In his turn, other organisms try to infer what is the emotional state of the organism expressing his/her emotions by means of the clues provided. Emotional experience refers to everything that an “organism” consciously perceives of the emotional state.

The human emotional state can be measured considering the valence and arousal dimensions [20], [28]. The valence dimension contrasts states of pleasure with states of displeasure. It refers to the intrinsic attractiveness or aversiveness of an object, event, or situation. The attractiveness/aversiveness is also known as reward/punishment pair. For example, every interaction whether between a person itself with others and/or with the environment produces some feedback. This feedback can arouse attractiveness or aversiveness. Interactions that arouse attractiveness possess positive valence and then produces positive emotions, whereas interactions that arouse aversiveness possess negative valence and then produces negative emotions. The arousal dimension contrasts states of low arousal with states of high arousal. It expresses the degree of excitement felt and generally produces a state of arousal, in which someone’s capabilities are temporarily enhanced, for example to prepare him/her to execute a possible action in response to an imminent danger (also known as “fight” or “fly”).

The most appropriate form to measure human subjective experience is through self-reports [20]. Self-reports are sensitive to valence and arousal dimensions and tends to capture the user’s emotional states. However, self-reports made on user’s momentary emotional state tend to be more valid than are self-reports of emotions made somewhat distant in time from the relevant experience.

2.3 Pervasive User-Generated Content

User-generated content, such as wikis, and videos, comes from ordinary people who contribute by authoring, editing, and sharing data, such as opinions, recommendations, and stories. It takes advantage of the collective intelligence and mass collaboration that generally yields publicly available data.

The recent advances in mobile, pervasive, and ubiquitous computing infrastructure, techniques, and tools along with the significant improvements on wireless networks and spatial location systems, laid the foundation for the rise of the so-called pervasive user-generated content [15]. Nowadays, people can leave passive (implicit) or active (explicit) digital footprints [12]. Passive digital footprints consists in footprints that people leave through interaction with an infrastructure, such as wireless network events generated by mobile phone calls and text messages. On the other hand, active digital footprints consists in footprints that come voluntarily from the users themselves, such as when they make spatiotemporal annotations in photos, and messages. Researches are starting to investigate these digital footprints in order to record, map, and visualize the social and physical components of the human dynamics when interacting not only with each other but also with the urban environment. In addition, once digital footprints can be stored, logged, and archived, it produces a history that can be used to reveal patterns of movements, activities as well as to identify and explain phenomena. For example, the digital traces, such as geo-referenced photos and the record of wireless network events generated by mobile phone users, have been used to map, analyse and visualize the behavior of tourists [12].

Considering the scenario described above as well as the widespread and increasingly use of mobile devices, users start becoming able to leave active or passive digital footprints of their autobiographical memories by capturing, storing, retrieving, and sharing data about themselves without the spatiotemporal constraints imposed by the desktop [26]. It would allow the creation of a powerful and abundant digital record of one's life history that can be preserved for future generations [5].

2.4 Related Work

Although there exists a variety of research on autobiographical memory namely in neuroscience and psychology, only in the last years computer scientists took notice of this topic. We can classify the current computational approaches related to autobiographical memories in two main categories.

The first category refers to the creation of computational models of autobiographical memories for artificial agents (namely Deutsch, Ho, Nuxoll, Tecuci, and Dodd). However, these efforts present some limitations. Although the work of Deutsch [7] incorporates emotions, this architecture does not take semantic memory into account. Ho [13] and Nuxoll [22] approaches do not take emotions nor semantic memories into account. Tecuci [30] also does not take emotions into account. Finally, the design of the system proposed by Dodd [8] is domain dependent.

The second category of computational approaches related to autobiographical memories refers to research on capturing, recording, and sharing different kinds of personal digital data (namely "Lifelogging", and Nold). A lifelogging system [29] is essentially a technological apparatus that aims to passively capture all one's digital media, including documents, images, sounds, and videos, such as the MyLifeBits [11] and Lee [18]. However, the lifelogging approach generally requires the use of post processing techniques to mine the vast amount of gathered data in order to extract meaningful information [10]. In addition, recent studies have been demonstrated that it is necessary to move away from the "capturing everything" approach to empower the user with the possibility of deciding what is of interest and so what needs to be captured and recorded [23], [24]. The starting point of Nold's book [21] is the description of the Bio Mapping project. This ongoing project investigates the implications of creating technologies that are able to record, visualize, and share people's intimate body-states. From a practical point of view, it consists of a portable and wearable tool that records data from a simple biometric sensor that measures the galvanic skin response and a GPS. The union of these devices make it possible to know the level of physiological arousal at a particular moment and space, offering the possibility of being able to record person's emotional state and visualize it in the form of an emotional map.

3 Ph.D. Proposal

Our work involves the creation of a computational model of autobiographical memory that incorporates emotional information and its respective causes, as well as the design and development of a user-friendly tool for allowing human agents to encode, retrieve, visualize, and share episodes. We consider that this model will provide the foundation for the development of a variety of novel applications such as the improvement of artificial agents possessing autobiographical memories and the spatiotemporal visualization of episodes according to their emotional classification.

We envision some research questions related to our work namely: (1) Can own's autobiographical memories help him/her in taking future decisions?; (2) Can other's autobiographical memories help him/her in taking future decisions?; (3) Is it possible to understand under which circumstances people feel positive or negative emotions?; (4) Can all autobiographical memories, especially the emotional information, be used for other purposes such as urban planning and traffic management?

3.1 Proposed Architecture

To enhance computational episodic memories with emotions, we follow the work of Nuxoll [22], and Tecuci [30], who pointed out some general requirements that any episodic memory system should implement namely three major phases which consists of encode, storage, and retrieval. As a result we propose the cognitive architecture [17] illustrated in Figure 1. The Autobiographical processing module is the core of the agent architecture. It consists of three processing modules namely the Encode module, the Storage module, and the Retrieval module. To store the information, we consider the use of the long-term procedural memory and the declarative memory. The declarative memory is, in turn, consisted of the semantic memory and the episodic memory.

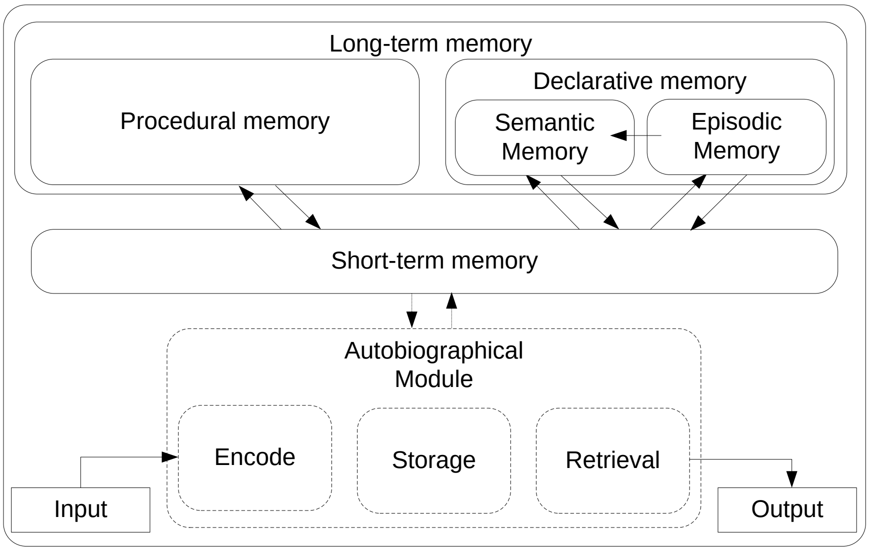

Figure 1: Proposed architecture for an agent with autobiographical memory

The Encode module is essentially the interface between the human agent and the artificial agent. We aim to design and develop a user-friendly personal digital assistant (PDA) tool for allowing human agents to annotate his/her emotions and its respective causes through self-report. Considering we want to encode, store, and retrieve micro-details related to the episodes, we propose an episode structure based mainly on the ESK (Event-Specific Knowledge) level of specificity described earlier. Therefore, in our context, an episode consists of one or more single events (an event is anything the human agent judges of interest). To capture the episodes, we follow the recent studies [23], [14] against the passively "capture everything" approach used by lifelogging systems. In our work, human agents can deliberately and voluntarily determine when an episode/event starts and ends, if an episode/event is related to a discrete or continuous date and time, and determine where an episode/event happens, and if it is related to a specific place or area.

The Storage module deals with the episode structure and dynamics. To store the episodic memory data, we propose structuring events in two components namely the contextual component and the emotional component, illustrated in Figure 2. The contextual component contains relevant contextual information related to the event. It is defined as a tuple $\langle Wa, Wo, We, Wr\rangle$, where Wa can be a simple textual description of the whole event, Wo is a set of "objects" which appropriate represents those who were related to the event; We is a start-end date and time representation; Wr is a set of spatial coordinates (latitude and longitude) which indicate where the event happened. The emotional component is defined as a tuple $\langle E, C\rangle$, where E is a set of pairs of real numbers in which the first number $\in [-1.0, 1.0]$ refers to the arousal, -1.0 denotes maximum negative arousal (sleepiness), and 1.0 denotes maximum positive arousal (excitement), and the second number $\in [-1.0, 1.0]$ refers to the valence, -1.0 denotes maximum negative valence (displeasure), and 1.0 denotes maximum positive valence (pleasure). C is a set of "multimedia objects" referring to the cause(s) of the felt emotion(s). The Storage module is also responsible for the process of consolidation and forgetting. We propose the consolidation process can be carried out considering the emotional classification of each episode/event provided by the human agent in terms of valence and arousal. For example, episodes/events classified with a certain level of arousal can be considered remarkable ones. Forgetting refers to the accessability and availability of memory. However, we currently do not consider forgetting as a relevant process for our work. Finally, the Storage module generates the semantic memory from the episodic memory. For example, based on a certain amount of episodes with positive emotions related to "going out for a dinner at restaurant A", the generalized knowledge "the human agent do like eating at A" can be yielded.

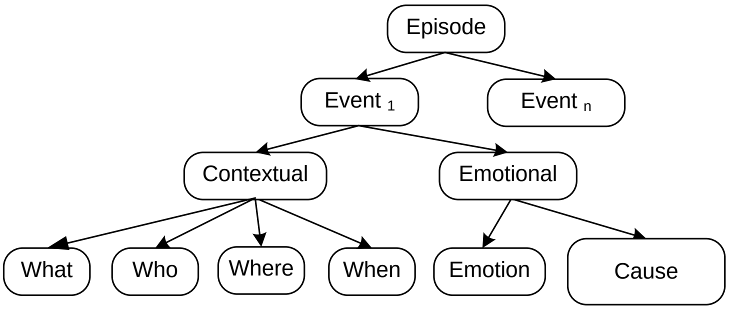

Figure 2: Proposed episode structure

The Retrieval module deals with the process of retrieving a given episode/event for further use. An episode/event can be retrieved both from contextual and emotional information that was previously stored. It provides the basis for the execution of a variety of "queries" such as, "in which spaces(places) I have felt positive emotions in the last week", or "which emotions I felt when I was accompanied with person A".

3.2 Validation

First of all, we need to evaluate the usability of the proposed PDA tool. Although we currently do not intend to focus on some particular audience, we are aware that it would be necessary to take into account some particular features depending on the chosen audience. In addition, we need to identify what is the "best" manner of using the PDA tool in the real world environment. We will invite a small group of users to carry out some preliminary tests with a prototype version of the PDA tool. A possible test scenario might be the realization of some tasks, such as to invite the users to visit a well known tourist attraction and to foster them to use the tool during the visit. We aim to identify which points need to be modified through observation in situ, interviews, questionnaires, and personal communications.

Moreover, we need to assess if the knowledge of autobiographical memories contributes in helping to take future decisions. To collect the data, we can perform comparative studies between groups empowered with the PDA tool and groups without it. We can create a set of predefined tasks, such as "visiting place A", "meeting at B", "go from A to B", and allocate the same task to both groups to create a similar and comparable scenario. For example, based on certain amounts of episodes with positive emotions related to "go from A to B", the artificial agent "living" in the PDA can identify and recommend that it is a good route in terms of pleasure (e.g. maybe there is a good landscape to be appreciated) but it is not the shortest route. Also, the human agent can visualize all autobiographical memories related to such route in the emotional map. Maybe the human agent have experienced episodes with positive emotions but the autobiographical memories of the others human agents pointed out (or are pointing out in real time) episodes with negative emotions. In addition, human agents not empowered but the PDA tool might forget if it was an episode with positive or negative emotions as well as they do not have access to the autobiographical memories of all human agents nor can visualize them in the emotional map. We also consider performing tests on some specific audience, such as people with episodic memory impairment (EMI) or elders.

Finally, we need to evaluate if all autobiographical memories, especially the emotional information, can be used for other purposes such as urban planning and traffic management. We will make the global emotional maps available for local authorities. We consider they will be able to identify and understand phenomena such as "a car accident that has just blocked a street", as well as to reveal some hidden phenomena such as "people have fear of walking on street A after 8:00 PM because there are some suspects in the area".

3.3 Expected Scientific Contributions

As a result we expect the following scientific contributions: (1) a new autobiographical model that incorporates emotional information and its respective causes; (2) a user-friendly personal digital assistant (PDA) tool, with an appropriate HCI, for allowing human agents to annotate his/her emotions and its respective causes through self-report; (3) a novel spatiotemporal visual representation of emotions; and (4) an architecture with an autobiographical memory module for an artificial agent that represents an human and can contribute to improve the quality of life of its "owner" by contributing to perform some tasks better.

References

[1] Boehner, K., DePaula, R., Dourish, P., and Sengers, P. (2007). How emotion is made and measured. *Int. J. Hum.-Comput. Stud.*, 65(4):275–291.

[2] Canadian Institute of Neuroscience, M. H. and Addiction (2009). The brain from top to bottom. http://thebrain.mcgill.ca.

[3] Committee, U. K. C. R. (2008). Road map for memories for life research. Technical report.

[4] Conway, M. A. and Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review*, 107(2):261–288. PMID: 10789197.

[5] Czerwinski, M., Gage, D. W., Gemmell, J., Marshall, C. C., Perez-Quinones, M. A., Skeels, M. M., and Catarci, T. (2006). Digital memories in an era of ubiquitous computing and abundant storage. *Commun. ACM*, 49(1):44–50.

[6] Damasio, A. (2000). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harvest Books, 1 edition.

[7] Deutsch, T., Gruber, A., Lang, R., and Velik, R. (2008). Episodic memory for autonomous agents. In *Human System Interactions, 2008 Conference on*, pages 621–626.

[8] Dodd, W. and Gutierrez, R. (2005). The role of episodic memory and emotion in a cognitive robot. In *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, pages 692–697.

[9] Faghihi, U., Fournier-Viger, P., Nkambou, R., Poirier, P., and Mayers, A. (2009). How emotional mechanism helps episodic learning in a cognitive agent. *0901.4963*.

[10] Gemmell, J., Bell, G., Lueder, R., Drucker, S., and Wong, C. (2002). MyLifeBits: fulfilling the memex vision. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 235–238, Juan-les-Pins, France. ACM.

[11] Gemmell, J., Lueder, R., and Bell, G. (2003). The MyLifeBits lifetime store. In *Proceedings of the 2003 ACM SIGMM workshop on Experiential telepresence*, pages 80–83, Berkeley, California. ACM.

[12] Girardin, F., Calabrese, F., Fiore, F. D., Ratti, C., and Blat, J. (2008). Digital footprinting: Uncovering tourists with User-Generated content. *IEEE Pervasive Computing*, 7(4):36–43.

[13] Ho, W. C. (2005). *Computational Memory Architectures for Autobiographic and Narrative Virtual Agents*. PhD thesis, University of Hertfordshire.

[14] Kalnikaite, V. and Whittaker, S. (2007). Software or wetware?: discovering when and why people use digital prosthetic memory. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 71–80, San Jose, California, USA. ACM.

[15] Krumm, J., Davies, N., and Narayanaswami, C. (2008). User-Generated content. *IEEE Pervasive Computing*, 7(4):10–11.

[16] LaBar, K. S. and Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nat Rev Neurosci*, 7(1):54–64.

[17] Langley, P., Laird, J. E., and Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, 10(2):141–160.

[18] Lee, M. L. and Dey, A. K. (2008). Lifelogging memory appliance for people with episodic memory impairment. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 44–53, Seoul, Korea. ACM.

[19] Macedo, L. M. M. L. (2006). *The Exploration of Unknown Environments by Affective Agents.* PhD thesis, University of Coimbra.

[20] Mauss, I. . and Robinson, M. . (2009). Measures of emotion: A review. *Cognition and Emotion*, 23:209–237.

[21] Nold, C. (2009). *Emotional Cartography - Technologies of the Self.* Softbook.

[22] Nuxoll, A. M. (2007). *Enhancing intelligent agents with episodic memory.* PhD thesis, University of Michigan.

[23] Petrelli, D., van den Hoven, E., and Whittaker, S. (2009). Making history: intentional capture of future memories. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 1723–1732, Boston, MA, USA. ACM.

[24] Petrelli, D., Whittaker, S., and Brockmeier, J. (2008). AutoTopography: what can physical mementos tell us about digital memories? In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 53–62, Florence, Italy. ACM.

[25] Picard, R. W. (1997). *Affective computing.* MIT Press.

[26] Pschetz, L. (2008). histories: supporting user generated history. In *CHI '08 extended abstracts on Human factors in computing systems*, pages 3693–3698, Florence, Italy. ACM.

[27] Reisenzein, R. (2007). What is a definition of emotion? and are emotions mental-behavioral processes? *Social Science Information*, 46(3):424–428.

[28] Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1178, 1161.

[29] Sellen, A. J., Fogg, A., Aitken, M., Hodges, S., Rother, C., and Wood, K. (2007). Do lifelogging technologies support memory for the past?: an experimental study using sensecam. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 81–90, San Jose, California, USA. ACM.

[30] Tecuci, D. G. (2007). *A Generic Memory Module for Events.* PhD thesis, The University of Texas at Austin.

[31] Tulving, E. (1985). *Elements of Episodic Memory.* Oxford University Press, USA.

[32] Tulving, E. (2002). Episodic memory: from mind to brain. *Annual Review of Psychology*, 53:1–25. PMID: 11752477.

Non-verbal behaviour and attribution of mental states

Sylwia Hyniewska
Télécom Paristech
37 rue Dareau, 75014 Paris, France.

Université de Genève
40, bld du Pont d'Arve, 1205
Genève, Suisse.
hyniewska@telecom-paristech.fr

Susanne Kaiser
Université de Genève
40, bld du Pont d'Arve, 1205 Genève,
Suisse.

Catherine Pelachaud
CNRS -Télécom Paristech
37 rue Dareau, 75014 Paris, France.

Abstract

The project has two aims: the study of mental state attributions to previously perceived non-verbal behaviours and the contribution to the non-verbal communication skills of embodied agents.
For the first task, short audio-visual clips presenting a person in a face-to-face context with another human have been evaluated, through a forced-choice questionnaire. The questionnaire was based on the appraisal theory's items and on the attribution of emotional labels. The appraisal theory enables the understanding of mental states in terms of successive evaluations and enables to predict the link between facial expressions and mental state attributions. In our case it will be used to predict the intuitive answers of participants to the observed expressions.
The second task is to improve the existing communicational capacities of 'embodied conversational agents'. Facial muscle movements will be transposed to a virtual agent, Greta. The interface of this agent will be used as a tool to validate the observations established in the first task. This application of the theory will enable us to run perceptive studies to verify and detail some aspects of the results. What is more, the behaviours verified by perceptive tests as expressing some particular mental states will be added to the agent's repertory, increasing the number of varied facial behaviours.

Keywords: Non-verbal communication, Facial expression, Appraisal theory, Embodied conversational agents.

1 Aims of the research

The main objective of the research is to explore the mental states attributions in relation with observed facial expressions. This study has two parts.
First, an observational study is lead on the perception of non-verbal communication cues. In a judgment task, participants observed a human in interaction with another one. They had to answer a questionnaire and evaluate the mental states of observed people. Questions related to cognitive evaluations of a situation as defined by the componential appraisal theory [1] and to the attribution of emotional labels enabled to judge the intuitive interpretation of observed expressions. Videos were coded with the FACS technique [2] by a certified coder. This coding based on the analysis of facial muscle movements and completed by an analysis of torso movements, enables to associate attributions of mental states to the perceived behaviour.
The second task is to transpose the action units described in the FACS system unto a virtual agent and to create a facial expression repertoire for the agent. These complex expressions, elaborated and defined in the first part of the research from human interactions will be described in a way to be usable by the agent. Precisions will be given concerning the evaluations and situations in which these expressions could be selected.
Expressions synthesised on the agent will be judged by participants. The attributions of mental states to expressions will be evaluated through the same paradigm as in the first study with humans. The duration, intensity and sequencing of facial action units will be manipulated and the impact of these manipulations on mental state attributions will be evaluated. This will enable to clarify the impact of the different elements contributing to the perception of internal states by others.
A particularity of this work lies in the evaluation of behaviour occurring in natural settings. Unlike the majority of studies which rely on videos of actors or other "acted" or stereotypical behaviour, the video corpus used for the study comes from a hidden camera. What is more, the situation is an emotional one, as it shows passengers declaring their

loss of luggage at an airport [3][4]. The use of such a corpus enables the synthesis of natural and non-stereotypic facial movements for virtual agents.

2 State of the art

2.1 Facial expressions

Although emotion has been investigated in relation to the voice, e.g. [5][6][7][8], body movements [9][10], posture [32] as well as more generally in multimodal expressions [12], the major weight of research focuses on the perception of emotion from the face. Facial expressions are a key element in the human communication, contributing to its efficiency by transmitting a large quantity of information [13]. According to some researchers [14][15] expressions can be seen as « rudiments of adaptive behaviour, which have acquired important signaling characteristics » [14]. Thus the expressive behaviour is a socially influenced message, prone to be regulated, as well as a true externalisation of internal states.

Since Charles Bell wrote about the intimate relation between the states of the mind and the body expressions (1844, cited by [16]) and Duchenne [17] has investigated ways in which individual muscles contribute to the perceived facial changes, a great deal of attention has focused on how emotions are communicated through facial expression.

2.2 Emotion theories

According to Scherer, emotional states are « almost always accompanied by a motor expression component » [12]. The readability of the body actions and poses enables to infere the internal states and attitudes of the person [12]. Among different emotion theories, two important approaches have been proposed, that diverge in their understanding of the emotions triggering and expression. Both propose explicit predictions for emotion-specific facial expressions, while one conceives emotions as categorical, the other as componential entities.

- discrete emotion theories [18][19][20][21] focus on a small number of so called basic emotions, in particular, anger, fear, joy, disgust, sadness, happiness, shame, and guilt. These are considered to result from innate neuromotor programs and to produce a fixed behavioural response. This expressive response is unitary in nature, emotion-specific and universally recognised
- componential appraisal theories to emotion [1][12][22][23][24] on the other hand stipulate that the individual elements of facial expressions (the micro-expressions) are determined by the appraisals of a given situation.

2.2.1 Discrete emotion theories

According to the discrete emotion theoreticians emotions are triggered by automatic mechanisms, such as neuromotor affect programs. These programs are believed to act independently from cognitive evaluations (e.g. [19]).

Studies following this approach focus on a few prototypical patterns. Tomkins [21], for instance, described these affect programs as leading to some expressive patterns specific to particular emotions. The number of these “basic emotions” is limited. As regards facial expressions, discrete (basic) emotion theory states that they are direct displays of internal states and that the ability to decode them in term of basic emotions is innate [25] and quick, thus considered unconscious [19].

2.2.2. Componential emotion theories

Cognitivist theoreticians following the componential approach to emotion counter the concept of discrete emotions resulting from automatic and biologically fixed programms [14][15][26]. They advance that the variability and complexity of emotions can be understood without any reference to basic emotions. According to those theoreticians, there is a great number of very differentiated emotional states that are captured by the labels only through a process of grouping of different states, through some kind of averaging and central tendencies. Scherer names these “averaged” states “modal emotions” (see for example [1]). The appraisal's theory's predictions for these modal emotions are the same as the ones suggested for basic emotions.

What is specific to the componential appraisal approach is to prone that emotions are the of cognitive evaluations (appraisals). An emotional state would result from the significance given to different elements of an event. An emotion is not defined and triggered directly by a situation or a stimulus, but depends from the relation established between a person and the surrounding environment. This relation is created through appraisal (for a review see [14]). Although this process relies on a succession of evaluations, it is important to note that some authors emphasise that this cognitive evaluation can happen in an automatic, fast and non conscious manner (see [24]).

2.3 Facial expression predictions

Scherer's model states a direct link between expressions and the underpinning appraisals (eg. [1]). According to Scherer, an emotion is composed of successive evaluations of a stimulus and of different, interconnected and synchronised (motor and physiological) changes that are linked to the sequence. This sequence is defined by a series of checks enabling the evaluation of a stimulus, whether it is internal or external. Micro facial expressions are associated to each appraisal check. These micro-expressions are described in terms of minimal facial muscle movements, that is facial Action Units (AU) as described by Ekman and colleagues in the Facial Action Coding System (FACS; [2]). The direct link between facial expressions and evaluation checks is described through AUs, each AU being attributed to a specific outcome of an evaluation check.

Thus a facial emotional expression is a superposition of particular micro-expressions resulting from all the successive evaluation checks (novelty, intrinsic pleasantness, goal conduciveness, coping potential, norm compatibility, etc.)

It is interesting to note that applying these appraisal checks predictions to conceive internal states enables to generate a multitude of different response combinations, leading to a great number of possible expressions. What is more, given that some evaluation combinations would be more frequent than others it would lead to the so called “modal emotions” enabling to predict prototypical facial expressions.

2.4 Embodied conversational agents

The constant development of communication technologies leads to an increasing need for virtual intelligence agents able to cope with affective aspects of interactions. Such affective abilities could facilitate the interaction with the user, but also enable contextually more adapted responses [27] and be perceived as more trustworthy [28].

Studies have shown that human interactions with virtual characters are similar to those developped with real humans [20][29][30]. This enables to hypothesise that a human emotion model can be used to modelise artificial emotions.

In order to be valid, such a model needs to include the external state of the agent, that is their expression of emotions, as much as their internal state [31][32]. Such an internal state could be defined for example by an input file defining “the general tendency of an agent” at a give time (communicativity, speed of responding, mood, etc.) or be deduced by a “dialogue system”.

Today, several research teams work on the elaboration of embodied conversational agents with interaction capacities. Embodied conversational agents are software entities with a virtual humanoid appearance. Some agents can express emotions or other internal states' characteristics.

A large number of teams uses the « OCC » model [33], which became a reference for emotion synthesis. Some agents internal states models rely on this computational model and others are inspired by it. According to this model, there are three types of emotions, triggered by the evaluation of three aspects: consequences of events, actions of agents and aspects of objects. Thus, an event enabling the realisation of a goal leads to joy, an agent's action not corresponding to the agent's convictions leads to shame and the perception of an aversive object leads to disgust. The “Virtual Hulman” agent developed by the Geneva MiraLab uses a model inspired by the OCC for the determination of the facial expressions of the agent. This embodied agent can interact with a user and have a basic conversation while expressing emotional states [34]. Although it has a complex internal state model with 24 emotions defined, its expressivity is restricted to the six basic emotion expressions.

Another approach, based on the discrete emotions theory, has been used by Bui [35] who uses a set of fuzzy rules to determine the blending expressions of the six basic emotions [18]. In Bui's work the fuzzy inference determines the degree of muscle contractions of the final expression, as defined by an input emotion intensity.

Some researchers, such as Paleari and Lisetti [31] or Malatesta et al. [36], on the other hand, have focused their work on the temporal relation between different facial actions predicted by the appraisal theory (e.g. see [1]). In [31] the different facial parameters are activated at different moments and the final animation is a sequence of several micro-expressions linked to cognitive evaluations. Malatesta and colleagues [36] have also used sequences predicted by Scherer's componential theory to create manually emotional expressions [34]. Differently from Paleari and Lisetti's work [34], in [36] each expression is derived from the addition of a new AU to the former ones. What is more, the authors [36] compare the additive approach with the sequential one, and find a greater recognition rate for the animations realised with the first approach [31].

It is in continuation of such appraisal componential approaches that the present studies are situated.

3 Problem definition

According to the appraisal theory, the emotional state of an individual is the result of successive appraisals, such as the novelty detection or the coping potential. According to Scherer [1] the same emotional label could be attributed to internal states characterised by slight variations in the outcome of appraisal checks. Thus, the facial expression could not always be linked directly to the emotional label, given that one label can be attributed to differentiated states/appraisal outcomes. It is the result of each appraisal check that is hypothesised to be directly related to particular facial movements, while expressions of one emotion would be more diversified.

We want to verify this hypothesis, by observing the link between seven appraisal attributions, action units (AU) and seven emotional labels. We have chosen the appraisal checks for which concrete facial action predictions have been formulated in the theory: 1) suddenness, 2) goal obstruction, 3) relevance (this is novel and important), 4) coping (mastery over situation) , 5) no coping, (no mastery over situation) 6) violation of internal standards and 7) violation of external standards (this is unfair and immoral).

For example, suddenness is linked to raised eyebrows and goal obstruction is linked to pressed lips. A more limited number of expressive patterns associated with appraisal checks is expected than with emotional labels.

We expect to confirm Scherer and Ellgring's predictions [1] for the five first appraisals. For the internal standards violation, no hypothesis is formulated because of a lack of unanimity in between researchers which all observe different facial changes [11].

The link between appraisal and emotions is also explored. We chose emotions directly associated to some studied appraisals [1]: control over the situation linked to anger, the lack of control to joy, and so on. Relief has been added to counter the bias of a unique positive emotion. Although there are no predictions for this emotion, it has been chosen as it is of particular interest for the virtual agent used.

In each clip, we expect that great attributions of some emotional labels will be linked to significative attributions of the corresponding appraisals, as predicted by theory. However we expect to observe some emotional label attributions without the presence of all contributing appraisals, keeping in mind that by definition emotional terms describe a group of similar but not identical states.

The intensity of emotional labels is expected to be higher for the more prototypical expressions, called modal emotions expressions by Scherer and having the same definition/characterisitcs as Ekman and Friesen's observations for basic emotions [18].

4 First study: preliminaries

In the first study, the facial expressions perceived in the videos by the participants have been evaluated through a questionnaire based on sequential evaluation checks [1] and completed with an attribution of emotional labels. For 24 video clips facial movements are being coded with the FACS system [2]. The facial action units are to be associated with the particular attributions, in order to show the link between AUs, appraisals and emotional labels. The study will be completed with a coding of body movement and postures, as well as with a basic acoustic analysis of the extracts where vocal and/or body activity is observed.

4.1 Material

Extracts from a video corpus [3][4] presenting images from a hidden camera have been selected. They present a face-to-face interaction, where passengers report they have lost their luggage to a hostess. The behaviour of the passenger is evaluated. The original clips from the corpus were one minute long, however as mental state changes appeared during that interval the clips had to be cut into segments presenting one state only per extract. Three experts were asked to evaluate each extract and annotate in time all the perceived internal states, whether these be appraisals, emotions, action tendencies or any other internal changes.

In the case of ambiguity, e.g. when one out of the three evaluators considered less changes in a clip than the other two and suggested a longer extract of a unified state, the clip was cut in a restrictive manner and the ambiguous segment was left out. The final extracts shown to the participants were from 5 to 50 seconds, with a majority falling in the 20-28 seconds.

4.2 Procedure

The study was realised individually on computers, guidelines being clearly provided on screen. The
The participants watched and evaluated from 6 to 42 short video clips, depending on their concentration level and their willingness to participate. After each video participants had to answer the same set of questions.

4.3 Hypotheses

An effect of observed/coded facial action units is expected on the attributions. Predictions are formulated for the facial expressions of appraisals (see Table 1) based on the appraisal theory [1]. We expect to observe patterns between attributed emotions and the attributed appraisal checks, as predicted in [1] (see table 2).

Characteristics of the event faced in the clip (evaluated by labelers)	**Facial Action Units expected to be observed in accordance**
Sudden	1+2
Goal attainment obstructive	17+23; 17+24
Relevant and discrepant	4, 7, 23, 17, gaze directed
Control and high power	4, 5 or 17, 23, 25 or 23, 24
Control but no power	1, 2, 5, 26, 20, 38, gaze directed+averted+directed
« Unfair »/external standards violated	10
Internal standards violated	14

Table 1. Expected association between AU[1] and cognitive evaluations. Predictions based on Scherer & Ellgring [1].

	Suddenness	Mastery	Goal obstruction	Relevance	Internal standard violation	External standard violation
Fear	YES	NO	YES	YES		
Joy		YES	NO			
Anger		YES	YES			YES
Sadness		NO	YES	YES		
Shame					YES	
Contempt						YES
Relief						

Table 2. Expected association between emotions and appraisal checks. Predictions based on Scherer (1999).

The appraisal attributions being given on a 7 point scale (from 0=totally disagree to 6=totally agree), an average significantly above 4 confirmed the presence of an appraisal and an average significantly below confirmed the absence of an appraisal in a given clip.
The participants had also to judge if the observed person was showing anger, fear, joy, sadness, shame, contempt and relief. Each emotion being evaluated on a scale from 0 (no emotion) to 6 (strong emotion), an appraisal is considered strongly attributed when the average >2.5 and significantly >1 (as the lower bound of the significance interval .95).

[1] Action Unit explanation:
1 + 2: brows raise 4: brows lower 5: upper lid raise 7: lids tighten 10: upper lip raise 14: dimpler 17: chin raise 20: lips stretch 23: lips tighten 24: lips press 25: lips part 26: jaw drop 38: nostril dilate

4.4 Partial Results

82 male computer science students evaluated 6-42 videos (mode= 13). Each extract was evaluated at least by 20 participants.

Each clip has its own pattern of attributions, which enables us to associate these to different AU patterns.

Given the fact that the number of clips coded until now is limited, the analysis of only one clip will be described here: clip 103b, the number being taken from the original corpus (103) and the letter added after splitting the clip into extracts containing only one internal state (b).

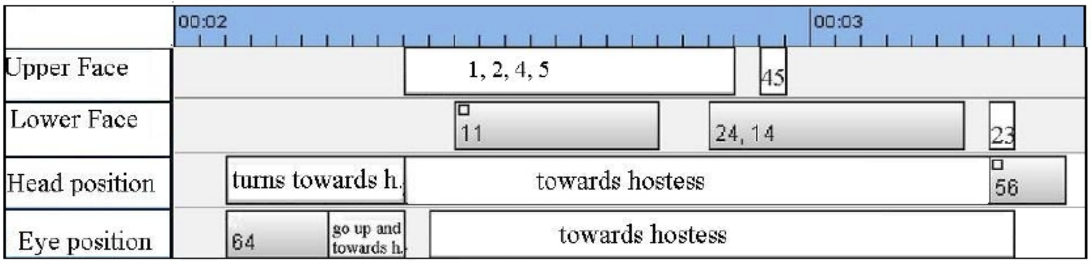

Figure 1. Summarised FACS[2] coding of clip 103b.

After having watched clip 103b, participants attributed mastery/power (average= 6.133 ; conf. interval 5.585-6.682), contempt (2.27 ; 1.61-3.73) and anger (1.6 ; .579- 2.621). No external standards violation has been attributed.

4.5 Result Interpretation for one clip

As expected, one observes an attribution of 'power' associated to the presence of lowered eyebrows and raised upper lids (AU 4+5). The presence of dimples (AU 14) is observed but it does not lead to an attribution of internal standards violations that was predicted by Scherer [1]. Here it could be associated more with some kind of external standard violations, as contempt has been attributed. These results show an analogy to Kaiser's study, which shows that dimples appear in situations of “injustice” [6].

In 103b dimples are linked to a pressure of the lips (AU 14+24) and are preceded by the deepening of the nasolabial furrow (AU 11). Such a pattern (AU 14+24) was observed by Michel and Unz [37] to be associated with contempt, as in our case. The authors deal with spectators' facial expressions during violent TV shows and observe frequent AUs 14+24, AU 14 and AUs 14+17 (dimples with a raised chin).

To summarise, the 103b expression is introduced by upper face AU (brows raised and pulled a little together, with upper lid raised) that are maintained for a second, followed by a marked blink (AU 45). The lower face expressions start a little bit later, however still during the activation of the first upper face units. AUs 14+24 are preceded by a deepening of the nasolabial furrow (AU 11) and followed by tensed lips (AU 23).

It is surprising to find a strong attribution of contempt but no significant attribution of external standard violation, although in theory this appraisal is one of those constituting the emotion of contempt in its modal expression. What is more, contempt being strongly attributed (average sign. >1.6) one could expect an association with a prototypical expression, with the predicted AUs and appraisals, which is not the case.

5 Continuation of the study

The majority of the clips still needs to be coded and the AUs patterns associated with the appraisal and emotional labels attributions.

To go further in the study of the link between behaviour and mental state attributions, different evaluations will be performed. First, some basic vocal analysis will be performed (MFCC, fundamental frequency, energy). In the extracts only few vocalisations occur and we want to control that these have no impact on the emotion and appraisal

[2] Action Unit explanation:
1: inner eyebrow raise 2: outer eyebrows raise 4: brows lower 5: upper lid raise 11: nasolabial furrow deepener 14: dimpler 23: lips tighten 24: lips press 45: blink 56: head tilt right

perception. Moreover, AU patterns determined in the first study will be synthesised with the Greta agent and evaluated using the same questions set as in the human video study. Then some modifications will be applied to the expressions and evaluated consequently. For example, the duration and the co-articulation of different AUs may be modulated and some AUs removed from the observed sequences to check if these are relevant for users' understanding of agents' states. This should clarify which aspects of agents' facial and body movements are essential for behaviour synthesis.

6 References

[1] Scherer, K. R. & Ellgring, H. (2007). Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion,* 7, 1, pages 113-130.

[2] Ekman, P., Friesen, W., Hager, J. C. (2002). *Facial Action Coding System*. Salt Lake City: Research Nexus.

[3] Scherer, K. R. & Ceschi, G. (1997). Lost luggage: A field study of emotion-antecedent appraisal. *Motivation and Emotion*, 21, pages 211–235.

[4] Scherer, K. R. & Ceschi, G. (2000). Criteria for emotion recognition from verbal and nonverbal expression: Studying baggage loss in the airport. *Personality and Social Psychology Bulletin.* 26, 3, pages 327–339.

[5] Scherer, K. R., Banse, R., Wallbott, H.G. & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, 15, pages 123-148.

[6] Grandjean, D. et al. (2005). The voices of wrath : brain responses to angry prosody in meaningless speech. *Nature Neuroscience*, 8, 2, pages 145-146.

[7] Schuller, B. (2002). Towards intuitive speech interaction by the integration of emotional aspects, SMC 2002, *IEEE International Conference on Systems, Man and Cybernetics*, Yasmine Hammamet, Tunisia.

[8] Belin, Fillion-Bilodeau & Gosselin (2008). The Montreal Affective Voices: a validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods*.

[9] Wallbott, H. (1998). Bodily expression of emotion. *European Journal of Social Psychology*, 28, pages 879-896.

[10] Pollick, F., Paterson, H., Bruderlin, A., Sanford, A. (2001). Perceiving affect from arm movement. *Cognition*, 82, pages 51- 61.

[11] Coulson, M. (2007). Attributing Emotion to Static Body Postures: Recognition Accuracy, Confusions, and Viewpoint Dependence. *Journal of Nonverbal Behavior,* 28, 2, pages 114-139.

[12] Scherer, K. R. & , Ellgring, H. (2007). Multimodal Expression of Emotion. *Emotion*, 7(1), pages 158–171.

[13] Mehrabian, A. (1971). Silent messages, Wadsworth, California: Belmont.

[14] Kaiser, S. & Wehrle, T. (2001). Facial expressions as indicators of appraisal processes. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.). *Appraisal processes in emotions: Theory, methods, research, pages* 285-300. New York: Oxford University Press.

[15] Kaiser, S. & Scherer, K. R. (1998). Models of 'normal' emotions applied to facial and vocal expressions in clinical disorders. In W. F. Flack, Jr. & J. D. Laird (Eds.). *Emotions in Psychopathology*. New York: Oxford University Press.

[16] Darwin, C. (1872/1998). *The expression of the emotions in man and animals* (3rd Ed.) New York: Oxford University Press.

[17] Duchenne, G. (1999). *The mechanism of human facial expression*. R. A. Cuthbertson, Ed. & Trans, Cambridge, England: Cambridge University Press.

[18] Ekman, P. & Friesen, W.V. (1975). *Unmasking the Face. A guide to recognizing emotions from facial clues.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

[19] Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. Cole (Eds.), *Nebraska Symposium on Motivation 1971*, Lincoln, NE: University of Nebraska Press, Vol. 19, pages 207-283.

[20] Reeves, & B. Nass, C. (1996). *The media equation: how people treat computers, television, and new media like real people and places.* Cambridge: Cambridge University Press.

[21] Tomkins, S. S. (1982). Affect theory. In P. Ekman (Eds.), ***Emotion in the human face,*** pages 353-395. Cambridge: Cambridge University Press (2nd ed.).

[22] Roseman, I. J., Smith, C. A. (2001). Appraisal Theory. In: K. Scherer, A. Schorr, T. Johnstone (Eds.). *Appraisal Processes in Emotion: Theory, Methods, Research,* Oxford: Oxford University Press, pages 3-19.

[23] 35a Turner T. J. & Ortony A. (1992). Basic emotions : can conflicting criteria converge ? *Psychological Review,* 99, 3, pages 566-571.

[24] Sander, D., Grandjean, D, Kaiser, S., Wehrle, T., & Scherer, K.R. (2007). Interaction effects of perceived gaze direction and dynamic facial expression: Evidence for appraisal theories of emotion. ***European Journal of CognitivePsychology,*** 19, 3, pages 470-480.

[25] Izard, C. E. (1994a). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. ***Psychological Bulletin,*** 115, pages 288-299.

[26] Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? ***Psychological Review,*** *97*, pages 315-331.

[27] Adler, R.S., Rosen, B. & Silverstein, E.M. (1998). Emotions in negotiation: How to manage fear and anger. ***NegotiationJournal,*** 14, pages 161-179.

[28] André, E., Rist, T., van Mulken, S., Klesen, M., Baldes, S. (2000). The Automated Design of Believable Dialogues for Animated Presentation Teams. In: Cassell et al. (Eds.): Embodied Conversational Agents, Cambridge, MA: MIT Press, pages 220-255.

[29] Schilbach et al. (2006). Being with virtual others : neural correlates of social interaction. ***Neuropsychologia***, 44, pages 718-730.

[30] Brave, S., Nass, C., Hutchinson, K. (2005). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. In *International Journal of Human-Computer Studies*, 62, pages 161–178.

[31] Paleari, M., Lisetti, C. (2006). Psychologically grounded avatars expressions. In: First Workshop on Emotion and Computing at KI 2006, 29th Annual Conference on Artificial Intelligence, Bremen, Germany. Mehrabian, A. (1971). Silent messages, Wadsworth, California: Belmont.

[32] Ochs, M., Niewiadomski R., Pelachaud, C. & Sadek, D. (2006). Expressions intelligentes des emotions. *Revue en Intelligence Artificielle RIA*, Special Edition "Interaction Emotionnelle", Vol. 20, pages 4-5.

[33] Ortony, A.; Clore, G. L.; and Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press. Malatesta, L., Raouzaiou, A., Karpouzis, K., Kollias, S.D. (2009). Towards modeling embodied conversational agent character profiles using appraisal theory predictions in expression synthesis. *Appl. Intell.*, 30 (1), pages 58-64.

[34] Egges, A, Kshirsagar, S. & Magnenat-Thalmann, N. (2002). Imparting Individuality to Virtual Humans, *First International Workshop on Virtual Reality Rehabilitation (Mental Health, Neurological,Physical, Vocational)*, Lausanne, Switzerland, pages 201-108.

[35] Bui, T.D. (2004). *Creating Emotions and Facial Expressions for Embodied Agents*. PhD thesis, University of Twente, Departament of Computer Science.

[36] Malatesta, L., Raouzaiou, A., Karpouzis, K., Kollias, S.D. (2009). Towards modeling embodied conversational agent character profiles using appraisal theory predictions in expression synthesis. *Appl. Intell.,* 30 (1) pages 58-64.

[37] Michel, B. & Unz, D. (2008). She smiles, he frowns? Does gender matter in the emotional processing of TV news? *12th European Conference on Facial Expressions*, University of Geneva, Switzerland.

Neurophysiological assessment of affective experience

Christian Mühl
Human Media Interaction,
University of Twente,
PO Box 217, 7500AE Enschede,
The Netherlands.
muehlc@cs.utwente.nl

Abstract

In the field of Affective Computing the affective experience (AX) of the user during the interaction with computers is of great interest. The automatic recognition of the affective state, or emotion, of the user is one of the big challenges. In this proposal I focus on the affect recognition via physiological and neurophysiological signals. Long-standing evidence from psychophysiological research and more recently from research in affective neuroscience suggests that both, body and brain physiology, are able to indicate the current affective state of a subject. However, regarding the classification of AX several questions are still unanswered. The principal possibility of AX classification was repeatedly shown, but its generalisation over different task contexts, elicitating stimuli modalities, subjects or time is seldom addressed. In this proposal I will discuss a possible agenda for the further exploration of physiological and neurophysiological correlates of AX over different elicitation modalities and task contexts.

Keywords: Emotions, Brain-Computer Interfaces, Affective Computing, Affective Signal Processing

1. INTRODUCTION

In this proposal I will outline the objective and structure of my PhD research project "Neurophysiological assessment of affective experience". In the project I will explore how different affective states are reflected by physiological and neurophysiological indicators. Furthermore, I will study possible methods to automatically determine these states from the measured sensor modalities. Finally, the use of automatic affect classification will be explored in various gaming scenarios. I will focus on electroencephalography to assess neurophysiological activity, and on cardiovascular indicators, galvanic skin response, facial muscle activity and respiration to assess the state of the autonomous nervous system.

To clarify my motivation I will start the proposal by a very short introduction to emotion research, traditional approaches of affect assessment and a discussion of alternatives to the use of (neuro-) physiological sensors, pointing out the problems with this undertaking and the results obtained so far. In the second part I will then proceed with a concise review of physiological and neurophysiological measurements associated with certain affective states and of studies exploring the use of those features for the classification of emotions. Thereby I will detail a number of unexplored questions, such as the generalisation of classifiers over context, time and subjects. The third part will then outline the methods that I propose to approach these issues. Finally, first results and observed problems from a preliminary analysis of a first experiment will be presented to provide a starting point for the discussion of the research project proposed.

1.1 DEFINITION OF THE SUBJECT AND MOTIVATION

There are several reasons for the study of emotional processes and their physiological and neural correlates. One of the most important is certainly the purely scientific interest that already motivated studies in psychophysiology and more recently in affective neuroscience. This fundamental curiosity is spurred by questions about the nature of emotions. As example it should suffice to mention the old debate about the structure of emotions. Here the debate between proponents of a small number of discrete emotions and those that propose a continuous affective space, spanned by two or more affective dimensions, such as arousal, valence, dominance, stretches already over decades.

With the development of the relatively new domain of Affective Computing [26] another source of interest to the study of emotion has been established apart from the purely theoretic interest, adding a perspective on possible applications regarding our knowledge about affective experience. Here one of the main interests lies in the prediction of subjective experience and the development of an objective measure of experience. The aim of this domain is to enrich the interaction with applications by the automatic recognition of the affective user state. Further areas of application are the evaluation of experience in application or product design [23].

Before reviewing some studies of (neuro-) physiological correlates of affective states and their automatic recognition we will briefly discuss traditional approaches to emotion assessment and motivate the use of physiological and neurophysiological signals for this purpose.

1.2 AUTOMATIC CLASSIFICATION VS TRADITIONAL METHODS TO MEASURE AX

The reliable measurement of affective states has been a known problem in the domains of psychology and psychophysiology for a long time. To assess emotional states and their interaction with other observations, researchers have developed several methods. One of the most widely used is the self-report, including techniques such as the Mood Adjective Checklist, the Profile Mood States, the Expanded Form of the Positive and Negative Affect Schedule, and the Differential Emotions Scale. A comprehensive overview can be found in [15]. Those self-reports are asking the subject (1) to observe and quantify their emotional experience, and (2) to honestly and accurately report it. These requirements are susceptible to several error sources. Firstly, retrospective reports can be distorted by effects of recency and duration neglect. Secondly, the subject's answers might be influenced by social desirability. Affect can also be measured by think aloud techniques and behavioural observations [23]. Think aloud techniques have the advantage to avoid the problems related to retrospective self-reports. However, they are unnatural and carry a high potential for distraction of the subject. Observations, e.g. by video recordings to code gestures, body language and verbalisations, are a rich data source, but the analysis is a time-consuming process not free of biases.

The advantages of an objective measure of the affective state via physiological signals are therefore obvious. It would be a continuous measure of the affective state of a subject, avoiding disturbances, and the risk of distortions known for retrospective reports. Furthermore, the ability to report emotions varies between subjects [30]. Objective measurement might allow one to assess emotions independent from the subject's ability to describe them. This would be interesting for basic and applied sciences. However, it is important to recognise that subjective reports remain in most cases the basis for the calibration of potential affect recognition systems as they provide the labels for the construction of physiological evaluation methods and the training of classifiers. Therefore, to gain from the advantages of physiological measures of affect, the problems innate to subjective reports have to be taken into account. Having motivated the search for an objective measurement of affective state we will now explore the advantages and disadvantages of the chosen sensors.

1.3 WHY TO USE PHYSIOLOGICAL AND NEUROPHYSIOLOGICAL SENSORS

I am focusing in this proposal on (neuro-) physiological attempts to classify affective experiences. While there is a multitude of physiological signals that might give valuable information about this experience, I explore mainly indicators of peripheral (e.g. muscle tension, galvanic skin response, and cardiovascular signals) and central nervous system functioning (i.e. electroencephalogram). Other physiological indicators of affective experience, e.g. visual and auditory analysis of subject behaviour, have also yielded promising results [10]. However, the use of (neuro-) physiological sensors has some advantages over those other, less intrusive sensors, as cameras or microphones. Firstly, one may observe state changes that are too subtle to find utterance in behavioural indicators. And, secondly, they appear to be less susceptible to voluntary deceiving. Thus, they are more robust indexes of affective state, as the behaviour can be easily manipulated, but the body state is harder to control [27].

The disadvantages, on the other hand, include the intrusiveness just mentioned, as it is yet not possible to measure (neuro-) physiological signals without the application of sensors to the body. Furthermore, there are variations of recording conditions, resulting from differences in hand washing, gel application, and sensor positions, which might lead to variations in the recordings themselves.

Another methodological problem arising in the study of (neuro-) physiological correlates of affective states is associated with the acquisition of the ground truth. While those studies that are centred on auditory and visual behaviour accompanying emotions can employ observer ratings to obtain the ground truth of a given sample, this is not possible when behaviour cannot be observed [27].

Despite the problems I believe that only the thorough understanding of emotional responses in body and brain can lead to robust classification of affective states. While there might be less intrusive alternatives for the observations of affective behaviour, the sensor technology in the focus of this proposal delivers important insights in otherwise non-observable facets of emotional states. Thereby valuable potential for the disambiguation of affective states might be gained. The next section will give an overview over physiological measurements and correlates of affective processes and first approaches of their classification. It will also reveal some hitherto unexplored issues that are in the focus of the proposed research.

2. RELATED WORK

In this section I will first show that the measured sensor modalities, physiological and neurophysiological, have been associated with emotional processes by several studies. Then I will discuss studies exploring their value for the training of automatic classifiers.

2.1 PHYSIOLOGICAL CORRELATES OF AFFECTIVE PROCESSES

The current project focuses on a limited range of physiological measurements to assess the state of the peripheral nervous system. Electromyography (EMG) assesses the somatic nervous system by the measurement of muscle activity. While the relation between muscle tension and emotional arousal seems straightforward, Cacioppo et al. [5] showed that facial EMG can differentiate valence *and* arousal. Magnée et al. [22] showed that the facial muscle activity reflects emotional processes, and is not just the result of facial mimicry.

The eccrine glands involved in the production of Galvanic skin response (GSR) are innervated by the sympathetic branch of the autonomous nervous system. GSR is therefore thought to be a trustworthy measure of sympathetic activation. Emotional arousal, for example, was shown to robustly express itself in the GSR response for affective pictures [21, 9] film clips [13], environmental sounds [4], and musical pieces [17]. However, it was found that emotional arousal in an erotic context is associated with a parasympathetic activation and thus not reflected by GSR.

Cardiovascular measures, like heart rate, blood pressure, blood volume and blood flow, assess the function of the heart and vasomotor activity. Lang et al. [21] measured electrocardiography (ECG) during the presentation of low and high arousal pictures. They found an increase of heart rate with emotional arousal. However, heart rate seems to be no robust measure, as exercise, stimulus, and design characteristics showed very influential on the heart rate. No heart rate increase was found for briefly presented arousing pictures or sounds [9]. And Aftanas and colleagues [1] measured even a decrease of heart rate for arousing pictures. Sammler et al. [29] observed a decrease for negatively valenced stimuli, but not for positive stimuli (distorted versus non-distorted music pieces). Heart rate variability (HRV), on the other hand, is supposed to be a more robust measure, at least in terms of exercise. From the three frequency ranges HRV can be divided into, the highest, respiratory sinus arrhythmia (RSA), which is associated with the effects of respiration on the heart rate, is supposed to be a good indication of parasympathetic activation. Interestingly, Frazier and colleagues [13] observed a decrease of RSA with increasing emotional arousal. Respiration has also been more directly associated with affective manipulations in different contexts [12, 14].

In general it might be said that it is difficult to make clear predictions of the response of a certain physiological measurement after affective stimulation. Autonomic response patterns can be highly subject- and context-specific. Nevertheless, some studies succeeded in identifying robust patterns for given experimental contexts. In the next section we will discuss neurophysiological responses associated with different affective states.

2.2 NEUROPHYSIOLOGICAL CORRELATES OF AFFECTIVE PROCESSES

In the last twenty years many studies explored the neurophysiology of emotions. The findings are manifold and complex. This might be due to differences in stimulus characteristics, experiment design, subject populations and finally the complexity of emotional responses in the brain. Most EEG studies on affect have focused on event-related responses to emotional stimuli. Those correlates are valuable for the spatial and temporal localisation of emotional processes, but only of limited value for the classification of affective states in a natural environment. Thus we will focus here on findings in the frequency domain, first those for valence and then for arousal.

According to the "hemispheric valence hypothesis" positive approach-related emotions are mainly processed in the left frontal cortex. Negative withdrawal-related emotions, on the other hand, are processed in the right frontal cortex. As activity in cortex is inverse to alpha activity, this translates to a decrease of alpha power in the right frontal cortex for positive emotions and in the left frontal cortex for negative [11]. However, Mueller and colleagues [25] did not find hemispheric interactions with valence for alpha or beta bands. Instead they found a left hemisphere increase of gamma activity for negative valence and a right hemisphere increase of gamma activity for positive valence. A role for fronto-medial theta power in emotional processes was suggested by Sammler et al. [29]. They observed an increase of theta for positive valenced music pieces.

Arousal, on the other hand, is supposed to activate neural structures, and therefore to decrease the overall level of alpha power and increase power in beta and gamma bands. Choppin [8] found this general alpha power decrease and a parietal beta power increase for arousing picture stimuli. Marosi et al. [24] showed an overall decrease of low alpha power for emotional sentences. However, Aftanas et al. [1] found an increase in the occipital low alpha band synchronisation for arousing picture stimuli. Furthermore, Keil and colleagues [16] found enhanced gamma responses toward arousing pictures. A similar gamma increase was shown by Mueller et al. [25] for emotional compared to neutral stimuli.

2.3 Classification of (neuro-)physiological correlates of affective processes

In the last sections plenty of evidence for the existence of (neuro-) physiological correlates of emotional processes was presented. This section will discuss studies that explore the automatic classification of affective states via these correlates. Physiological signals were classified with very high accuracy rates. Benovoy et al. [3] showed that four different emotions, elicited by method acting and visualisations, could be separated by their physiological characteristics with an accuracy of 90%. Kim et al. [18] even achieved for four musically induced affective states an average classification ratio of 95%. Neurophysiological signals, on the other hand, only reached relatively low classification accuracies. The differentiation between three emotional states (calm neutral, excited positive, excited negative) via neurophysiological signals was with an accuracy of 67% possible, for two classes with 76 - 79% accuracy [6]. Comparable rates were achieved for arousal classification in single subjects using affective pictures to induce emotions [7]. The direct comparison of physiological and neurophysiological studies is rather difficult as they apply different methods for the affect elicitation and different signal intervals (minutes vs. seconds).

Once the subject-specific distinguishability of different emotional states is shown the question arises, whether a classifier can be generalised to function for different subjects. In the above-mentioned study of Kim et al. [18] the authors also trained a subject-independent classifier that differentiated emotional states by physiological signals with an accuracy of 70%. Kim and colleagues [19] trained a classifier on the physiological signals of 50 children aged between seven and eight years. They achieved correct classification ratios of 61% and 78 % for four (sadness, anger, stress, and surprise) and three (all but surprise) classes, respectively. To date no study compared subject-specific and general classifiers build from neurophysiological features. However, Choppin [8] found a great variation between subjects' EEG responses co-occurring with affective states, suggesting a strong subject-specificity of neurophysiological emotional responses.

Another aspect of classifier generalisation to consider is whether the patterns of affective states that are collected over time can be used to obtain a stable classifier. In other words, are the potential day-to-day variations in our physiological responses accompanying affective states hindering the training of reliable classifiers? This question was addressed indirectly by two studies. Picard and colleagues [27] collected the physiological data over several weeks from the same subject. A classifier that was trained on the data discriminated among eight classes of emotion with an accuracy of 81%. Interestingly, this high accuracy was achieved despite a great variation between the daily measurements, probably originating from differences in sensor-placement and daily background mood.

Also Kim et al. [18] collected their physiological data over a period of several weeks. As already reported they recognised four affective states with 95% and 70% accuracy with a subject-dependent and -independent classifier, respectively. The high classification ratios show that autonomic nervous system patterns that co-occur with affective states are stable enough to train time-independent classifiers on them. Again no EEG study invested the generalization of a classifier over several days or weeks.

Summarising, it might be said that there is strong evidence for robust (neuro-) physiological features associated with affective processes. Classification studies employing physiological measurements achieved high rates and showed that physiological patterns of affective states generalise over subjects and time. While classification of affective states was shown for neurophysiological measurements, no study to date explored its generalisation over subjects and time. Furthermore, most studies were conducted in a limited and controlled experimental context. This is especially true for the few EEG studies and leaves several fundamental questions about the viability of (neuro-) physiological affect classification open: (1) generalisation over different stimulus modalities and (2) over different contexts. The next part will introduce the current state of the discussed research project and the broad agenda, which is intended to deal with at least some of the mentioned issues.

3. CURRENT AND PLANNED RESEARCH

In the last section several unanswered questions regarding the generalisation of affect classification were presented. Here I will outline approaches that are suited to answer some of the questions to a certain degree. The first experiment is already conducted and the data currently analysed. It will also be in the focus of the discussion section. Therefore the first experiment will be described in more detail, while the other ideas will only be sketched.

3.1 ELICITING AFFECT VIA MULTIMODAL STIMULI

In a first experiment the influence of different affect eliciting modalities was studied. On the one hand the question is how well the classification of affect generalises over different elicitation contexts, in this case auditory or visual. On the other hand I am interested in the classification of the affect eliciting modality itself. To study the effects that the different modalities have on neurophysiological affective responses, 180 multimodal stimuli were combined from the auditory and visual affective stimuli sets IADS and IAPS. The stimuli were combined in a way that allowed a grouping into auditory negative and positive, visual negative and positive, and multimodal neutral. The auditory negative consisted of a negative auditory stimulus and a neutral visual stimulus. The auditory positive group contained positive auditory and neutral visual stimuli. This way the affect elicitation was supposed to result from the auditory stimulus. Correspondingly, the visual negative and positive stimuli were created from a neutral auditory and a valence-holding visual stimulus. The multimodal neutral stimuli consisted of a neutral auditory and a neutral visual stimulus. This group was important as a control group, which enables the analysis of the specific effects of positive and negative stimulation, respectively. While the grouping is based on the distribution of the stimuli on the valence axis, the group differences on the arousal axis are kept comparable to avoid confounding effects. To assess the effect of the stimuli on the participant's affective state, a self-assessment manikin (9 point Likert scale) for the dimensions of valence (SAMv) and arousal (SAMa) is used. Each trial had therefore the following sequence of presentations: (1) a pre-stimulus period of 2 seconds only showing the fixation cross, (2) 6 seconds in which the multimodal stimulus is presented, (3) 2 seconds in which only the fixation cross is presented, (4) SAMv, (5) SAMa, and (6) 5 seconds of black screen. Thus a trial took about 20 seconds.

In case clear features and good classification results are yielded, the experiment could be extended by a temporal dimension. The repeated recording of participants over a longer period can inform about the generalization of (neuro-) physiological pattern over time, i.e. how robust a classifier would work after some time is passed.

3.2 ELICITING AFFECTIVE STATES VIA COMPUTER GAMES - PACMAN

For the elicitation of emotions in a more natural (task) context a modified version of the game Pacman was used [28]. The experiment described in that paper induced frustration by manipulation of the game flow and control responsiveness. Other manipulations of game speed or opponent intelligence might induce states of boredom, flow, and stress, to assess (neuro-) physiological differences between them and explore methods for their classification.

3.3 ASSESSING AFFECTIVE STATE VIA PROBE STIMULI – TESTING A N400 APPROACH

Another approach to emotion recognition could be based on the response of the central nervous system to certain probe stimuli. For example Allison and Polich [2] use probe stimuli to assess work load in video games via P300 ERP analysis. This probing is made possible by a negative relationship between attention response (indicated by P300 amplitude) and work load. For the analysis of emotional responses I plan to explore the N400 ERP. This potential is associated with semantic processing and appears when a (probe) stimulus semantically mismatches a preceding stimulus. Koelsch et al. [20] have shown that the effect of the N400 occurs in very broad contexts, for example when words expressing moods are succeeding excerpts of classic music. This might qualify the N400 response for the assessment of affective experience in movies or games by the exclusion of non-matching states.

3.4 APPLYING AFFECT CLASSIFICATION AND NEUROFEEDBACK APPROACHES IN GAMES

To study the dynamics of conscious and unconscious affective feedback and its effect on the user experience in a computer game I will use the Pacman game just mentioned, which will be controlled in a traditional manner and by affective neurofeedback. The affective state of the user thereby integrated by a game logic to manipulate one or more parameters of the game. As a simple feature representing affective state one can use alpha activity over central or frontal electrodes. However, the described research on (neuro-) physiological correlates of affect is supposed to reveal further potential methods and features to determine the affective state of the user. These will provide more precise and robust means of affect classification.

4. Discussing relevant methodological issues

I would like to use the discussion to bring some specific issues related to my first experiment to the attention of the committee. First I will lay out problems related to the grouping of the trials using the self-assessment method. Then I will address the conflicts for experiment design introduced by the use of physiological *and* neural measurements. Finally, the issue of temporal precision in the analysis of neurophysiological features will be discussed.

As already mentioned, the self-assessment is the basis of the classifier training, as it determines the ground truth for the training samples. Analysis of the mean stimulus valences suggested that the emotion induction did work (Table 1 left). The mean values behaved according to the group membership (N+: positive, Nn: neutral, N-: negative). However, for many stimuli the induced emotions differed from the emotions the stimuli were supposed to induce. This was also reflected by participants' reports after the experiments. For example, a starving African child on a blue blanket was perceived as cared for and elicited a calm and rather positive response, while it was intended to elicit a negative reaction. These deviations from the original grouping of the stimuli are natural taking the individual differences between participants into account. Those differences are already reflected in the standard deviations that characterise the (norm) ratings of the individual stimulus sets, IAPS and IADS. Therefore, the overall grouping of trials into the three emotion classes, independent from the elicitating modality, is not straightforward. Grouping the trials according to the ratings, we observe a trend towards the middle, thus toward the neutral class (S1n), while the positive (S1+) and negative (S1-) classes are underrepresented in the data (Table 1 upper right). However, by assuming each rating that deviates from the middle of the Likert scale by one scale unit towards one end of the scale to result from a negative or positive affective response, the S2 grouping is obtained (Table 1 lower right). Here the responses are equally distributed over all 3 classes (S2+, S2-, and S2n), as the neutral class is narrowed down to 1 Likert point. Of course, this results in smaller differences of the mean valences between the emotion conditions.

Group	Valence mean (std)	Arousal mean (std)
N+	5.29 (1.58)	4.01 (1.84)
Nn	4.49 (1.35)	3.75 (1.86)
N-	3.04 (1.70)	5.02 (2.03)
S1+	6.79 (0.64)	3.71 (1.94)
S1n	4.4 (0.74)	3.76 (1.78)
S1-	1.79 (0.75)	5.88 (1.59)
S2+	6.22 (0.81)	3.79 (1.89)
S2n	4.47 (0.13)	3.41 (1.70)
S2-	2.37 (0.98)	5.34 (1.79)

	N+	Nn	N-	sum
S1+	270	104	38	412
S1n	386	519	313	1218
S1-	60	97	369	526
sum	716	720	720	2156
S2+	426	216	92	734
S2n	162	308	140	610
S2-	128	196	488	812
sum	716	720	720	2156

Figure 1. The left table shows the mean and standard deviations of the self-assessments for the emotion conditions according to the different grouping methods (N, S1, S2). The right table shows the relations between the intended grouping of trials and the grouping according to the 3-point sized neutral class (upper right) and the 1-point sized neutral class (lower right).

The combination of auditory and visual stimuli complicates the matter further. In the experiment I constructed the emotional stimuli from one emotion inducing part and one neutral part. A preliminary analysis suggests that the planned fragmentation of the trials into groups of 30 stimuli for each emotion-modality pair does not coincide with the self-assessments. Therefore the direct comparison of the effect of an eliciting modality for a certain emotional reaction, positive or negative, is not viable if the self-assessment data is used as ground truth. However, the comparison of the effect of the eliciting modality of emotional reactions in general might still deliver valuable information. Eventually, a classification of an emotional reaction might not only be helped by a better understanding of modality specific parts, but might guide the determination of potential emotion eliciting events.

A more general problem of all approaches to correlate neurophysiological and physiological measurements to affective processes are the different requirements of the recording methods. The low signal-to-noise ratio of EEG signals requires many trials of affective responses. As neurophysiological processes are fast paced and relatively short this high number of trials can be gained by a fast-paced presentation of affective stimuli. However, the response of physiological sensors, as ECG or GSR, is rather sluggish and needs several seconds to start and several more to reach its peak. Thus, both sensor groups have conflicting requirements. A fast-paced presentation of many affective stimuli is also contraindicated by a potential habituation of the participants to the emotional content of the stimuli. A longer and more intense induction of affective states might be a viable EEG paradigm if the lengthy trials could be split in subtrials, resulting in a higher number of features per condition. However, the features extracted in this fashion could be highly dependent or, on the other hand coincide with different consecutive cognitive processes. Furthermore, the analysis of neurophysiological responses toward affective stimuli is often focused on the few 100 milliseconds after stimulus presentation. This makes sense in the context of the analysis of fast affective processes that follow stimulus presentation. The component process theory [30], for example, assumes that affective responses are a chain of evaluation processes conducted partially automatically within very short intervals. Each of these processes is supposed to have its own neural signature. Therefore, an evaluation of neural responses that takes several seconds into account sees only a rough summation of different response correlates, instead of the individual affective responses. However, without the use of probe stimuli as is the case in natural gaming scenarios, no information about onset of the affective process is available and thus an analysis with temporal precision impossible.

5. Conclusion

In this proposal I motivated and outlined a project for the study of (neuro-) physiological correlates of affective experience and their classification. It was shown that physiological and neurophysiological measurements can inform about affective states and are suited for automatic classification approaches. However, several open questions regarding the (task) context-, (elicitation) modality-, subject-, and time-independent classification of affective experience were detailed. I proposed three approaches to investigate the (neuro-) physiological correlates over different elicitation-modalities and task context.

Acknowledgements. The authors gratefully acknowledge the support of the BrainGain Smart Mix Programme of the Netherlands Ministry of Economic Affairs and the Netherlands Ministry of Education, Culture and Science.

References

[1] Aftanas, L. I., Reva, N. V., Varlamov, A. A., Pavlov, S. V., and Makhnev, V. P. (2004). Analysis of evoked EEG synchronization and desynchronization in conditions of emotional activation in humans: temporal and topographic characteristics. *Neuroscience and behavioral physiology*, 34(8), pages 859-867.

[2] Allison, B. Z. and Polich, J. (2008). Workload assessment of computer gaming using a single-stimulus event-related potential paradigm. *Biological psychology*, 77(3), pages 277-283.

[3] Benovoy, M., Cooperstock, J. R., and Deitcher, J. (2008). Biosignals analysis and its application in a performance setting - towards the development of an emotional-imaging generator. *IEEE International Conference on Bio-Inspired Systems and Signal Processing*, pages 253-258, Funchal, Portugal.

[4] Bradley, M. M. and Lang, P. J. (2000). Affective reactions to acoustic stimuli. *Psychophysiology*, 37(2), pages 204-215.

[5] Cacioppo, J. T., Petty, R. E., Losch, M. E., and Kim, H. S. (1986). Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. *Journal of personality and social psychology*, 50(2), pages 260-268.

[6] Chanel, G., Asl, K. A., and Pun, T. (2007). Valence-arousal evaluation using physiological signals in an emotion recall paradigm. *In Proceedings of the IEEE SMC and International Conference on Systems, Man and Cybernetics, Smart cooperative systems and cybernetics: advancing knowledge and security for humanity*, pages 2662-2667, Montreal, Canada.

[7] Chanel, G., Kronegg, J., Grandjean, D., and Pun, T. (2005). Emotion assessment: Arousal evaluation using eeg's and peripheral physiological signals. *Technical report*, Computer Vision and Multimedia Laboratory, Computing Centre, University of Geneva, Switzerland.

[8] Choppin, A. (2000). EEG-based human interface for disabled individuals: Emotion expression with neural networks. *Unpublished master's thesis*. Information processing, Tokyo institute of technology, Yokohama, Japan.

[9] Codispoti, M., Bradley, M. M., and Lang, P. J. (2001). Affective reactions to briefly presented pictures. *Psychophysiology*, 38(3), pages 474-478.

[10] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1), pages 32-80.

[11] Davidson, R. J. (1992). Anterior cerebral asymmetry and the nature of emotion. *Brain and Cognition*, 20(1), pages 125-151.

[12] Etzel, J., Johnsen, E., Dickerson, J., Tranel, D., and Adolphs, R. (2006). Cardiovascular and respiratory responses during musical mood induction. *International Journal of Psychophysiology*, 61(1), pages 57-69.

[13] Frazier, T. W., Strauss, M. E., and Steinhauer, S. R. (2004). Respiratory sinus arrhythmia as an index of emotional response in young adults. *Psychophysiology*, 41(1), pages 75-83.

[14] Gomez, P., Shafy, S., and Danuser, B. (2008). Respiration, metabolic balance, and attention in affective picture processing. *Biological Psychology*, 78(2): pages 138-149.

[15] Gray, E. and Watson, D. (2007). Assessing positive and negative affect via self-report. In J. A. Coan, and J. J. B. Allen (Eds.), *Handbook of Emotion Elicitation and Assessment*. New York: Oxford University Press, USA.

[16] Keil, A., Müller, M. M., Gruber, T., Wienbruch, C., Stolarova, M., and Elbert, T. (2001). Effects of emotional arousal in the cerebral hemispheres: a study of oscillatory brain activity and event-related potentials. *Clinical Neurophysiology*, 112(11), pages 2057-2068.

[17] Khalfa, S., Isabelle, P., Jean-Pierre, B., and Manon, R. (2002). Event-related skin conductance responses to musical emotions in humans. *Neuroscience letters*, 328(2), pages 145-149.

[18] Kim, J. and André, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12), pages 2067-2083.

[19] Kim, K., Bang, S., and Kim, S. (2004). Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, 42(3), pages 419-427.

[20] Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., and Friederici, A. D. (2004). Music, language and meaning: brain signatures of semantic processing. *Nature Neuroscience*, 7(3), pages 302-307.

[21] Lang, P. J., Greenwald, M. K., Bradley, M. M., and Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30(3), pages 261-273, 1993.

[22] Magnée, M. J., Stekelenburg, J. J., Kemner, C., and de Gelder, B. (2007). Similar facial electromyographic responses to faces, voices, and body expressions. *Neuroreport*, 18(4), pages 369-372.

[23] Mandryk, R. L., Inkpen, K. M., and Calvert, T. W. (2006). Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & IT*, 25(2), pages 141-158.

[24] Marosi, E., Bazán, O., Yañez, G., Bernal, J., Fernández, T., Rodríguez, M., Silva, J., and Reyes, A. (2002). Narrow-band spectral measurements of EEG during emotional tasks. *The International Journal of Neuroscience*, 112(7), pages 871-891.

[25] Mueller, M. M., Keil, A., Gruber, T., and Elbert, T. (1999). Processing of affective pictures modulates right-hemispheric gamma band EEG activity. *Clinical Neurophysiology*, 110(11), pages 1913-1920.

[26] Picard, R. W. (1997). *Affective Computing*. Cambridge: MIT Press, USA, 1997.

[27] Picard, R. W., Vyzas, E., and Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10), pages 1175-1191.

[28] Reuderink, B., Nijholt, A., and Poel, M. (2009). Affective pacman: A frustrating game for brain-computer interface experiments. *In Intelligent Technologies for Interactive Entertainment*, volume 9, pages 221-227.

[29] Sammler, D., Grigutsch, M., Fritz, T., and Koelsch, S. (2007). Music and emotion: Electrophysiological correlates of the processing of pleasant and unpleasant music. *Psychophysiology*, 44(2), pages 293-304.

[30] Sander, D., Grandjean, D., and Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural Networks*, 18(4), pages 317-352.

Affect driven Creativity Support Tools

Priyamvada (Pia) Tripathi
Department of Computer Science
and Engineering
Arizona State University, Tempe,
AZ, USA
pia@asu.edu

Abstract

The relationship between affect and creativity represents an intriguing opportunity for creativity support tools. A fine grain model that links affect with creativity can lead to development of feedback environments that maximize a user's effectiveness and creative output. In this dissertation proposal, I present formulation of affective feedback environments that aim at emotional support for the users through feedback. This formulation is designed for enhancing group creativity. Preliminary results were obtained from an empirical study conducted on group affect and creativity in a leading industrial research laboratory. Five members in a research group reported their daily affect and creativity using an online social science survey. In addition, they used a PDA liken social sensing device called socio-scopes that captured their affective and social behavior patterns. The results show that creativity is positively correlated with positive affect. Furthermore, person's speech and movement profiles are also significantly correlated with creativity. More studies are planned for future that will refine these results. Based on these empirical investigations, I propose that creativity support tools will benefit by actively considering the impact of affect on the creative process. Physiological sensors can serve an important role in providing this measurement. The proposed research lays the foundation for future work in development of such affect driven real time measurement and modulation systems for creativity support.

Keywords: Affective computing, Creativity, CreativeIT, Group work, Social signal processing.

1 INTRODUCTION

Creativity has been defined as any product, process, or personality that is novel and appropriate [1]. For a long time, scientists and laymen alike treated creativity as anomaly of human nature placing it outside the context of everyday use. In recent years, however, due to the efforts of scientists such as Amabile [2], Csikszentmihalyi [3] and others [4], everyday creativity has now attained a much deserved scientific focus. In popular literature as well, it is contended that this creative class is leading way to a creative economy that will be driven by myriads of innovations on multiple levels [5]. A 2003 report by National Research Council [6] emphasized a similar need for information technology (IT) to support innovation across users, groups, and organizations. The boom in user-centric applications such as facebook®, youtube®, twitter®, whose success can be little explained by their functionality or usefulness alone, is also indicative of the need of applications that are scientifically grounded in the subjective reality of the user. There is a tremendous need for user-centric technologies that promote individual and collective creativity.

Empirical approaches have themselves undergone tremendous change since Guilford [7] first emphasized the need to study creativity in his annual address at the American Psychological Society more than 50 years ago. Scientists have found unique ways to access person's nonverbal or subconscious states through physiological and biometric sensing [8]. For example, skin conductance recording, voice tone recognition, facial expression analysis, gesture and posture signals as well as behavioral interaction patterns such as location of the person as well social network can be measured through existing tools [9]. These tools are both theoretically grounded and empirically validated. Social science instruments also increasingly reflect the growth from initial understanding of creativity as a pure reasoning process to more complex interactive spectrum within the user's internal and external context.

Amabile's [2] study of affect-creativity interrelationship using her KEYS scale reflects this shift by incorporating these contextual constructs in the self-report survey.

The relationship between affect and creativity represents an intriguing opportunity for user centered IT tools that foster creativity. In this paper, I present preliminary guidelines for affect driven feedback systems that are derived from fine grained analysis of multimodal data. Physiological and social sensing sensors are used in conjunction with social science instruments to provide reliable data for mining a multilevel model of affect-creativity relationship.

2 BACKGROUND AND RELATED WORK

Affect is an umbrella term that includes moods, feeling states, and emotion [10]. Studies show that there is deep and complex relationship between various types of affect and its influence on the creative process . Several experimental investigations into the relationship between affect and creativity have shown that positive affect is directly related to divergent thinking (defined as ability to generate variety of ideas) and cognitive flexibility (ability to see things with a new perspective). This relationship has been found in naturally occurring positive affect such as through self report measures [11] as well as experimentally induced affect such as through showing 5 minutes of comedy film or giving a small gift [12]. Isen and her colleagues [12] have found that children, adolescents, as well as adults alike give more category words or show improved problem solving ability with positive affect. They have also shown that positive affect is associated with more cognitive flexibility and heightened sensitivity to differences. Hirt and colleagues [13] showed that participants worked longer and better and more creatively if the given task was told to be worked on as long as the participants enjoyed it. On the other hand, participants who were given a sufficiency goal, stopped as soon as they had "done enough". In these studies positive affect is shown to influence the cognitive processes of the person and therefore led to improved performance. Kaufman and Vosburg [14] found that positive affect was in fact detrimental to an insight task while negative affect was beneficial. They argued that positive affect encourages a satisfying strategy while negative affect facilitates optimizing strategy. Similarly, Melton [15] induced mood by means of cartoons and did not find that positive mood led to superior performance. He attributed the poor performance to the idea that participants in positive moods expend less effort on a task.

In addition to these studies at individual level affect, there have been a few direct investigations into manipulated affect in group situations. Forgas [16] induced positive, negative, and neutral moods in individual or group participants who were asked subsequently to rate nine person categories along a number of dimensions. Similar to individual studies, the mood influenced the judgment of categories with happy (sad) individuals making more positive (negative) choices compared to controls. Moreover, being in a group amplified the positive bias and lowered the negative bias. Staw et al [17] have reported that stressful environments led to more rigid cognitive processing similar to models of individual affect mentioned above. Hertel and Fiedler [18] showed that both group cooperation and competition was accentuated in the positive mood group when compared to negative mood group. Hertel and his colleagues [18] have also suggested that heuristic processing is more likely in positive mood group while negative mood group is inclined towards more systematic processing. Carnevale and Probst [19] showed that negative mood in a group induced due to suggestion of conflict led to narrow and rigid thinking thereby reducing creativity. Positive moods however show readiness of action and stimulus elaboration thus leading to sharing of information and promoting participation [20]. Thus, at group level, negative moods encourage thinking in depth (narrow rigid thinking) while positive moods encourage thinking in breadth (shallow but wide) making the relationship between affect and creativity dependent on context.

For the purposes of feedback, emotional interaction between individual can be categorized into three levels: (1) observable and communicative emotions that are expressed verballyor behaviorally such as tone of the speech, facial expressions, gestures, posture. These are conscious acts of communication, (2) Intermediary which act as backchannels to cognitive content and interpretation and are expressed physiologically such as heart rate and speech intonations, blood pressure, (3) unobservable or private emotions. In research studies (1) and (2) have been known to both contradict and support each other on different occasions. Thus, the efforts in affective computing to disambiguate and interpret social behaviors in terms of emotional responses may play a crucial role in HCI. This relationship can also lead to improved mapping between actual experience of the user and expected behavior through human computer interactions. For example, induced or sustained positive affect in an application may encourage its use. This relationship can also be exploited to design applications that promote creativity explicitly or embody craft that requires creativity.

In the proposed research, we hope to test whether the interaction of affect and cognitive styles can be managed by enhancing affective awareness between and among group members. It has been shown that higher self-awareness leads to self-directed adaptation of behavior [21]. It may also allow higher attention to groups processes, goals, and strategies (West [22] defined this as reflexivity). This is related to work by Gersick and Hackman [23] who found that work groups can break dysfunctional habitual routines by self-reflection. Similarly, group self-monitoring can enhance the understanding of breakdowns of creativity related to both cognitive or affective factors and lead to prevention of breakdowns. The challenges with regard to use of affective feedback towards group creativity are many. Some of these include: (a) when to interrupt, (b) when to emphasize individual affect vs. group affect, (c) what degree of manipulation should be allowed to effectively manage and maintain group cohesiveness. Prior to application, one needs to study and understand the relationship between creativity at work and personal and team affect.

Many IT tools integrate emotional support for the creative user. This is accomplished through emoticons, affect laden messages, humor, etc. There have also been a few attempts towards active affective support for the user as well. For example, Klein et al [24] tested participants who experienced an artificially induced frustration while playing games. After five minutes of delay, one set of participants got a questionnaire that ignored their emotions, another set got a questionnaire that allowed them to vent their emotions, and last set was given questionnaire with affective feedback and support. They found that subjects who were given affect-support questionnaire got greater relief than either of the other two groups. Similarly, Partala and Surakka [25] found that performance improved significantly due to affective interventions using synthetic speech with positive emotional content.

3 PROPOSED OBJECTIVES

The specific objectives of this dissertation are:

(1) To conduct empirical studies that provide fine grain data on relationship between affect and creativity at both individual and group level,

(2) To provide validity of using hybrid methodologies such as physiological sensing, behavioral sensing, and social sensing for understanding the relationship between affect and creativity

(3) To develop empirically driven model that describes the relationship between affect and creativity

(4) To develop a framework for affect based creativity support tools

(5) To test the validity of affect based creativity support tools through carefully designed intervention.

4 RESEARCH FRAMEWORK

This research is divided into two broad phases. In research phase#1, objective (1) – (3) are tackled and in research phase # 2, objective (4) and (5) are targeted. Figure 1 gives an overview of the proposed research.

4.1 RESEARCH PHASE# 1: HYBRID METHODOLOGIES

This research combines the strengths of the emerging methodologies in social sciences with advanced computational tools for sensing human behavior in digital and physical environments. This hybrid methodology presents us data at a finer granularity of multiple behavioral and affective elements for creativity research in IT research organizations. Existing social science survey tools such as Amabile's KEYS scale [2] are used to identify factors that influence and interact with creativity at individual and team levels. Previous research [26, 27] has successfully employed empirical setups that make use of physiological sensing and wearable computing to understand and predict the relationships between low level signals and high level behavioral constructs. For example, methods from affective computing have been able to distinguish affective state at 81% throughout everyday activities [26] while machine learning tools that incorporate Human Eigen behaviors and Coupled Hidden Markov Models (CHMMs) have been shown to account for 96% of the variance of behavior of typical individuals [27].Therefore, this research leverages the strengths of social science survey tools, affective sensing, and social sensing methodologies towards new findings on affect-creativity interrelationships.

4.1.1 Study # 1: Studying relationship between affect and creativity in a research environment

Participants

Five people (mean age = 32.4, range= 26-38, all had at least an undergraduate degree) involved in a creative project lasting four weeks at a leading IT research laboratory in USA wore socioscopes (wearable devices that are worn around neck and capture accelerometer, voice, and network data) during their workdays and completed online survey based on Amabile's KEYS scale at the end of workday. The total number of hours participants worked on the project for the duration of the study was reported to be 653.25 with an average of 7.17 hours per work day. Out of these hours, approximately 26.6% of average time (standard deviation 19.4) was spent in working with the team members. On average, team had 3 members in attendance.

Apparatus

The survey is an adaptation of survey employed by Amabile et al [11]. The daily questionnaire includes items related to affect, team work and external environmental factors for that particular day. The survey has 15 questions in total, out of which two were open ended. A 7 point Likert scale with 0 being not at all and 7 being extremely is used for self ratings on numerical measures. For the open ended question, they were asked to describe the major activities they were engaged in that were relevant to the target project on that day. Second, they were asked to "briefly describe at least one event from today that stands out in their mind as relevant to the target project, their feelings about this project, their work on this project, their team's feelings about this project, and their team's work on this project".

In addition, participants are given socioscopes provided by Dr. Pentland's lab (MIT, Boston, USA). Socioscopes are a type of wearable computing devices such as PDA and cell phone to locate a person, upper body movement and ambient audio using embedded speakers and 2d accelerometer that are invisible to the user. Thus, socioscopes track location and analyze elements of participants' social interaction through bi-directional IR, accelerometers, and low-resolution microphone analysis [30]. No personal data is recorded. For example, in speech recording, only digital 'tone' is recorded and the time the person is speaking and not the content of the speech.

Procedure

All participants were informed that we are studying work flow issues in team work. Participants were not told that the purpose of the study was to see relation between their affect and creativity. This was done so as to not interfere with their natural affective responses. In addition, all participants are given participation ID through which they corresponded for the duration of the study. The participants were also made aware that their responses will remain anonymous and will be evaluated by researchers independent of the work involved. This encouraged them to respond freely on the survey. They were given socioscopes as well as online daily survey. Participants devoted 10-15 minutes on every work day to complete the end-of-the-day survey.

Results

Participated self-reported constructs included positive affect (happy, satisfied, enjoyment of work), Negative affect (frustrated, distracted, satisfied (reverse scored)), Intrinsic motivation (felt motivated, challenged, own internal pressure), extrinsic motivation (rewards that might earn, team feedback, supervisors feedback, organizational resources). Creativity was measured through constructs of direct creativity, imagination, and idea-flow. In addition we looked at the narratives for the open ended questions for qualitative assessment.

The items were subjected to principal component analysis (PCA) using SPSS 17.0. Prior to analysis, suitability was assessesed by inspecting correlation matrix that revealed the presence of many coefficients of 0.3 and above. The Kaiser-meyer-oklin value was 0 .726 exceeding the recommended value of 0.6 and Barlett's test was statistically significant supporting factorability of correlation matrix. PCA revealed the presence of 2 components with eigenvalues exceeding 1 explaining 40.2 percent and 20.6 per cent of variance respectively. Thus two components were retained and verimax rotation with Kaiser normalization was performed to aid in interpretation. The rotated solution had a simple structure with both components showing strong loadings and all variables loading substantially on one variable. When the survey was checked, the components corresponded to positive and negative affect respectively.

Subsequently, Pearson-moment correlation coefficients was computed between these two sets of variables corresponding to positive and negative affect with the independent variables : team creativity and personal creativity. It was found that all but one positive variable (relaxed) correlated significantly. No negative affect variable correlated significantly (at .01 level) with the independent variables.

The goal of the scocioscope analysis was to find whether there are global trends on creativity, positive affect, negative affect and tag data. Socioscope analysis was centered on the following questions: (1) Is there a correlation between creativity and movements? (2)Is there a correlation between creativity and speech? (3) Are movements and speech correlated?

The accelerometer data from the socioscopes was analyzed as follows. First, following features were extracted: magnitude, mean, variance, energy, spectral entropy, pairwise correlation between the three axes, the first ten FFT coefficients and exponential FFT bands. These are calculated for 5 element sliding window across the accelerometer profile after which all values are normalized to be between 0 and 1. After this was done for each file, the mean and standard deviation value of each of these values were used to develop a users' profile. Each of the variables were normalized separately and then a correlation between each of the variables and the normalized creativity, positive affect and negative affect value was calculated.

From the speech data, the gross words spoken per file by each subject was calculated and the mean and standard deviation of the spoken words was found. Creativity correlated significantly with acceleration profiles which means that people who show movement are more likely to feel creative. Creativity also significantly correlated to speech which means people who spoke more felt more creative. Speech was positively correlated with positive affect while it had no impact on negative affect.

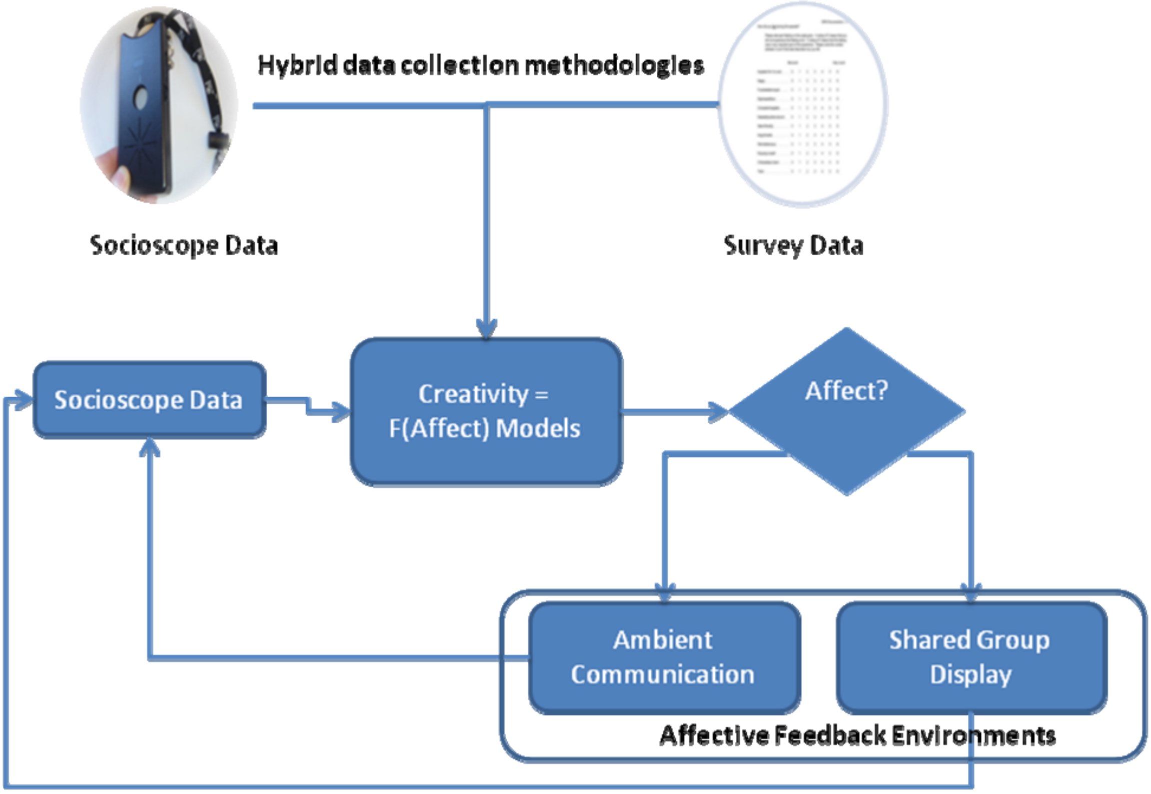
Hybrid data collection methodologies
Socioscope Data
Survey Data
Socioscope Data
Creativity = F(Affect) Models
Affect?
Ambient Communication
Shared Group Display
Affective Feedback Environments

Figure 1 Overview of the Proposed Research

4.2 Research Phase # 2: Affect driven feedback systems

This section describes how focused affective interventions can facilitate team interaction and promote team creativity. First, the characteristics of affective feedback environments are listed and then empirical methodology that can be used to validate these principles is presented.

Affect driven feedback systems are defined by following characteristics:

(1) They promote self-directed behavior enhancing motivation, personal engagement, and self-awareness. Thus, affective feedback is designed to work with the subjective background of the user. For example, they can help a user understand that he or she is frustrated by reflecting the same emotion. This is similar to feedback in group displays that has been used to motivate participants to modify their speaking and turn taking behavior by embodying

these behaviors through color coding, or size of the participant circle. Affective behavior that is sensed physiologically and behaviorally can be communicated symbolically to the user through icons or color codes.

(2) Affect driven feedback systems are based on empirically derived adaptive models of affect-creativity interrelationship. Since they are theoretically grounded in the understanding of individual and team affect, they will provide a more direct relationship between members and technology. This is achieved by using computational engines that mediate the relationship between displays and sensed behavior.

(3) Affect driven feedback systems promote team level awareness. They aim at seamless integration as partners and facilitators within system as a whole. Thus, they aim to promote and preserve ‘flow’, defined as a psychological ‘high’ with heightened feelings of involvement and concentration, at individual and group level.

(4) Ideally, affective feedback will match skill levels, and affective state, to task typology and stage of project completion. Research shows that specific types and levels of affect are more conducive for divergent thinking or convergent thinking. Hence, though behavioral sensing, we can encourage team members to do certain type of task versus another to suit the given affective states. This strategy may also include moderating extreme affect that may be distracting for the overall productivity or manipulating affect from a negative continuum to positive continuum.

4.2.1 Research Questions

The role of affect and cognitive styles can be managed by enhancing awareness between and among group members. These investigations can be summarized as following research questions:

R1: Affective intervention will have significant impact on group cohesiveness factors measured by leveling of affect (lesser variance) and higher ratings of team level positive affect.

R2: Affective intervention will lead to increased creativity. This will be measured by observation and output of the proposed task. Creativity will be measured by number of ideas generated, novelty of ideas generated, and value of ideas generated. Consensual assessment technique will be employed to judge creative output produced during the session.

4.2.2 Data Collection

The platform is closely based on affective sensing platform used by Burleson to drive affective agent [28] in combination with socioscopes proposed by Pentland [27]. The computational engine is driven by multimodal sensors such as pressure chair, skin conductance recording. Following physiological Sensors record physical state of the participants in an unobtrusive and natural manner:

1. Posture chair: It gives ratio of forward to backward posture, activity level. The chair has some pressure sensors underneath that record how much pressure is being applied in each zone (the surface of chair is divided into smaller regions) and these differences in pressure are used to calculate posture and activity. Pressure chair gives our pressure maps in the form of 8-bit binary image for each pressure unit present in the seat and the back. After preprocessing, the pressure array is modeled using Gaussian mixture model (GMM) for classification of activity (low, medium, and high).

2. Skin conductance: The skin conductance sensor is embedded into a bracelet (looks like a rubber band). We will use this data to measure arousal of the members.

3. Socioscopes: This will give us how much did each member speak and in what tone.

Thus, in total we have activity, skew, mean, and variance from pressure mouse (4 features), activity and ration of postures from pressure chair (2 features) and skin conductance (1 feature). These features are fed to a multimodal sensing pattern recognition engine that using support vector machines to derive final affect state (positive, negative, neutral).

4.2.3 Experimental Procedure

In a 50 minute session on a weekly basis, the performance of the experimental as well as the control group will tested on a open ended creative task. Minimum guidance will be provided on how the task is conducted.

Participants will be asked to devise solutions to the homeless problem faced by cities and society. This task is has been used by Massetti [29] who suggested that creativity should be measured on tasks that do not require any prior training or specific knowledge. Participants are selected because they had no prior knowledge of issue of homelessness.

During the intervention session, the participants are first given overview of the display and sensors. They are shown the display with affective feedback and are given 5-10 minutes to interact with it. Second, they are asked to fill a pre-test survey on their demography, affect levels, and team perception. Subsequently, they will start the main problem solving task. For this, they are given the problem solving task on a sheet of paper. A video recording will be done as to avoid interference of the group flow and ideation. Subsequent analysis will be based on transcription of this video.

4.2.4 Analysis

Creativity scores will be generated by averaging their number of ideas produced, novelty rating, and value of ideas (each rated on a 0-7 Likert scale with 0 least and 7 maximum). Analysis of variance techniques will be employed to test the significance of intervention by testing interactions of intervention x group affect, intervention x creativity, intervention x flow. Furthermore, multivariate analysis will be conducted to find a the nature of relationship between the intervention ad team affect from pre-intervention survey as independent variable and creativity and team affect post intervention as dependent variables. Software satisfaction will be calculated from averaging subjects' responses on the measures of computer comfort, software likability, better decision support, and ease of use.

5 CONCLUSION

Research has been conducted in social science studying various factors that may interact with the creative outcome. On the other hand, research in computational sensing has now made tremendous successes in sensing the psychological state of the user. Applications such as meeting mediator [30] and visual shared displays [21] suggest that these sensing tools can lead to automated systems that aim to promote individual and group productivity. Creativity plays a significant role in survival of organizations as well as individuals. Creativity can be seen ranging from low-level everyday creativity such as using a chair for climbing and high level creativity such as designing a new tool. Understanding of the underlying processes of creativity especially those that are directly related to successful engagement of a person in creation will lead to improved outcomes in both activities of everyday living as well as long range organizational goals such as products and services. Indirectly, these tools will also result in greater subjective satisfaction with improved concentration, collaboration, productivity. In the long term, it is expected that this approach will lead to adaptive reflective technologies that stimulate collaborative activity, reduce time pressure and interruption, mitigate detrimental processes pertaining to group work, and increase individual and team creative activity and outcomes.

6 AKNOWLEDGEMENT

I would like to thank my advisor Dr. Winslow Burleson for his continued supervision and support. I would also like to acknowledge Mary Czerwinski (Microsoft Corporation), Alex (Sandy) Pentland (MIT), Taemie Kim (MIT) as collaborators on this project. The project is being supported by NSF SGER: Creativity in IT research organizations (Award #0846148)

REFERENCES

[1] Sternberg, R.J. and T.I. Lubart, *The Concept of Creativity: Prospects and Paradigms*, in *Handbook of Creativity*, R.J. Sternberg, Editor. 2007, Cambridge University Press. p. 3-15.

[2] Amabile, T., *Creativity in context.* 1996: Westview Press.

[3] Csikszentmihalyi, M., *Creativity: Flow and the psychology of discovery and invention*. 1997: Harper Perennial.

[4] Sternberg, R.J., ed. *Handbook of Creativity*. 2007, Cambridge University Press.

[5] Florida, R.L., *The rise of the creative class: and how it's transforming work, leisure, community and everyday life*. 2002: Basic Books.

[6] Mitchell, W.J., A.S. Inouye, and M.S. Blumenthal, *Beyond Productivity: Information Technology, Innovation and Creativity* Committee on Information Technology and Creativity, National Research Council. 2003, Washington, D.C.: National Academics Press.

[7] Guilford, J.P., *Creativity.* American Psychologist, 1950. **5**(9): p. 444-454.

[8] Vinciarelli, A., M. Pantic, and H. Bourlard, *Social signal processing: Survey of an emerging domain.* Image and Vision Computing. In Press, Corrected Proof.

[9] Picard, R.W., et al., *Affective Learning - A Manifesto.* BT Technology Journal, 2004. **22**(4): p. 253-269.

[10] Russ, S., *Affect, Creative Experience, and Psychological Adjustment* 1999: Psychology Press.

[11] Amabile, T.M., et al., *Affect and Creativity at Work.* Administrative Science Quarterly, 2005. **50**: p. 367-403.

[12] Isen, A., *On the relationship between affect and creative problem solving*, in *Affect, Creative Experience and Psychological Adjustment*, S.W. Russ, Editor. 1999, Brunner/Mazel: Philadelphia. p. 3-18.

[13] Hirt, E.R., et al., *Processing goals, task interests, and mood-performance relationship: a mediational analysis.* Journal of Personality and Social Psychology, 1996. **71**: p. 245-261.

[14] Kaufmann, G. and S.K. Vosburg, *'Paradoxical' Mood Effects on Creative Problem-solving.* Cognition & Emotion, 1997. **11**(2): p. 151-170.

[15] Melton, R.J., *The role of positive affect in syllogism performance.* Personality and social pyschology bulletin, 1995. **21**: p. 788-794.

[16] Forgas, J.P., *Affective influences on individual and group judgments.* European Journal of Social Psychology, 1990. **20**(5): p. 441 - 453.

[17] Staw, B.M., L.E. Sandelands, and J.E. Dutton, *Threat-rigidity effects on organizational behavior.* Administrative Science Quarterly, 1981. **26**: p. 501-524.

[18] Hertel, G., et al., *Mood effects on cooperation in small groups: Does positive mood simply lead to more cooperation?* Cognition & Emotion, 2000. **14**(4): p. 441- 472.

[19] Carnevale, P.J. and M. Probstt, *Social values and social conflict in creative problem solving and categorization.* Journal of personality and social psychology, 1998. **74**(5): p. 1300-1309.

[20] Frijda, N.H., *The emotions*. 1986: Cambridge University Press.

[21] DiMicco, J.M. and W. Bender, *Group Reactions to Visual Feedback Tools,* in *Persuasive Technology,* Y. de Kort, et al., Editors. 2007, Springer Berlin / Heidelberg. p. 132-143.

[22] West, M.A., *Reflexiity and work group effectiveness: A conceptual integration,* in *Handbook of work-group psychology,* M.A. West, Editor. 1996, Wiley: Chichester, UK. p. 555-579.

[23] Gersick, C.J.G. and J.R. Hackman, *Habitual Routines in Task-Performing Groups.* Organizational Behavior and Human Decision Processes, 1990. **47**: p. 65-97.

[24] Klein, J., Y. Moon, and R. Picard, *This computer responds to user frustation: theory, design, and results.* Interacting with Computers, 2002. **14**: p. 119-140.

[25] Partalaa, T. and V. Surakka, *The effects of affective interventions in human–computer interaction.* Interacting with Computers, 2004. **16**: p. 295-309.

[26] Kapoor, A., W. Burleson, and R.W. Picard, *Automatic prediction of frustation.* International Journal of Human-Computer Studies, 2007. **65**: p. 724-736.

[27] Pentland, A., *Automatic mapping and modeling of human networks.* Physica A: Statistical Mechanics and its Applications, 2007. **378**(1): p. 59-67.

[28] Burleson, W., et al. *A platform for affective agent research.* in *Workshop on Empathetic Agents, Third International Joint Conference on Autonomous Agents and Multi-Agent Systems*. 2004. New York, NY.

[29] Massetti, B., *An Empirical Examination of the Value of Creativity Support Systems on Idea Generation.* MIS Quarterly, 1996. **20**: p. 83-97.

[30] Kim, T., A. Chang, and A.S. Pentland. *Enhancing Organizational Communication using Sociometric Badges*. in *IEEE 11th International Symposium on Wearable Computing (Doctoral Colloquium).* 2007. Boston MA.

List of authors

A

Acosta, Jaime C. 49

B

Baccan, Davi D'Andréa 73
Beetz, Micheal 41

F

Favre, Sarah 25

G

Gedikli, Suat 41

H

Hyniewska, Sylwia 81

K

Kaiser, Susanne 81
Košir, Andrej 33

M

Mühl, Christian 89
Macedo, Luís 73

N

Novielli, Nicole 1

P

Pammi, Sathish 57
Pelachaud, Catherine 81

R

Radig, Bernd 41
Riaz, Zahid 41
Riek, Laurel D. 65
Robison, Jennifer 17

S

Strapparava, Carlo 1

T

Tasič, Jurij 33
Tkalčič, Marko 33
Tripathi, Priyamvada (Pia) 97

Y

Yu, Changrong 9

Z

Zhou, Jiehan 9