AISB 2004 Convention:

Motion, Emotion and Cognition

AISB The Society for the Study of Artificial Intelligence and the Simulation of Behaviour

Proceedings of the AISB 2004

Symposium on Language, Speech and Gesture for Expressive Characters

29 March – 1 April, 2004

ICSRiM, University of Leeds, Leeds LS2 9JT, UK

www.leeds.ac.uk/aisb www.icsrim.org.uk

AISB 2004 Convention

29 March – 1 April, 2004
ICSRiM, University of Leeds, Leeds LS2 9JT, UK
www.leeds.ac.uk/aisb www.icsrim.org.uk

Proceedings of the AISB 2004
Symposium on Language, Speech & Gesture for Expressive Characters

Published by

The Society for the Study of Artificial Intelligence and the Simulation of Behaviour

http://www.aisb.org.uk

ISBN 1 902956 39 0

Contents

The AISB 2004 Convention ……. ii
K. Ng

Symposium Preface ……. iii
R. Aylett, M. Cavazza, P. Olivier

Expressive speech characteristics in the communication with artificial agents ……..…… 1
Kirsten Fischer

A probabilistic hierarchical framework for active appearance model ..……..…… 12
Lukasz Zalewski & Shaogang Gong

Modelling character emotion in an interactive virtual environment ..…… 20
El Jed Mehdi, Pallamin Nico, Dugdale Julie and Pavard Bernard

Artistically based computer generation of expressive motion ……..…… 29
Michael Neff & Eugene Fiume

Speaking and acting - interacting language and action for an expressive character …… 40
Sandy Louchart, Daniela Romano, Ruth Aylett & Jonathan Pickering

Expressive characters and a text chat interface ……...…… 48
Marco Gillies, Barry Crabtree & Daniel Ballin

Speaking with emotions …… 58
Elisabetta Bevacqua, Maurizio Mancini & Catherine Pelachaud

Reusing motion data to animate visual speech ……..……...…… 66
James D. Edge Manuel A. Sánchez Lorenzo & Steve Maddock

Influences on Embodied Conversational Agent's Expressivity …...…..…… 75
Vincent Maya, Myriam Lamolle & Catherine Pelachaud

Developing a virtual ballet dancer to visualise choreography ……..…… 86
Royce Neagle, Kia Ng & Roy Ruddle

Virtual human signing as expressive animation …… 98
John Glauert, Richard Kennaway, Ralph Elliott & Barry-John Theobald

Defining the gesticon: language & gesture coordination for interacting embodied agents ... 107
Brigitte Krenn & Hannes Pirker

Artificial companions for older people …… 116
Gopal Vaswani, David Benyon, Stewart Cringean, Oli Mival & Greg LePlatre

"To be or seem to be; that is the question" ……...…… 124
Christopher Newell & Alistair Edwards

To tell or not to tell...building an interactive virtual storyteller ……..…… 132
Andre Silva, Guilherme Raimundo, *Celso de Melo & Ana Paiva*

Pragmatics of Body-Speech Coordination in Multi-Modal Expression …...…… 136
Satinder P. Gill & Masahito Kawamori

The AISB 2004 Convention

On behalf of the local organising committee and all the AISB 2004 programme committees, I am delighted to welcome you to the AISB 2004 Convention of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour (SSAISB), at the University of Leeds, Leeds, UK.

The SSAISB is the oldest AI society in Europe and it has a long track record of supporting the UK AI research community. This year, the underlying convention theme for AISB 2004 is "*Motion, Emotion and Cognition*", reflecting the current interest in such topics as: motion tracking, gesture interface, behaviours modelling, cognition, expression and emotion simulation and many others exciting AI related research topics. The Convention consists of a set of symposia and workshop running concurrently to present a wide range of novel ideas and cutting edge developments, together with the contribution of invited speakers:

- Prof Anthony Cohn
  Cognitive Vision: integrating symbolic qualitative representations with computer vision;
- Prof Antonio Camurri
  Expressive Gesture and Multimodal Interactive Systems;
- Dr David Randell
  Reasoning about Perception, Space and Motion: a Cognitive Robotics Perspective; and
- Dr Ian Cross
  The Social Mind and the Emergence of Musicality;

not to mention the many speakers invited to the individual symposia and workshop, who will made the Convention an exciting and fruitful event.

The AISB 2004 Convention consists of symposia on:

- Adaptive Agents and Multi-Agent Systems;
- Emotion, Cognition, and Affective Computing;
- Gesture Interfaces for Multimedia Systems;
- Immune System and Cognition;
- Language, Speech and Gesture for Expressive Characters; and the
- Workshop on Automated Reasoning.

The coverage is intended to be wide and inclusive all areas of Artificial Intelligence and Cognitive Science, including interdisciplinary domains such as VR simulation, expressive gesture, cognition, robotics, agents, autonomous, perception and sensory systems.

The organising committee is grateful to many people without whom this Convention would not be possible. Thanks to old and new friends, collaborators, institutions and organisations, who have supported the events. Thanks the Interdisciplinary Centre of Scientific Research in Music (ICSRiM), School of Computing and School of Music, University of Leeds, for their support in the event. Thanks to the symposium chairs and committees, and all members of the AISB Committee, particularly Geraint Wiggins and Simon Colton, for their hard work, support and cooperation. Thanks to all the authors of the contributed papers, including those which were regretfully not eventually accepted. Last but not least, thanks to all participants of AISB 2004. We look forward to seeing you soon.

Kia Ng
AISB 2004 Convention Chair
ICSRiM, University of Leeds,
School of Computing & School of Music, Leeds LS2 9JT, UK
kia@kcng.org www.kcng.org

Proceedings of the AISB 2004
Symposium on Language, Speech & Gesture for Expressive Characters

Symposium Preface

Research into expressive characters, for example embodied conversational agents, is a growing field, while new work in human-robot interaction has also focussed on issues of expressive behaviour. With recent developments in computer graphics, natural language engineering and speech processing, much of the technological platform for expressive characters – both graphical and robotic – is in place.

However, progress has been hampered by the need to integrate work in various sub-fields of psychology, in natural language processing, speech and in computer graphics, carried out by many different groups in communities that rarely intersect. Other areas, such as integrating gesture and facial expression and affective state with language and speech, are less developed but vital to progress.

The symposium aims to bring together psychologists, experts in natural language and speech technologies, researchers in embodied agents (graphical and robotic), affective computing and computer graphics and animation researchers

The Organising Committee

Expressive Speech Characteristics in the Communication with Artificial Agents

Kerstin Fischer*
*University of Bremen
FB10: Sprach- und Literaturwissenschaften
Postfach 330440
28334 Bremen
tel: +49-421-218-9735
fax: +49-40-42883-2515
kerstinf@uni-bremen.de

Abstract

This paper deals with emotional speech characteristics in human-computer- and human-robot interaction. The focus is on the users' involuntary expression of emotion in reaction to system malfunction, which may cause severe problems for automatic speech recognition and processing. Investigating different user groups is shown to be a useful method for determining what makes speakers respond emotionally and for understanding the interpersonal differences that can be observed in reaction to system malfunction. Which aspects may be involved is illustrated by discussing the example of the personal relationship between user and system as evident from the different forms of address that can be found in the corpora. We shall draw on corpora of human-computer and human-robot communication involving children and adults from both sexes. However, it will be demonstrated that the major factor that determines the users' expressive behaviour is their conceptualisation of the artificial agent and the situation. Knowledge about this is then used to develop means for influencing the speakers' attitude towards the system, which can be shown to change their expressive speech characteristics in situations of system malfunction much more effectively than by assessing the linguistic behaviour directly. Thus, interestingly, what turns out to be most suitable for guiding the user into a linguistic behaviour that is understandable for an automatic speech processing system is employing aspects of expressive speech as well.

1 Introduction

Not all emotional expression in the communication with artificial agents may be welcome. Of course expressive speech is a natural characteristic of conversational behaviour in human-human situations. In conversation, topics and purposes are free to vary, speakers may report on emotionally engaging events, and co-participants may display their involvement in order to align with the speaker. Communication with artificial agents in contrast is usually task-oriented and domain-restricted. With the exception of *eliza*-like systems, human-computer interaction (HCI) and human-robot interaction (HRI) are generally carried out for a particular, usually pre-defined, purpose, and thus topics are not free to vary. Emotional topics may thus only arise if planned for by the system designers.

Emotional expression may then occur for three reasons: either it is designed to be part of the domain, or it is used as an accompanying feature to make also domain- and task-specific interaction more human-like, or it occurs as a not planned reaction by system users to particular behaviours of the system, such as system malfunction. In this paper, I will address on the one hand the expressive characteristics of speech humans direct to dialogue systems in situations of miscommunication, the problems these emotional characteristics cause for automatic speech processing systems, and the consequences of these findings for the design of dialogue systems, especially for dialogue management and personality modelling. Thus, it will be discussed how applications can be enriched so as to deal with the problems outlined. On the other hand, it will be discussed in how far, and under which conditions, expressive behaviours from the side of the artificial agent can be used to make the communication more human-like.

The paper thus aims at the descriptive adequacy of the linguistic phenomena involved, and it can be evaluated by a measurable increase or decrease of expressive properties in reaction to particular dialogue acts.

2 Data and Method

The data employed in this study are elicited on the basis of a particular methodology. The method consists in eliciting data of human-computer and human-robot interaction by keeping as many variables constant as possible and by systematically varying only particular aspects of the com-

municative situation. This means that the system's output, as much as the robot's behaviour, are controlled by means of predefined schemata. This method ensures the interpersonal comparability of the data and the identification of those contextual features that determine the users' linguistic behaviour. Another aspect of the method is the repeated use of system malfunction in order to get the users to reformulate their utterances and thus to uncover their hypotheses about what may have caused the communicative problem. A simple example would be that in case of a miscommunication if the speaker starts speaking very loudly, she displays her hypothesis about her communication partner as someone who needs to be talked to loudly. For details on the method, see Fischer (2003).

The data for this study stem from two sources. On the one hand, Wizard-of-Oz data have been elicited that provide an independent method for identifying emotional speech characteristics.[1] In particular, there are 64 German and 8 English dialogues of appointment scheduling with an average length of 18-33 minutes. There are about 248 turns per dialogue elicited in a Wizard-of-Oz scenario (simulated human-machine dialogues), and questionnaire results show that speakers have not doubted to be speaking to a real system. A particular feature of the corpus is the control for inter- and intrapersonal variation which is achieved by employing a fixed schema according to which 'system' output is generated. The system malfunctions simulated are misunderstanding, failed understanding, generation/synthesis errors, and extra long processing time. The sequences of system output are repeated at least three times throughout the dialogues so that we can compare how speakers react to a particular dialogue act, say, for the first, the third and the fifth time, and thus have an independent measure of changes in speaker attitude (since nothing else changes within these dialogues). Table 1 shows a short extract from the fixed schema of system utterances. The turn ID allows the identification of the occurrence of the turn within each dialogue, and as sequences 2101-2103 and 3101-3103 illustrate, the sequences of system output are repeated throughout the dialogue three to five times. All dialogues were annotated for prosodic, lexical, and conversational peculiarities (Fischer, 1999a).

On the other hand, in the framework of a larger project investigating human-robot interaction (SFB/TR8 'Spatial Cognition' at the Universities of Bremen and Freiburg), a constantly growing body of HRI dialogues are being elicited; by controlledly varying particular situational variables the goal is to find out which situational parameters influence the different ways of speaking to a robot.

The two corpora of German and English HRI used here were elicited in the following two settings: The first set of dialogues was elicited in a joint attention scenario where the participants' task was to verbally instruct Sony's Aibo

Table 1: Fixed Schema of Computer Output

ID	Dialogue Act	System Output
2101	Nonsense	What for date whatthehell bla.
2102	Nonsense	Bla rabartibla blurb.
2103	Request proposal	Please propose a date.
2201	Reject proposal	This time is already occupied.
2202	Misunderstanding	Vacation time is June 15 to July 20.
2203	Request proposal	Please propose a date.
2301	Misunderstanding	7th of February is a Sunday.
2302	No understanding	I did not understand.
2303	Misunderstanding	The weekend is already occupied.
2304	Misunderstanding	It is impossible to meet at 4am.
2305	Reject proposal	This time is already occupied.
2306	Misunderstanding	Friday suits me well.
2307	Misunderstanding	1rst of March is already taken.
2308	Accept proposal	I have noted the appointment.
3101	Nonsense	What for date whatthehell bla.
3102	Nonsense	Bla rabartibla blurb.
3103	Request proposal	Please propose a date.

(see Figure 1) to move to a particular object pointed at by the leader of the experiment and through a sequence (a parcour) of such object localisation tasks.[2] Again the robot's behaviour was predetermined, involving some malfunctions in order to get users to present several different solutions to the same problems. The robot's behaviour was actually manipulated by a wizard according to a fixed schema. The method thus relies heavily on speakers' reformulations because those show what their hypotheses are about what could have gone wrong (Fischer, 2003) and thus their mental models of the system. We have furthermore access to dialogues recorded at the University of Erlangen (Batliner et al., 2004), for which the setting was coordinated with ours, yet which involves children, not adults, as in our data.

The second set of HRI dialogues was elicited in a scenario in which the users' task was to instruct the robot, a pioneer 1, to measure distances between two objects from a set of seven pointed at by the leader of the experiment. In this case, the users typed their instructions into a notebook. The robot's output was also predetermined, consisting for the most part either of error messages or of messages naming the distances between the objects to be measured. 21 German speakers participated in this experiments, about half of which were computer scientists.

[1] The transcription conventions are <P> for pause, <B> for breathing, <L> for syllable lengthening, and interpunctuation (,.?) for level, falling and rising intonation contours respectively.

[2] The transcription conventions are: -, –, and (1sec) for the pauses, (at=loud) (/a) for attributes of speech, such as loudness, and interpunctuation (,.?) for level, falling and rising intonation contours respectively.

Figure 1: Sony's Aibo Robot

3 Emotional Speech in the Communication with Artificial Agents

The analyses of emotional peculiarities in HCI allow us to present a typology of the expressive characteristics of speech in the communication with artificial agents that are due to system malfunction. Such an account is non-trivial since most research on expressive properties of speech has been carried out on data produced by actors (cf. Tischer, 1993; Batliner et al., 2003), and very little research has been conducted on real situations of system use (but see Oviatt, 1995; Oviatt et al., 1998a,b; Levow, 1999). Using actors' speech has the methodological advantage that it is clear which strings of speech are supposed to be emotional. In actual situations it may be very difficult to lay one's finger at the moment in which a speaker gets emotionally involved (Batliner et al., 2000). Oviatt and collaborators solve this problem by focussing on particular speech acts, such as repetitions. In the current study, we have the possibility, because of the particular design of the corpus, to identify the emotional properties of speech on independent grounds, namely on the basis of changes within the linguistic behaviour in reaction to the same system behaviour at different points in time.

In the following example, the same speaker, e058, is confronted with the same system utterances three times. While in the earlier phase of the dialogue the speaker uses metalanguage to discuss the misunderstanding, he uses an ironic comment [*yes, great*] some minutes later, and towards the end of the dialogue he uses many prosodic peculiarities and an impolite evaluation [*you should be fired*]:

(1) s0582201: dieser Termin ist schon belegt. [*this date is already occupied*]

e0582201: FÜNFter erster. [*5th of January*]

s0582202: die Urlaubszeit ist fünfzehnten Juni bis zwanzigsten Juli. [*vacation time is from 15th of June to 20th of Juli*]

e0582202: ja, das hat ja auch nicht viel damit zu tun, da wir uns im Januar befinden, ne? [*well this has not much to do with the fact that we are in January, right?*]

s0582203: bitte machen Sie einen Vorschlag. [*please make a proposal*]

e0582203: fünfter Januar. [*5th of January*]

(2) s0584101: dieser Termin ist schon belegt. [*this date is already occupied*]

e0584101: ACHTZEHN bis zweiundzwanzig Uhr. [*6 to 10pm*]

s0584102: die Urlaubszeit ist fünfzehnten Juni bis zwanzigsten Juli. [*vacation time is from 15th of June to 20th of Juli*]

e0584102: <B> ja, klasse. <P> Dienstag, zwölfter erster, ACHTzehn bis zweiundzwanzig Uhr. [*yes, great. Tuesday, 12th of January, 6 to 10pm*]

s0584103: bitte machen Sie einen Vorschlag. [*please make a proposal*]

e0584103: Dienstag, zwlfter erster, achtzehn bis zweiundzwanzig Uhr. [*Tuesday, 12th of January, 6 to 10pm*]

(3) s0586206: dieser Termin ist schon belegt. [*this date is already occupied*]

e0586206: sechster <P> Januar, <P> zwanzig bis zweiundzwanzig Uhr. [*6th of January, 8 to 10pm*]

s0587102: die Urlaubszeit ist fünfzehnten Juni bis zwanzigsten Juli. [*vacation time is from 15th of June to 20th of Juli*]

e0587102: dich sollte man feuern. <B> sechster <P> Januar , <P> zwanzig bis zweiundzwanzig Uhr. [*you should be fired. 6th of January, 8 to 10pm*]

s0587103: bitte machen Sie einen Vorschlag. [*please make a proposal*]

e0587103: se<L>chster Ja<L>nua<L>r, <P> <;<zwa<L>nzig bi<L>s zweiundzwa<L>nzig Uhr> ;with very low voice>. [*6th of January, 8 to 10pm*]

Thus, using the method proposed, the prosodic peculiarities, lexical means and conversational strategies, such as metalanguaging or the use of repetitions can be correlated with particular staes of the dialogues and thus with changes in speaker attitude.

3.1 Prosodic Peculiarities in Reaction to System Malfunction

The phonetic and prosodic peculiarities identifiable in our German corpus (cf. also Pirker and Loderer, 1999) are very similar to those identified in a number of studies for English (Levow, 1998, 1999; Oviatt, 1995). Thus, speakers were found to employ the following prosodic strategies:

- hyper-articulation
- syllable lengthening (e.g. Mon<L>day)
- pauses (between words and syllables, e.g. on <P> Thurs<P>day)
- stress variation
- variation of loudness
- variation of intonation contours
- laughter or sighing

In order to design automatic speech processing systems that can deal with really occurring speech properties, it is necessary to know what can be expected.

For instance, the example shows that the speaker regards slow speed, syllable lengthening (<L>), pausing (<P>), and strong emphasis (capital letters) as something that makes her speech easier to process for her communication partner:

(4) e4077101a: you didn't even let me finish. how do you know it's if it's occupied or not? <B> Monday <P> the eLEVenth <P> of JANuary <P> at twelve pm.

s4077102: vacation time is from the tenth of June till the fifteenth of July.

e4077102: no<L>, no <P> <<;slow> JAN<L>uary.> <P> <Swallow> <B> JANuary the <:<B> elEVenth:> <P> <B> at TWELve pm.

s4077103: will you please make a suggestion for an appointment?

e4077103: <Swallow> okay. <Swallow> let's try JANuARy <B> the <:<B> e<L>LEVenth:> <P> <B> at <P> TWELve pm.

If recognisers and dialogue managers are not designed to take into account the peculiarities that arise in real situations of miscommunication, since situations of communicative problems often trigger even more peculiarities, a vicious circle may arise that may lead to interruptions of the communication and, in the worst case, to the loss of a customer. That is, if malfunction occurs, and the speaker employs speech peculiarities either to increase the understandability of her utterances or to express her anger, the system, not being trained on such features, will understand even less Levow (1998).

Consequently, speech recognisers need to be adapted to the actually occurring, for instance, by being trained on data that include the phenomena arising. Alternatively, speech recognisers trained particularly on emotional data can be employed that are used as soon as emotional language is detected. This presupposes means to identify the moment at which the speaker get emotionally engaged (Batliner et al., 2003, see). Another possibility is to try to use dialogue management to calm down the angry user and to find other ways to prevent those emotional characteristics that are problematic for automatic speech processing systems from occurring; this is the perspective taken in this study.

3.2 Conversational Peculiarities in Reaction to System Malfunction

In emotional speech phonetic and prosodic characteristics are only one aspect of the expressive behaviour; speakers are furthermore also found to employ a number of conversational strategies that are peculiar to the situation of communicative problems, such as the repetitions investigated by Levow (1998, 1999); Oviatt (1995). Moreover, these behaviours may be easier to identify than increases of prosodic peculiarities (see Glockemann, 2003), in case emotional involvement in the speaker is to be monitored (Batliner et al., 2003). Here, what can be found are reformulations, additional specifications, meta-linguistic statements, new proposals without any relevant relationship to the previous utterances, thematic breaks, repetitions, and evaluations.

Reformulation

e4032301: the fifth of January, Tuesday <P> an appointment for five hours.

e403s4032302: I did not understand.

e403e4032302: an appointment on Tuesday January fifth <P> for five hours.

Metalanguage

e4022306a: Tuesday, January fifth, from eight o'clock until one o'clock.

s4022307: the first week of March is already occupied.

e4022307: I mean January fifth.

Additional Information

s4015104: please make a proposal.

e4015104: okay. <P> twelfth of January ninety-nine?

s4015201: I did not understand.

e4015201: the twelfth of January nineteen-ninety-nine?

New Proposal without Relevant Sequential Relation

e4022201: then the<L> twenty-second? <P> at <P> eight in the morning? <P> until two in the afternoon?

s4022202: vacation time is from the tenth of June till the fifteenth of July.

e4022202: <B> <P> uhm <P> on January fifth, at eight o'clock?

Thematic Breaks

e4072302: <Swallow> <B> <P> I have time on Thursday the twenty-first of January <P> <B> at two pm.

s4072303: the weekend is already occupied.

e4072303: <B> <Smack> okay. <P> <B> let's try <P> <B> okay, I have a another suggestion. how about Monday, <P> the eighteenth of January <P> <B> at <B> twelve pm?

Repetition

e4024101: January fourteenth, <P> from six until ten at night?

s4024102: vacation time is from the tenth of June till the fifteenth of July.

e4024102: January fourteenth, from six until ten at night?

Evaluation

e4075206: <Swallow> okay. <B> you're very busy for a computer. <P> <B> how about <B> Sunday the seventeenth of January <B> at <B> ten am?

The features described are likely to be due to emotional arousal; one reason may be that almost all speakers answered in the questionnaire that they filled out after the dialogues that they have been emotionally involved.

Second, there are systematic changes in the course of the dialogues, regarding their prosodic, conversational, and lexical properties. As shown in Figure 3, for instance, the conversational strategies are correlated with different phases of the dialogues, such that reformulations etc. occur more in earlier phases of the dialogue, associated with co-operative linguistic behaviour, whereas repetitions, rejections, and evaluations are located at the other end of the spectrum, occurring typically towards the end of the dialogues. Thus, when repetitions occur (see Levow, 1998, 1999), then this fact points to some dissatisfaction of the user already Fischer (1999b,c). The same holds for the prosodic and the lexical properties of the users' speech in the current corpus. For instance, regarding lexical material used, expressions with the German equivalent of *shit* are twice as frequent in the second half of the dialogues. Expressions with *Gott* [*god*] occur five times as frequently in the second half of the dialogues.

Third, if the lexical, conversational and prosodic strategies described were only due to the speakers' attempts to make themselves more understandable, there would be no prosodic peculiarities if the speakers feel that understanding is not at issue. For instance, rejections of proposals are understood as relevant answers:

(5) s0323202: dieser Termin ist schon belegt. [*this date is already occupied.*]

e0323202: aha. <B> wenigstens eine korrekte Antwort. [*uhuh. <B> at least a correct answer.*]

Although the speakers feel understood when the system rejects their proposals, the prosodic peculiarities of turns following rejections also increase throughout the dialogues (Fischer and Batliner, 2000). The changes observable can therefore not be solely attributed to different strategies to increase one's understandability, and thus emotional arousal must be involved as well.

Like speech recognizers, dialogue managers need to be designed to deal with the conversational characteristics of speech in the context of communicative problems, both regarding the recognition of the respective dialogue acts, and the capability to respond appropriately. Very few studies address the prevention or resolution of communicative problems in dialogue systems (cf. Batliner et al., 2003). Two aspects are relevant here: on the one hand the employment of expressive speech by the artificial agent, which can help calm down the user significantly, on the other hand the development of personality modelling on the basis of interpersonal differences observable in the data.

4 User Groups

If we now ask which measures can be taken to calm down an angry user and to prevent the situation from escalating, many possible behaviours of the dialogue manager could be used to guide the conversation even in the case of miscommunication. We may proceed by investigating whether all users behave in the same way and if not, what we can learn from those who display fewer problematic characteristics.

Sociolinguistic variables that have been found to often influence speech are age, gender, and social class. Unfortunately our data do not allow conclusions about social class, but about age and gender. As we shall see in the discussion of these two variables, it is more the attitude

the speakers display towards their communication partner and their perception of the situation than any extralinguistic speaker characteristics that may be relevant for user modelling. Thus, the following three subsections address age, gender, and speaker attitude respectively.

4.1 Age

In the use of expressive characteristics in HRI the speakers' age may be relevant; it is intuitively plausible that children may approach a robot in a much more playful way than an adult may. Accordingly, in experiments with children carried out in a similar set up like ours at the University of Erlangen, many more expressive speech characteristics can be found, compared to our experiments with adults:

(6) Ohm_21.062: Aibo steh auf [*Aibo get up*]

Ohm_21.063: brav so ist es brav lauf lauf geradeaus [*nice this is nice run run straight ahead*]

Ohm_21.064: lauf geradeaus Aibo hopp geradeaus laufen [*run straight ahead Aibo hopp straight ahead*]

Ohm_21.065: lauf [*run*]

Ohm_21.066: lauf Aibo Aibo lauf geradeaus [*run Aibo Aibo run straight ahead*]

Ohm_21.067: Aibo [*Aibo*]

Ohm_21.068: hörst du nicht geradeaus laufen Aibo [*don't you listen run straight ahead aibo*]

Ohm_21.069: brav so ist es brav [*nice this is nice*]

Ohm_21.070: lauf weiter [*go on*]

Ohm_21.071: Aibo steh auf [*Aibo get up*]

Ohm_21.072: Aibo [*Aibo*]

Ohm_21.073: steh auf [*get up*]

Ohm_21.074: steh auf [*get up*]

Ohm_21.075: b"oser Hund [*bad dog*]

Ohm_21.076: lauf [*run*]

Ohm_21.077: so ist es brav lauf [*this is nice run*]

Ohm_21.078: lauf weiter laufen [*run go on*]

Ohm_21.079: Aibo lauf lauf [*Aibo run run*]

Ohm_21.080: lauf Aibo [*run Aibo*]

Ohm_21.081: lauf [*run*]

Ohm_21.082: Aibo lauf [*Aibo run*]

Ohm_21.083: sitz Aibo Aibo sitz [*sit down Aibo Aibo sit down*]

Ohm_21.084: sitz [*sit down*]

Ohm_21.085: Aibo sitz [*Aibo sit down*]

Ohm_21.086: sitz Aibo [*sit down Aibo*]

In this example, the child uses the robot's name numerous times to get the robot's attention and to make it attend to his instructions. Furthermore, we find several evaluations of the robot's behaviour, such as *good dog* or *bad dog*. We also find interjections *na*, *oh Gott, ach herrje* or, as in this example, the secondary interjections *hopp* and *komm*.

According to Brown and Gilman (1962), address forms reveal aspects of the relationship between communication partners along the dimensions of power and solidarity. In particular, the informal T-forms can be distinguished from the more formal V-forms, the former expressing more equal and more solidary relationships, the latter being used in hierarchical and less solidary relationships. German distinguishes two forms of address, the informal *Du* and the formal *Sie*. The child in the above example, as all the other children as well, employs the T-form, expressing solidarity and an equal relationship, and very frequently the robot's (first) name.

The use of the robot's name is very typical for all the dialogues with the children. It may be an indicator that the children are building up a much stronger personal relationship with the robot than the adults in our corpus in a very similar setting did. Such a strong personal relationship becomes apparent in the next example in which the speaker, another boy, points out this relationship to the robot as a motivation to follow his instructions:

(7) Ohm_27.174: Aibo tanz [*dance*]

Ohm_27.175: mach's für mich bitte [*do it for me please*]

Ohm_27.176: oh wie lieb [*oh how kind*]

Actually, there is no child among the 26 children recorded who would not use the robot's name. There are two children who use it only in situations of 'disbehaviour', and all others use it consistently throughout the dialogues. Correspondingly, Batliner et al. (2004) report the name *Aibo* to be the most frequent word in the German child-aibo data.

Furthermore, many features of human-to-human dialogues occur in the dialogues with the children that are very rare in adult human-computer interaction. One such example are discourse particles, among them interjections, and modal particles. Thus, Batliner et al. (2004) report the modal particle *mal* and the secondary interjection *komm* to be even among the ten most frequent

words. This is quite surprising since the numbers of discourse and modal particles usually decrease in human-computer interaction (Hitzenberger and Womser-Hacker, 1995). The function of modal particles is to relate the current utterance to an assumed common ground, whereas the function of discourse particles is to mark an utterance as non-initial (Fischer, 2000a). The modal particles used by another child in the following example are *mal* and *schon*:

(8) Ohm_25.006: okay Aibo jetzt lauf mal [*okay Aibo now start running*]

Ohm_25.007: komm schon Aibo lauf [*come on Aibo run*]

Ohm_25.008: Aibo lauf [*Aibo run*]

In contrast, in our English and German adult-aibo dialogues, very few adults addressed the robot. For instance, the following speaker employs it after a successfully carried out task:

(9) turn right. –

turn right ? (2secs)

(at=loud)move forward(/at). -

move forward. (2secs)

(at=slow)ok(/at). –

(at=quiet)good robot(/at)

Furthermore, discourse particles and, in the German data, modal particles are extremely rare. A typical example from a German adult-aibo dialogue is the following:

(10) VP: rechts [*right*]

R: noncomply: straight ahead

VP: rechts, rechts, rechts [*right, right, right*]

R: comply: right

VP: vorwärts [*straight ahead*]

R: comply: straight ahead

VP: vorwärts [*straight ahead*]

R: comply: straight ahead

VP: vorwärts [*straight ahead*]

R: comply: straight ahead

VP: links [*left*]

R: comply: left

VP: vorwärts [*straight ahead*]

R: comply: straight ahead

VP: vorwärts [*straight ahead*]

However, it is possible that the differences between adults and children observable in these data are not exclusively due to speakers' age. In the adult-aibo scenario, the users were confronted with an experimental situation in which they had to give verbal instructions while there were three people, the leader of the experiment and two students operating the camera and taking notes, were present. In contrast, in the child-aibo data, the children had had the chance to get 'familiar' with the robot, and the robot was explicitely introduced to them by its name. In English child-aibo data using the same scenario, the British children displayed a very different linguistic behaviour (Batliner et al., 2004), the data being more similar to our adult-aibo dialogues. Part of the story may thus be that the children in the German dialogues with aibo experienced the experimental situation as very playful, which is supported by the fact that they all reported afterwards that they had had fun (even though the robot was as 'malfunctioning' as the robot in the other scenarios) and that they were made acquainted with the robot before the experiments.

Looking at another set of HRI data provides further evidence that variation between speakers cannot be simply traced back to extralinguistic factors, such as speaker age. In the second set of HRI data used here, adult users typed instructions into a notebook to instruct the robot and thus were much more private in their interaction with their artificial communication partner than in the previous scenario. Here it turns out that many users are much more playful. Some users address the robot like the children do, for instance:

(11) VP17-1: hallo roboter [*hello robot*]

sys:ERROR

VP17-2: hallo roboter [*hello robot*]

sys:ERROR

VP17-3: Die Aufgabe ist, den Abstand zu zwei Tassen zu messen. [*the task is to measure the distance to two cups.*]

sys:ERROR 652-a: input is invalid.

VP17-4: miß den Abstand zur Tasse genau vor Dir [*measure the distance to the cup right in front of you*]

sys:69,8 cm

(12) VP9-1: hallo roboter [*hello robot*]

sys:ERROR

VP9-2: wie kann ich entfernungen messen? [*how can I measure distances?*]

The following example stems from a speaker who uses not only the name of the robot and the T-form, but who also employs the modal particle *mal*:

(13) VP7-1: hallo roboter [*hello robot*]

sys:ERROR

VP7-2: miss mal die entfernung, roboter [*please measure the distance, robot*]

sys:ERROR

VP7-3: siehst du die tassen? [*do you see the cups?*]

However, in the following example, even though the user employs the T-form, the direct address is used as if it was an insult:

(14) VP4-28: Welchen Abstand haben die Tassen links hinten und hinten zueinander ? [*Which distance do the cups back left and back ?*]

sys:Unverstaendliche Eingabe. Bitte formulieren Sie neu. [*non-understandable input. Please reformulate.*]

VP4-29: Das ist nicht unverständlich, sondern sonnenklar, du Roboter. [*that's not non-understandable but completely clear, you robot.*]

sys:Bitte umformulieren! [*please reformulate!*]

VP4-30: Spinner. [INSULT]

However, not all users address the robot. As the following example shows, users sometimes believe that the robot does not even know where it is itself:

(15) VP10-1: das erste objekt ist das, das dir am nächsten ist [*the first object is the one which is closest to you*]

sys:ERROR

VP10-2: das erste objekt steht genau vor dir [*the first object is right in front of you*]

sys:ERROR

VP10-3: erstes Objekt: das am nächsten vor dem Roboter. [*first object: that is closest in front of the robot.*]

To sum up, in these dialogues users consistently use the informal T-form, and several users address the robot directly. Even though in later phases of the dialogues the robot uses 'natural language' and thus the formal V-form, users stick to the T-form.

In contrast, in the human-computer appointment scheduling dialogues, in which the computer's first utterance employs the V-form, the speakers consistently employ the V-form as well. Sometimes they switch to the T-form in the middle of the dialogues, explaining afterwards that they have thought about it and that they found that it does not make sense to use the formal form of address in the communication with an artificial agent. However, the following example shows that speakers may also use the different forms of address to mark on-stage versus off-stage talk:

(16) e0372301: nein, der siebte erste ist ein Donnerstag, du dumme Maschine. na, egal. machen wir den achten ersten als Termin ab. sind Sie damit einverstanden? [*no, the seventh of January is a Thursday, you (T-form) stupid machine. well, who cares. let's take the eighth of January. Do you (V-form) agree?*]

To conclude, it cannot be shown that a particular relationship with the robot, as evidenced by the forms of address used, can be directly related to particular user groups. That is, a relationship of solidarity and equality can not only be found in the child-aibo dialogues, but also in adult-computer and adult-pioneer interaction, where it may even be used consciously as a resource to mark an informal, or an off-topic, relationship. Although at first sight age seems thus to play an important role regarding the emotional involvement of the speakers, the perception of the situation and the artificial agent's output seem to be at least as relevant, as the different behaviours of the adults in the two human-robot andin the human-computer scenarios show. Thus, although we can see a bias of children towards approaching the human-robot interaction itself more playfully, it seems to be particularly important to predict how the users UNDERSTAND the situation.

4.2 Gender

A second often relevant extralinguistic variable is gender, and thus it may help to predict users' linguistic behaviour on the basis of their sex. However, if we investigate the use of conversational strategies between men and women in the human-computer dialogues, we find very few significant differences.

In the human-robot scenario in which users had to type instructions into a notebook in order to get the robot to measure distances between objects, different behaviours between males and females could be observed - yet, there was also the unfortunate coincidence that most of the male users were computer scientists, while most of the females were not. In those few cases in which a female was also a computer scientist, her language usage was much more similar to the male computer scientists than to the other females. Thus, the more important variable seems to be experience with artificial agents, but our data do not permit any reliable claims in this respect.

In contrast, our corpus of human-computer interaction allows us to make statistically valid assertions about the impact of the variable gender. For instance, in the use of reformulations (see Figure 2), women reformulate

slightly longer than the males (the difference in phases 3 and 4 being significant), but display the same linguistic behaviour in the last phase of the dialogue. To reformulate one's utterances for the system is co-operative behaviour, and thus women seem to be slightly longer co-operative than males are.

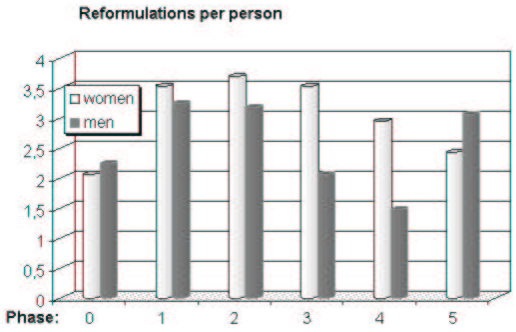


Figure 2: Reformulations by Women and Men

Conversely, issuing new proposals irrespective of the system's utterance is non-co-operative behaviour, and here women can be found to start a little later with this behaviour, as shown in Figure 3. Thus, for phase 3, there are significant differences, yet again women and men finish in the same way. Thus, the gender-specific behaviour just seems to be determined by a bit more patience with the system from the female side.

Figure 3: New Proposals by Women and Men

To conclude, gender differences, defined by the extralinguistic variable sex, do not seem to be particularly relevant in the communication with computers, besides the fact that the two sexes employ particular characteristics in different phases. This is partly in contrast with other findings on gender in human-computer interaction (see Fischer and Wrede, 1997). Nevertheless, much interpersonal variation in the dialogues can be observed, and so the question remains, if it is not age or gender, what this variation can be explained by.

4.3 Attitude

The last two sections have shown that extralinguistic factors may not be very helpful for predicting expressive speech characteristics. Instead, it may be useful to investigate the attitude that speakers employ towards the artificial agent. For instance, the occurrence of the prosodic peculiarities outlined above depends significantly on the speakers disposition. Thus, those who experience the situation as amusing will show significantly fewer prosodic peculiarities than those who experience it as annoying (Fischer, 2000b). Consequently, different user types can be identified on the basis of attitude.

I have developed a simple method for identifying different attitudes towards the system within the first utterance of the interaction; these attitudes can be shown to have direct consequences on the expressive characteristics of the speech directed towards the system. Thus, depending on the users' reaction to the question '*how do you do?*,' three user types can be distinguished: those who treat the computer as a tool and propose the first task instead of an answer, those who laugh and say '*fine*', and those who laugh, say '*fine*' and then ask the system '*and how do you do?*'. These three user groups can be significantly distinguished on the basis of their conversational behaviour on all linguistic levels.

Examples are the use of metalanguage and new proposals (that are not related to what has been previously discussed) in Figures 4 and 5. Here, we distinguished between players, those who pretend to have a normal conversation with the computer and thus ask it back politely about its well-being, and all others. In contrast to the two gender groups discussed above, here the differences between the two groups are significant for almost all phases of the dialogues.

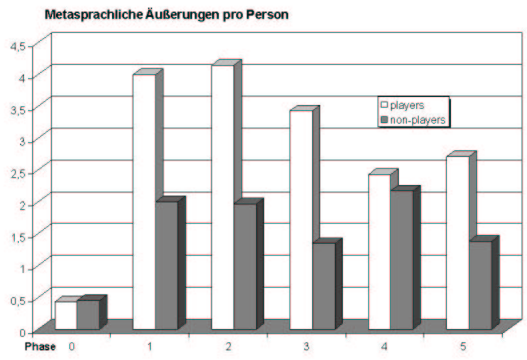


Figure 4: Metalinguistic Utterances by Players and Non-players

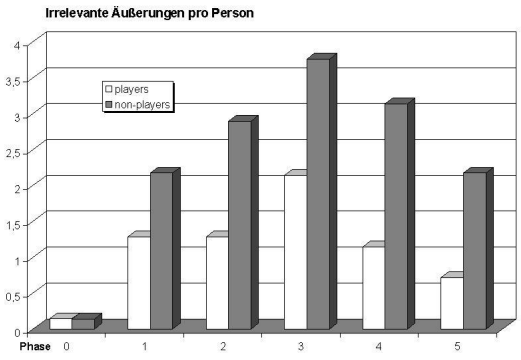


Figure 5: New Proposals by Players and Non-players

The way speakers conceptualise the situation and the artificial agent thus contributes significantly to the way they design their linguistic behaviour towards the system.

To sum up, in this section we have investigated some of the determinants of speakers' linguistic behaviour towards the system and in particular the role of extra-linguistic factors and the conceptualisation of the system as a communication partner. The results obtained, the fact that children are in general more likely to develop a playful and intimate relationship with a robot than adults are, that females tend to be more patient with artificial communication partners than males but otherwise do not differ very much in their linguistic behaviour, and that the speakers' attitude towards the system and the perception of the situation may be decisive for the users' design of their utterances and thus the expressive speech characteristics observable, can now be used to develop ways to guide the users' expressive speech behaviour.

5 Expressive Characteristics in System Output Design

In this section, two aspects will be discussed: on the one hand the role of expressive speech characteristics in system output in general, on the other the employment of system utterances to prevent those emotional features that are problematic for automatic speech processing.

Regarding the latter problem, in the HCI dialogues elicited, first steps towards influencing the users behaviour were taken. By experimenting with different system utterances, from directives like 'please speak more clearly, but not hyper-clearly' to excuses by the system, very different user reactions were obtained (cf. also Fischer and Batliner, 2000): while changes on the surface of the linguistic behaviour in reaction to directives were short-lived, and the peculiarities after the directive even increased, an excuse by the system for the communicative failures lead speakers to calm down immediately and for a longer time. Thus, approaching the users' linguistic behaviour on the level of the interpersonal relationship seems to be the most useful way of obtaining a linguistic behaviour that is unproblematic for automatic speech processing. This finding corresponds to the results obtained in this paper: Since the speakers attitude towards the system plays such a central role regarding the expressive properties of their speech, a useful way of guiding the users linguistic behaviour into something the system can deal with best may be located on the level of the attitude as well. Thus, using expressive characteristics in the robots or computers output may be one way to address the users emotional behaviour.

However, even if the system is embodied and employs expressive characteristics, it is not guaranteed that speakers will consider emotional expression as an appropriate part of the communication. In our HRI dialogues, some speakers took off their head sets which they wore for instructing the robot when they had to laugh about the robots behaviour. Thus, they regarded emotional expression to be off-topic. Another reason to be careful about implementing expressive behaviours in artificial agents may be that human-like properties of artificial agents may raise too high expectations. Thus, if human-like properties are being used, it has to be kept in mind that these must be functioning well - otherwise the opposite effect has been reported (see Bruce et al., 2001; Kanda et al., 2001).

6 Conclusion

Emotional speech, if it occurs unplanned in the context of system malfunction, may constitute a great problem for speech recognition and dialogue management. However, in order to prevent such problems, conversational behaviours by the artificial agent can be employed that may calm down an angry user. A number of such behaviours were proposed. Interestingly, the most effective behaviours are expressive behaviours as well. However, which measures should be taken depends essentially on the users' attitude towards the system, and thus personality modelling constitutes an important first step towards the users guidance by means of dialogue mana gement. How this can be done implicitly and online was exemplified in this paper as well.

Acknowledgements

Most of the research presented here was carried out in the framework of the SFB/TR8 'Spatial Cognition' at the University of Bremen, funded by the German Research Foundation. Many thanks furthermore go to Anton Batliner from the University of Erlangen for his co-operation in the corpus design and for his sharing of the data.

References

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to find trouble in communication. *Speech Communication*, 40(1-2), 2003.

A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russel, and M. Wong. 'you stupid tin box' – children interacting with the aibo robot: A cross-linguistic emotional speech corpus. In *Proceedings of LREC 2004*, 2004.

A. Batliner, R. Huber, H. Niemann, E. Nöth, J. Spilker, and K. Fischer. The recognition of emotion. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin etc.: Springer, 2000.

R. Brown and A. Gilman. The pronouns of power and solidarity. *American Anthropologist*, 4(6):24–29, 1962.

A. Bruce, I. Nourbakhsh, and R. Simmons. The role of expressiveness and attention in human-robot interaction. In *Proceedings of 2001 AAAI Fall Symposium*, 2001.

K. Fischer. Annotating emotional language data. Technical Report 236, Verbmobil, 1999a.

K. Fischer. Discourse effects on the prosodic properties of repetetitions in human-computer interaction. In *Proceedings of the ESCA-Workshop on Dialogue and Prosody, September 1rst - 3rd, 1999, De Koningshof, Veldhoven, The Netherlands.*, pages 123–128, 1999b.

K. Fischer. Repeats, reformulations, and emotional speech: Evidence for the design of human-computer speech interfaces. In Hans-Jörg Bullinger and Jürgen Ziegler, editors, *Human-Computer Interaction: Ergonomics and User Interfaces, Volume 1 of the Proceedings of the 8th International Conference on Human-Computer Interaction, Munich, Germany.*, pages 560–565. Lawrence Erlbaum Ass., London, 1999c.

K. Fischer. *From Cognitive Semantics to Lexical Pragmatics: The Functional Polysemy of Discourse Particles*. Mouton de Gruyter: Berlin, New York, 2000a.

K. Fischer. What is a situation? *Gothenburg Papers in Computational Linguistics*, 00-05:85–92, 2000b.

K. Fischer. Linguistic methods for investigating concepts in use. In Thomas Stolz and Katja Kolbe, editors, *Methodologie in der Linguistik*. Frankfurt a.M.: Peter Lang, 2003.

K. Fischer and Anton Batliner. What makes speakers angry in human-computer conversation. In *Proceedings of the Third Workshop on Human-Computer Conversation, Bellagio, Italy*, 2000.

K. Fischer and B. Wrede. Discourse particles in female and male human-computer-interaction. In Milton Keynes De Montford University, editor, *Women into Computing*, pages 36–49, 1997.

M. Glockemann. Methoden aus dem bereich des information retrieval bei der erkennung und behandlung von kommunikationsstörungen in der natürlichsprachlichen mensch-maschine-interaktion. Master's thesis, University of Hamburg, 2003.

L. Hitzenberger and C. Womser-Hacker. Experimentelle Untersuchungen zu multimodalen natürlichsprachigen Dialogen in der Mensch-Computer- Interaktion. *Sprache und Datenverarbeitung*, 19(1):51–61, 1995.

T. Kanda, H. Ishiguro, and T. Ishida. Psychological analysis on human-robot interaction. In *IEEE International Conference on Robotics and Automation ICRA*, 2001.

G.-A. Levow. Characterizing and recognizing spoken corrections in human-computer dialogue. In *Proceedings of Coling/ACL '98*, 1998.

G. A. Levow. Understanding recognition failures in spoken corrections in human-computer dialogue. In *Proceedings of the ESCA-Workshop on Dialogue and Prosody, September 1rst - 3rd, 1999, De Koningshof, Veldhoven, The Netherlands.*, pages 123–128, 1999.

S. Oviatt. Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, (9):19–35, 1995.

S. Oviatt, J. Bernard, and G.-A. Levow. Linguistic adaptations during spoken and multimodal error resolution. *Langauge and Speech*, 41(3-4):419–442, 1998a.

S. Oviatt, M. MacEachern, and G.-A. Levow. Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication*, 24:87–110, 1998b.

H. Pirker and G. Loderer. I said "two ti-ckets": How to talk to a deaf wizard. In *Proceedings of the ESCA Workshop on Dialogue and Prosody, September 1rst - 3rd, 1999, De Koningshof, Veldhoven, The Netherlands*, pages 181–186, 1999.

B. Tischer. *Die vokale Kommunikation von Gefühlen*. Weinheim: Beltz, 1993.

A Probabilistic Hierarchical Framework for Expression Classification

Lukasz Zalewski and Shaogang Gong*
*Department of Computer Science, Queen Mary, University of London
London, E1 4NS, UK
[lukas|sgg]@dcs.qmul.ac.uk

Abstract

We address the issue of understanding facial expressions through statistical modelling and analysis without the need for any temporal information whatsoever. We introduce a hierarchical decomposition of a human face into different subcomponents where each of them is modelled using Probabilistic PCA (PPCA). Classification is performed by fusing all the subcomponent information with a Hybrid Bayesian Network (HBN) to provide parameterised output which we use to animate the avatar.

1 Introduction

A human face can exhibit complex and intricate expressions. Facial expression changes are dependent on many factors such as muscle contractions, current emotional state and its implied context. Also facial expressions are individually independent: no two people exhibit the same expression in the same way. These factors make modelling and recognising facial expressions a challenging task.

Bettinger et al. (2002) used AAM (Active Appearance Model) as the underlying basis of their model, sample mean shift and variable length Markov model, to learn the relationships between trajectories of facial expressions, Devin and Hogg (2001) combined AAM with sound as their framework to produce sequences of a talking head. Both approaches do not deal with the expression classification directly. Cohen et al. (2002) used a model based on the motion vectors of Bezier volumes. These vectors were then used in conjunction with a multi-level HMM to classify expression from image sequences. They also experimented with static Bayesian Networks (BN). Chuang et al. (2002) used statistical appearance representation (similar to Cootes and Taylor (2001)) to represent facial expression configurations, then a factorised bilinear model to synthesise existing sequences with different expressions during the speaking process. Tian et al. (2001) used FACS (Facial Action Coding System Ekman et al. (1972)) and a neural network to perform detailed classification of facial expressions. Their approach does not deal with the self occlusion and relies on the detailed geometrical measurements to describe different features which is unreliable.

In this work we wish to model the semantics of a set of low-level facial behaviours, or states which include neutral, smile, grin, surprise, fear, sadness and anger. We aim to model the intrinsic inner-expression relationships by placing hierarchical constraints to bootstrap the process to help in classification of facial expressions. In contrast to Tian et al. (2001); Chuang et al. (2002), we provide one compact and unified probabilistic framework for such a task. Our facial appearance under varying expressions is based on a statistical appearance model originally introduced by Cootes and Taylor (2001). We extend the basic definition of the AAM model to implicitly incorporate parameters for large pose variations into the statistical distribution. Our model is also equipped with a pose estimator to bootstrap the tracking process during large pose changes.

Facial expression classification is achieved by two components: 1) hierarchical shape model, onto which the current instance is projected, where the face is decomposed into the root component consisting of jaw outline, centroids of the eyes and mouth and nose outline. The children are defined as left eye and left eyebrow (eyeL), right eye and right eyebrow (eyeR) and mouth. The children are modelled using PPCA (Section 2.1) and built with frontal view only, letting the pose parameters, such as rotation and translation be inherited from the root component; 2) Hybrid Bayesian Network which fuses all the information to produce the final output. Figure 1 depicts an overview of our system.

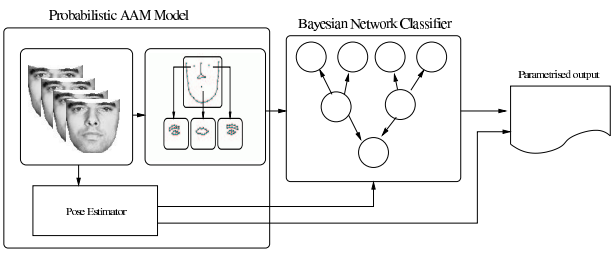


Figure 1: General overview of our system.

2 Framework

The basic representation of an AAM is only able to cope with frontal/near-frontal views ($[-20^o, 20^o]$ in yaw). At the extreme pose changes due to occlusion during the warping process distorts the texture, creating large residuals and causing tracking failure. To overcome this problem we use the pose estimator (Section 2.2) to obtain yaw rotation and mirror the warped image when necessary ($-15^o < yaw > 15^o$). A similar approach was used by Dornaika and Ahlberg (2003). Figure 2 shows the original images from a sequence (top row), frontal view warped texture vectors, with visible distortions (middle row) and pose corrected frontal view texture vectors (bottom row).

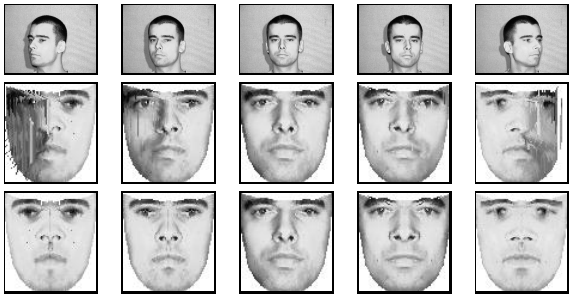


Figure 2: Distortions due to the pose changes and self-occlusion. Top row: original images, middle row: frontal view warped images, bottom row: pose corrected frontal view morphed images.

Unfortunately, mirroring provides only an approximation to the true representation of the face at extreme views. To further improve the tracking process we introduce a pose corrected weight vector such that the original texture difference $\Delta\mathbf{T} = \mathbf{T}_{im} - \mathbf{T}_m$ becomes $\Delta\mathbf{T}_{corr} = \mathbf{T}_w \otimes \Delta\mathbf{T}$, where $\mathbf{T}_{im}$, $\mathbf{T}_m$ are the texture instances in the image and model frame respectively, $\mathbf{T}_w$ is the pose dependent weight vector drawn from the normal distribution and $\otimes$ is component-wise multiplication. Figure 3 shows different representations of $\mathbf{T}_w$ with respect to different yaw rotation values.

Also during the training stage we find the relationship between yaw rotation and the model component responsible for yaw changes by fitting second order polynomial to the data (we are only interested in yaw rotation it is the most likely cause of self-occlusion). During the model fitting stage we use it to provide a model prediction, such that:

$$t_p = a\alpha_h^2 + b\alpha_h + c \tag{1}$$

where α_h is the yaw rotation for the current hypothesis, a, b, c are the coefficients of the polynomial and t_p is the predicted pose parameter such that $t_p \in [-E_h std_{pse}, +E_h std_{pse}]$ with E_h being the residual error of the current hypothesis and std_{pse} being the standard deviation for the given pose component. Experimental results of the pose corrected AAM tracking are presented in Section 4.

Figure 3: Different weight vector representations for different yaw rotation values.

2.1 Probabilistic PCA

PCAs lack of probability distribution makes it-ill suited for the Bayesian framework. Tipping and Bishop (1998) reformulated PCA as the maximum likelihood solution using a latent variable model such that the observed variable $\mathbf{t}$ is given by:

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{2}$$

where $\mathbf{x}$ is the latent variable such that $P(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I}_q)$ and $\mathcal{N}$ denotes a Gaussian distribution, $\mathbf{W}$ is the parameter matrix whose columns define the principal subspace of the data, $\boldsymbol{\mu}$ is the d-dimensional vector, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$ where σ^2 is the noise variance, $\mathbf{I}$ is the identity matrix and $\mathcal{N}$ represents a Gaussian distribution. Then

$$P(\mathbf{t}|\mathbf{x}) = \mathcal{N}(\mathbf{t}|\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \sigma^2\mathbf{I}_d) \tag{3}$$

Marginal distribution of the observed variable $\mathbf{t}$ is

$$P(\mathbf{t}) = \int P(\mathbf{t}|\mathbf{x})P(\mathbf{x})d\mathbf{x} = \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}) \tag{4}$$

where covariance matrix $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}_d$. The above model represents a constrained Gaussian distribution controlled by $\boldsymbol{\mu}, \mathbf{W}$ and σ^2. A maximum likelihood solution for the parameters is given by:

$$\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{t}_i \tag{5}$$

$$\mathbf{W}_{ML} = \mathbf{U}_q(\boldsymbol{\Lambda}_q - \sigma^2\mathbf{I}_q)^{\frac{1}{2}}\mathbf{R} \tag{6}$$

$$\sigma^2_{ML} = \frac{1}{d-q}\sum_{i=q+1}^{d}\lambda_i \tag{7}$$

where $\mathbf{t}_i$ is the i-th d-dimensional feature vector from the data set, $\boldsymbol{\Lambda}_q$ is the diagonal matrix containing the q-largest eigenvalues λ_i, $\mathbf{U}_q$ is the matrix containing the q-largest eigenvectors and $\mathbf{R}$ is an arbitrary orthogonal rotation matrix.

2.2 Pose Estimator using PPCA

The pose estimator provides us with continuous 3D pose estimation based on a probabilistic framework. We use a sparse set of training samples, that cover only part of the view sphere, $(-40^o, 40^o)$ around yaw and $(-20^o, 20^o)$ around pitch (using 10^o intervals) and are able to estimate the pose for a much larger, continuous view sphere. The model is also able to generalise to a much denser shape model for appearance synthesis at virtual views. Additionally we do not need to utilise any temporal information such that the estimation is done on-the-fly frame-wise in real time and the system is able to cope with very large jumps and discontinuities in pose change.

Given a d-dimensional multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$ its marginal q-dimensional marginal multivariate distribution (where $q << d$) is also Gaussian (Krzanowski, 1988). Let $\mathbf{B}$ be a qxd dimensional identity matrix ($\mathbf{B} = \mathbf{I}$). Then the marginal q-component multivariate probability distribution function (p.d.f) f_q is given by:

$$f_q \sim \mathcal{N}(\mathbf{B}\boldsymbol{\mu}, \mathbf{B}\mathbf{C}\mathbf{B}^T) \tag{8}$$

Following the concept of the marginal p.d.f we define the cumulative distribution function (c.d.f) Φ, such that for a q-dimensional random variable $\mathbf{x}$ the Gaussian c.d.f is given by:

$$\Phi(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} f_q(\mathbf{x}) d\mathbf{x} \tag{9}$$

We are mostly interested in the c.d.fs that are closely related to the components responsible for the yaw and pitch rotation. Let f_{my}, f_{mx} be marginal p.d.fs and Φ_{my} and Φ_{mx} be marginal c.d.fs corresponding to pose changes. For a given shape $\mathbf{t}$ the estimate of the yaw rotation r_y is given by Equation (10), where a_1, a_2, a_3, a_4 are coefficients of a cubic polynomial estimated during the training stage, p_y is the marginal cdf for the yaw rotation and ϵ_y is the error term defined by constant weighted by the marginal probability f_{my}:

$$\begin{aligned} r_y &= a_1 * p_y^3 + a_2 * p_y^2 + a_3 * p_y + a_4 + \epsilon_y \\ p_y &= \Phi_{my}(\mathbf{t}) \\ \epsilon_y &= f_{my}(\mathbf{t}) * const \end{aligned} \tag{10}$$

The estimate of the pitch rotation r_x is given by Equation (11) where b_1, b_2, b_3, b_4 are coefficients of a cubic polynomial estimated during the training stage, p_x is the marginal cdf of the pitch rotation and ϵ_x is the error term defined by constant weighted by the marginal probability f_{mx}:

$$\begin{aligned} r_x &= b_1 * p_x^3 + b_2 * p_x^2 + b_3 * p_x + b_4 + \epsilon_x \\ p_x &= \Phi_{mx}(\mathbf{t}) \\ \epsilon_x &= f_{mx}(\mathbf{t}) * const \end{aligned} \tag{11}$$

To find the relationship between the angles and the c.d.fs, we use the posterior distribution of the PPCA model. PPCA is described in Section 2.1. We have found that such a probabilistic framework provides much more accurate estimation than one using conventional PCA (e.g. by finding the relationship between the projected parameters and angles). Our model is able to generalise to a denser shape model (Figure 4). This is achieved by down-

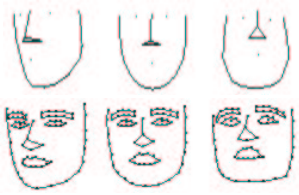


Figure 4: Denser shape model (74 landmarks) driven by the sparse set using 14 landmark points.

sampling the larger PDM to the required size by calculating the centroids of the eyes and mouth and selecting the subset of the jaw outline.

2.3 Hierarchical Shape Representation

We define a hierarchical decomposition of the shape as follows: The jaw outline, nose and centres of the eyes and mouth form the root of our hierarchy. As leaves, or children, we have eye and eyebrow pairs and mouth. Figure 5 shows an example of such decomposition.

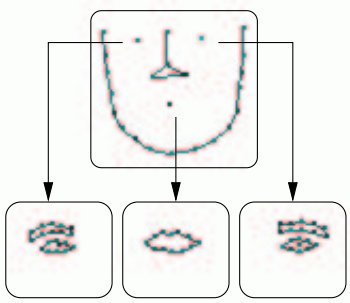


Figure 5: The top row corresponds to the highest point in the hierarchy (root), the middle row corresponds to the leaves.

We allow pose parameters to be incorporated into the root model, and let the children to be frontal view only, with the pose parameters (rotation, translation) inherited from the parent. If the instance of the model at the current frame j has the rotation parameters given by $\mathbf{p}_j = (\alpha, \beta, \gamma)^T$ representing yaw, pitch and bank rotation respectively, the frontal representation $\mathbf{F}_j^i$ for a given instance $\mathbf{M}_j^i$ where $i \in \{eyeL, eyeR, mouth\}$ is then given by:

$$\mathbf{F}_j^i = \mathbf{R}(\mathbf{p}_j)(\mathbf{M}_j^i - \mathbf{T}^i) \tag{12}$$

where $\mathbf{T}^i$ is the translation obtained from the root for i-th leaf. Thus our face can be represented as a combination of the subcomponents. Given the instances of the hierarchical subcomponents, any arbitrary view/expression can be represented as:

$$\mathbf{x}_{final} = B(\mathbf{x}_{root}, \mathbf{x}_{eyeR}, \mathbf{x}_{eyeL}, \mathbf{x}_{mouth}) \quad (13)$$

where B is a shape blending function, $\mathbf{x}_{root}$ is the root model and $\mathbf{x}_{eyeR}, \mathbf{x}_{eyeL}, \mathbf{x}_{mouth}$ are right eye, left eye and mouth models respectively. Our motivation for choosing hierarchical representation of the shape model is as follows: We strongly believe that the shape is individually independent (given appropriate normalisation) and can be efficiently utilised to capture manifolds of the facial expressions.

We also noticed that the modes of variation for each of the components correspond to their intrinsic functionality. For example for mouth they are mouth open, mouth closed and mouth grin. We associate corresponding marginal and cumulative probability distributions with each of the functionalities. Instead of defining facial expressions as holistic entities, we represent them as a combination of intrinsic functionalities of the subcomponents (expression implied facial feature independence has been exploited by Donato et al. (1999); Zalewski and Gong (2004)). So any facial expression can be defined as:

$$expression = state_{mouth} + state_{eyeR} + state_{eyeL}$$

The advantages of this are two-fold: First of all, each of the expressions is defined in a more intuitive and quantitative way. Secondly, such a representation allows us to account for similar expressions (smile with eyes open, or smile with eyes closed) without any additional overhead. Given probability distributions for each of the subcomponents, we obtain final classification by fusing all the information through a Hybrid Bayesian Network (Section 3), hence producing a parameterised form of expression definition.

3 Expression Classification

Bayesian Networks allow us a way for data fusion in a probabilistic fashion, and have been successfully used in face recognition and classification related tasks (Yand et al., 2002). As the basis of the classifier, we adopt a Hybrid Bayesian Network (HBN) (Figure 6).

Round nodes correspond to continuous states, square ones to the discrete states. Shaded nodes are observed, unshaded ones are hidden. In the design of this HBN we took into consideration psychophysical evidence implied by the human perception of facial expressions (Ekman, 1973; Ekman et al., 1972). Such a layout gives us the means to describe the states of each of the subcomponents at a high abstraction level that in turn can be used as a parameterised output for animation purposes.

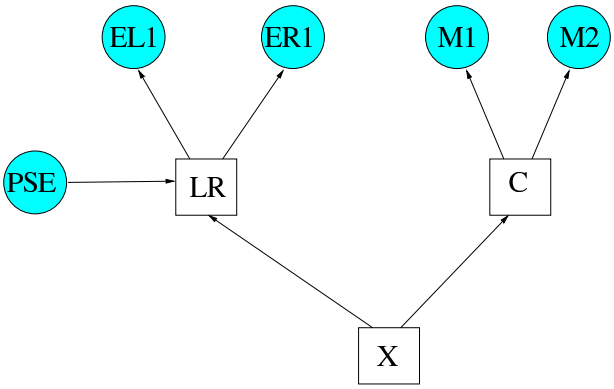


Figure 6: The Hybrid Bayesian Network used for a paramtrised expression definition.

We derive logical sections within the net that characterise the functionalities of the different facial components. These are eye components defined by likelihoods $P(EL|LR)$, $P(ER|LR)$, and mouth component defined by likelihoods $P(M1|C)$, $P(M2|C)$, which are drawn from our hierarchical model such that

$$\begin{aligned} P(EL|LR) &= f^1_{eyeL}(\mathbf{t}_{eyeL}) \\ P(ER|LR) &= f^1_{eyeR}(\mathbf{t}_{eyeR}) \\ P(M1|C) &= f^1_{mouth}(\mathbf{t}_{mouth}) \\ P(M2|C) &= f^2_{mouth}(\mathbf{t}_{mouth}) \end{aligned}$$

where f^j_i defines the posterior marginal probability distribution for the j-th principal component, $i \in \{eyeL, eyeR, mouth\}$ and $\mathbf{t}_i$ is the input vector. The prior $P(PSE)$ is drawn from the marginal distribution of the pose model such that:

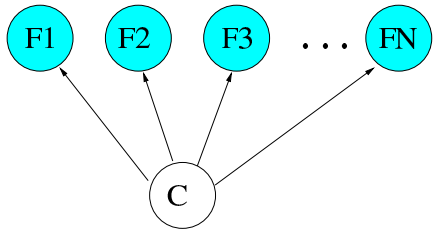
$$P(PSE) = f^1_{pose}(\mathbf{t}_{pose})$$

and accounts for the missing features in extreme pose changes. If one of the features becomes occluded, the remaining visible feature, not the combination of both, will be used for classification.

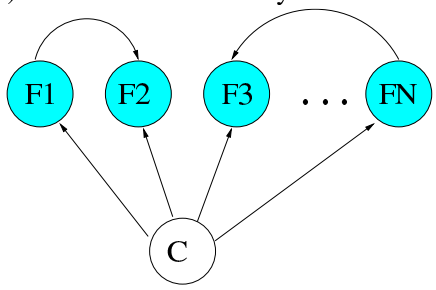
We can think of the output nodes as descriptors of different features on the face, which are independent of each other (this is implied by orthogonality of our distribution spaces). As Cohen et al. (2002) pointed out, the main limitation of feature based Naive Bayesian Classifiers is the independence of the features given the expression which might not be true in real life scenarios. (Figure 7 (a)) depicts such a Naive Bayesian Classifier, where $F1, F2, F3, ..., FN$ define different features and C is the expression class. To overcome that limitation they suggested use of a TAN classifier (Tree Augmented Naive) where dependencies are represented as arcs between different features, and its structure in defined is the learning stage (Figure 7 (b)).

In our case, to account for dependencies amongst different facial features we introduce hidden nodes LR and

C. During training these nodes will capture the possible dependencies among different feature inputs (LR for EL, ER and C for $M1, M2$).



(a) Feature-based Naive Bayesian Classifier



(b) Feature-based TAN Classifier

Figure 7: Different representations of BNs.

To perform final classification we choose the hypothesis that maximises the posterior given the evidence Θ and the net structure $\mathbf{m}$:

$$P(\mathbf{m}|\Theta) = P(X|LR, C, A, B, EL, ER, M1, M2, PSE, \Theta) \propto P(X, LR, C, A, B, EL, ER, M1, M2, PSE|\Theta)$$

To train the Hybrid Bayesian Net we performed supervised training based on 813 hand labelled samples representing continuous sequences of various changes in facial expressions.

4 Experiment

For AAM model training we used a set consisting of 1790 images and shapes (74 landmarks), which included seven basic expressions (neutral, smile, grin, sadness, fear, anger, surprise) and large variations in pose. For the frontal view, a hierarchical decomposition shape model, a training set of 700 shapes was used. For the pose estimator, 640 different and much sparser shapes (14 landmarks) were used. All the training samples were hand labelled beforehand.

The outline of our algorithm is as follows:

- Perform colour segmentation on the input image to find a rough position of the head and remove unnecessary background information. For colour segmentation, the HSV (Hue,Saturation,Value) colour model is used, which carry sufficient discriminative information for such a task.
- Fit the pose-corrected AAM model representation to the image. Obtain a pose estimate through from the shape model (repeat both until convergence).
- Obtain final pose estimate.
- Project the current instance of the model onto the hierarchical shape model.
- Classify the expression using the Hybrid Bayesian Network and obtain a parameterised output.
- Animate Avatar according to the obtained parameterised output.

The Avatar animation was performed using a morph-based approach (Noh and Neumann, 1998), defined by:

$$Expression = \sum_i \mathbf{w}(i)\mathbf{\Gamma}(i) \qquad (14)$$

where $\mathbf{w}$ defines the morph weight vector such that $\sum_i \mathbf{w}(i) = 1$ and $\mathbf{\Gamma}$ defines a set of morph bases. To test our improved AAM representation we used a test sequence 0 containing 415 frames and large pose changes. Figure 8 shows the first few frames. Top row corresponds to the original AAM formulation, with visible loss of focus at large pose variation, and the bottom row corresponds to the pose corrected AAM, where the focus is not greatly affected by pose changes.

To test the expression classifier we used two sequences containing 750 and 530 frames respectively. We compared this with the BN given in the Figure 7 (3 input nodes, taking the parameter vectors from the hierarchical distribution). We obtained the following classification rates:

	Our HBN	Cohen et al. (2002)
test sequence 1	88%	82%
test sequence 2	83%	80%

Figure 11 shows selected frames from the sequence 1 experiment. Within each of the boxes the left image corresponds to the currently tracked image frame with the AAM mask superimposed on it. The image on the right corresponds to the synthetic avatar animated according to the classified expression. Figure 9 shows the corresponding classification results for test sequence 1 and Figure 10 for test sequence 2.

5 Conclusion

In this paper we have extended the basic AAM approach to cope with large pose variations by introducing pose-based constraints upon the tracking process. We also introduced shape based hierarchical decomposition of a human face into independent components, such that their

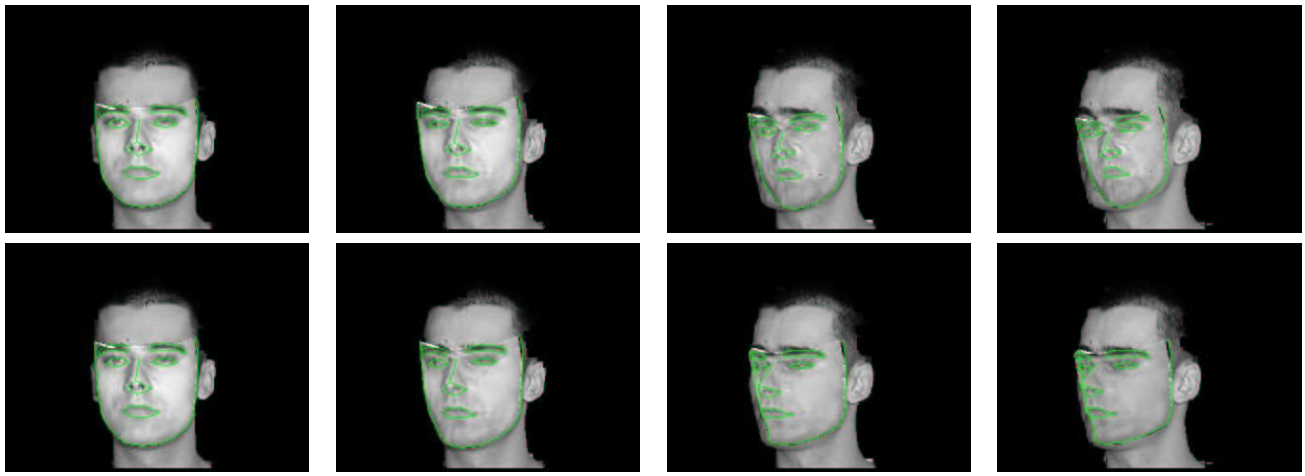


Figure 8: Selected frames from the experiment on the AAM fitting onto the extreme pose view. Top row corresponds to the original AAM formulation and bottom row to the pose corrected AAM.

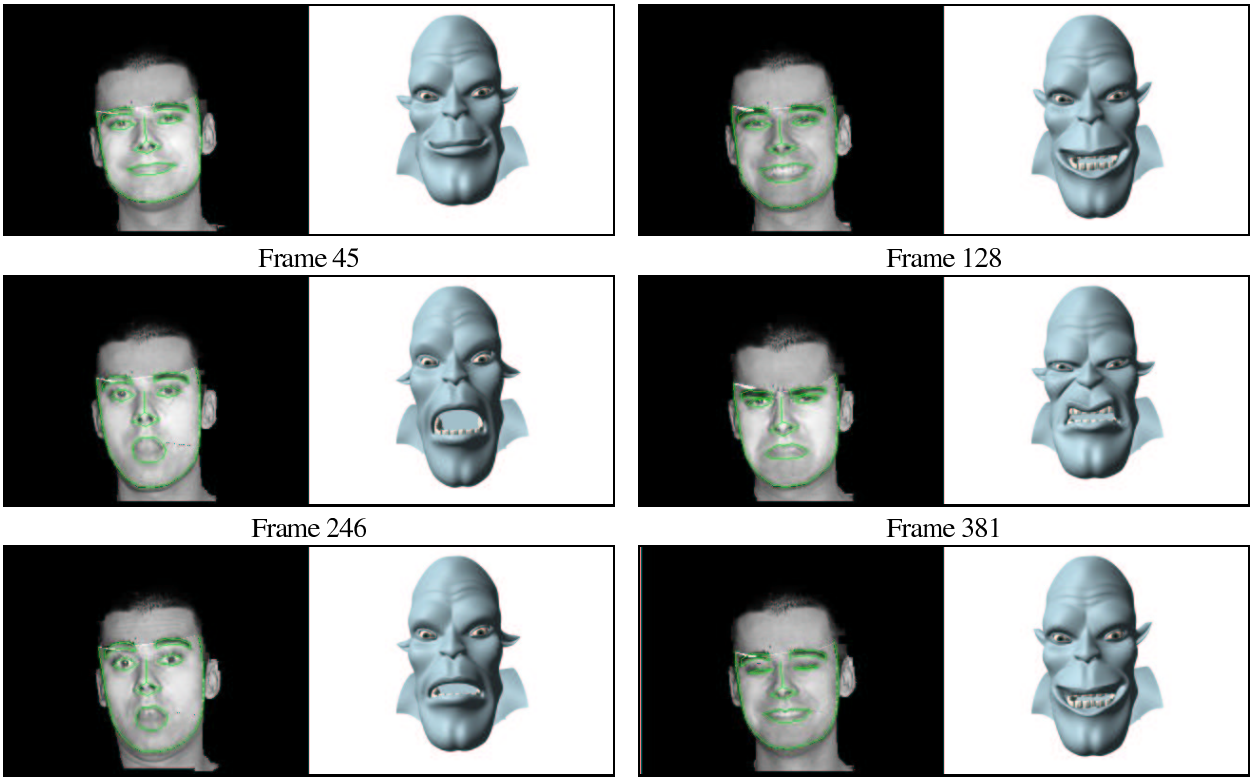


Figure 11: Selected frames from the experiment on expression classification and avatar animation (test sequence 1). Each of the images shows tracked frame with AAM mask superimposed on it (left) and corresponding synthesised avatar (right).

combinatorial form can be used to define an arbitrary facial expression, and the probabilities obtained from their distributions in conjunction with Hybrid Bayesian Network (HBN) can serve as a basis for expression classification. Our future work includes investigation into Dynamic Bayesian Networks (DBNs) and their use in behaviour context modelling.

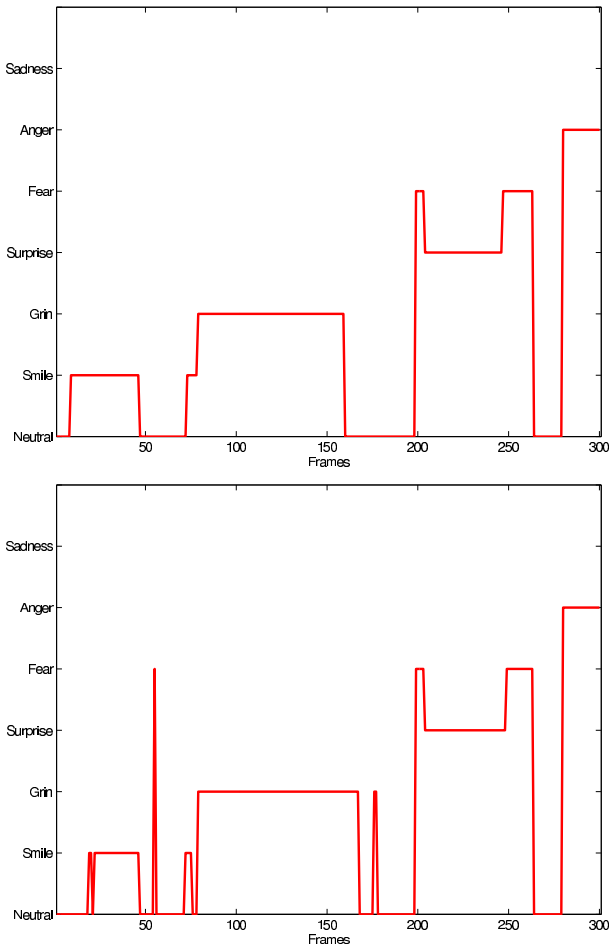


Figure 9: Expression classification for test sequence 1: top row corresponds to our HBN approach, bottom row to the Cohen et al. (2002) approach.

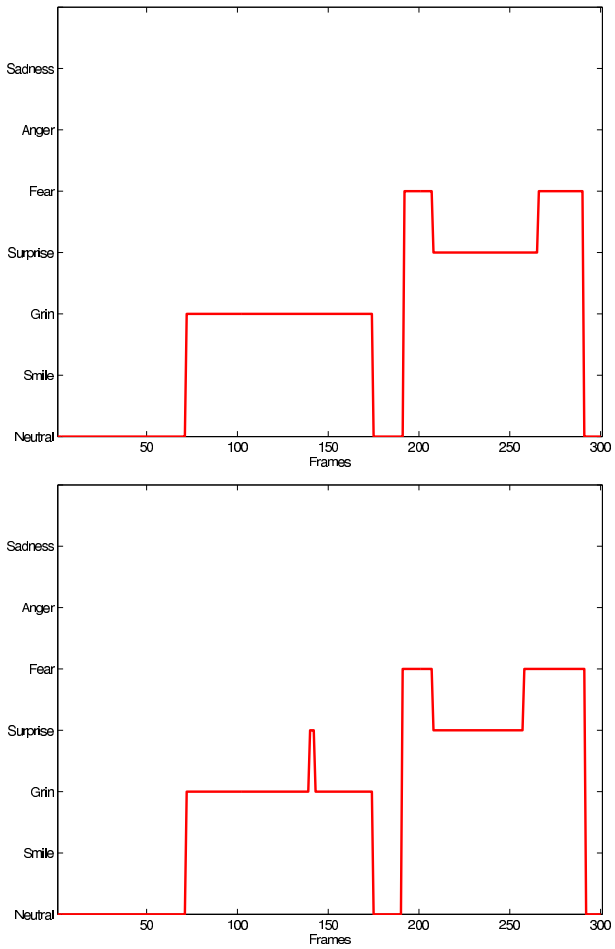


Figure 10: Expression classification for test sequence 2: top row corresponds to our HBN approach, bottom row to the Cohen et al. (2002) approach.

Acknowledgements

We would like to thank Jose Galan for enlightening discussions on Bayesian Networks.

References

F. Bettinger, T. F. Cootes, and C. J. Taylor. Modelling facial behaviours. In *BMVC*, volume 2, pages 797–806, 2002.

E. S. Chuang, H. Deshpande, and C. Bregler. Facial expression space learning. In *10th Pacific Conference on Computer Graphics and Applications*, Beijing, 2002.

I. Cohen, N. Sebe, L. Chen, A. Garg, and T. S. Huang. Facial expression recognition from video sequences. *International conference on Multimedia and Expo*, 2:121–124, 2002.

T. F. Cootes and C. J. Taylor. Statistical models of apperance for computer vision. Technical report, University of Manchester, Manchester, UK, 2001.

V. E. Devin and D. C. Hogg. Reactive memories: An interactive talking head. In *BMVC*, 2001.

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

F. Dornaika and J. Ahlberg. Efficient active appearance model for real-time head and facial feature tracking. *2003 IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 173–180, October 2003.

P. Ekman, editor. *Darwin and facial expressions: A century of research in review*. Academic Press New York, 1973.

P. Ekman, W. V. Frieser, and P. Ellsworth. *Emotion in the human face*. Pergamon New York, 1972.

W. J. Krzanowski. *Principles of Multivariate Analysis*. Oxford University Press, 1988.

J. Noh and U. Neumann. A survey of facial modeling and animation techniques. Technical report, University of Southern California, 1998.

Y. Tian, T. Kanade, and J. F. Cohon. Recognising action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):1–19, February 2001.

M. E. Tipping and C. M. Bishop. Mixture of probabilistic component analysers. Technical report, Dept of Computer Science and Applied Mathematics Aston University, Birmingham B4 7ET, UK, 1998.

M. Yand, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 34–58, 2002.

L Zalewski and S. Gong. Synthesis and recognition of facial expressions in virtual 3d views. In *Proc. 6th IEEE International Conference on Automatic Face and Gesture Recognition*, Korea, 2004.

Modelling character emotion in an interactive virtual environment

El Jed Mehdi, Pallamin Nico, Dugdale Julie and Pavard Bernard
* GRIC-IRIT
Groupe de Recherche en Ingénierie Cognitive - Institut de Recherche en Informatique de Toulouse
UPS-CNRS (UMR 5505),118, Route de Narbonne 31062 Toulouse Cedex. France.
{eljed | pallamin | dugdale | pavard}@irit.fr

Abstract

Expressing emotions plays a critical role in increasing the believability of virtual characters. Our aim is to develop an emotional model that can be integrated into characters in an interactive virtual reality environment. We have adopted mood, personality and emotion in an approach where the human users animates their own virtual character. This approach allows the user to employ all of their social and cultural skills in both decision-making and in interaction with other agents. The model so far developed is currently being integrated into a virtual reality training tool for fireman.

1 Introduction

Expressing emotions plays a critical role in increasing the believability of virtual characters. Our aim is to develop an emotional model that can be integrated into characters in an interactive virtual reality environment in order to improve their expressiveness and situatedness.

Such a model will allow virtual actors to interact using non-verbal communication skills influenced by their emotional state, personality and mood in order to provide a sufficient feeling of immersion and thus reproduce an efficient simulation of human activities.

This aim poses problems of representation of this non intentional and emotional dimension on a virtual actor. Until now these problems have been treated essentially in the paradigm of implementing fully autonomous agents (Breazeal, 2003), (Gratch and Marsella, 2001).

This paradigm is clearly representational because of its aim to explicitly model all dimensions of cognition or social interaction. However, this approach is strongly limited in terms of realism because of the inherent complexity of modelling human social interactions and our lack of understanding of complex human cognitive expertise.

Our approach instead is based on the mixed agent paradigm as proposed by (Pavard and Dugdale, 2002) where agents who have an important role (important interaction with other agents in the environment) are animated by humans. What is interesting in this approach is that it allows the user to interact in the virtual reality environment according to his or her own perceptions, objectives, knowledge, history and culture.

2 State of the art

There is currently a large research effort devoted to creating virtual characters able to communicate in a human-like way. Much of this work is concerned with the implementation of personality and emotion in a virtual character to enable it to interact with humans using natural language, gestures and facial expressions in an expressive and appropriate way.

One of the most widely accepted models of emotion was proposed by Ortony, Clore and Collins and is commonly known as the OCC Model (Ortony et al., 1988).

Systems such as 'Emile' (Gratch, 2000) implement the OCC model to embody emotion in their virtual characters. This system and others (e.g. (Marsella and Gratch, 2001)) are strongly Artificial Intelligence (AI) oriented and are based on a planning system that generates emotions according to an assessment of risks and which progresses towards a series of predefined goals. More recently, Marsella and Gratch have integrated coping strategies with the OCC model to try to explain how people deal with strong emotions (Marsella and Gratch, 2002).

Concerning the implementation of personality, Ball and Breese (Ball and Breese, 2000) have experimented with Bayesian Belief Networks (BBN) to model emotion and personality. A similar approach has been adopted by Kshirasagar (Kshirsagar and Magnenat-Thalmann, 2002) who also uses BBN for modelling personality and emotion but introduces an additional layer of "mood" in the model. A generic model for personality and emotion has also be presented by (Egges et al., 2004). Models of personality and emotion have been used as well to emphasize different aspects of the affective agent-user interface in three projects at the DFKI (André et al., 1999).

In this paper we have adopted a different approach in modelling emotion, mood, personality and behavior. In this approach, the human user animates their own virtual character. This approach allows the user to employ all of their social and cultural skills in both decision-making and in interaction with other agents. In this way, we avoid the problem of implementing a complete cognitive sys-

tem and we focus only on the reduced set of events that can induce an emotional change on the character.

The main advantage of this approach is its simplicity in the ease of implementation as well as the consequent reduction of speed of execution allowing its use in Real Time virtual training.

3 Emotional model

In this section, we present our definitions of emotion, personality and mood and describe our approach in managing the interaction between these different dimensions.

This model is based on the generic model for personality and emotion developed in MIRALab (Egges et al., 2004) but differs by its implementation process of computing and updating emotions and the way we manage the interaction between emotion, personality and mood.

3.1 Emotion

Emotions are one of the most prominent characteristics of interpersonal interactions and they play an essential role in the comprehension of social phenomena.

The emotional state of a character can change over time. It is represented by a set of emotions whose intensities are changing continually.

We define E_t as the emotion state at time *t*. E_t is an *m*-dimensional vector where all *m* emotion intensities are represented by a value in the interval [0,1].

$$E_t = \begin{pmatrix} e_1 \\ e_2 \\ \cdots \\ e_m \end{pmatrix} \forall i \in [1,m] : e_i \in [0,1]$$

We adopt the emotional categories proposed by the OCC Model, but regroup the emotions into four basic emotional classes: SATISFACTION, DISAPPOINTMENT, ANGER and FEAR. These classes were chosen because they reflect the most frequently occurring emotions in the real-life fire-fighting training environment that we are trying to simulate in virtual reality. A second and important consideration is the reduction in computation that is achieved by using these four basic emotional classes.

In our model, E_t is an *4* dimensional vector:

$$E_t = \begin{pmatrix} e_{satisfaction} \\ e_{disappointment} \\ e_{anger} \\ e_{fear} \end{pmatrix}$$

As an example, a character that is experiencing anger and little fear is represented as:

$$E_t = \begin{pmatrix} 0 \\ 0 \\ 0.8 \\ 0.3 \end{pmatrix}$$

These emotions affect the avatar state in three main ways:

1. An affective reaction expressed by a facial animation which conveys the predominant emotion (e.g. fear).

2. A behavioural reaction expressed by a set of gestures or/and actions (e.g. protecting the head with the hands).

3. A cognitive reaction expressed by a change in the perceptive abilities of the character (e.g. narrowing of the visual perception focused on a salient object).

3.2 Personality

Personality is an important research theme within the psychology domain, however, there is currently no general agreement on the definition of personality. In general, most descriptions of personality emphasize the distinctiveness quality of individuals and point out how this quality corresponds to stable psychological traits that are reflected in the behaviour and emotional responses of a person.

We represent a personality as a static *n*-dimensional vector where *n* represent the dimensions of the personality.

$$P = \begin{pmatrix} p_1 \\ p_2 \\ \cdots \\ p_n \end{pmatrix} \forall i \in [1,n] : p_i \in [0,1]$$

p_i represents the dimension of the personality and its value is in the interval [0,1].

One of the most widely accepted models of personality is the Five Factor Model (FFM) (McCrae and John, 1992) in which openness, conscientiousness, extraversion, agreeableness, and neuroticism are considered to be the five dimensions providing a description of personality.

Personality is represented in our model by a vector of five dimensions whose values are in the interval [0,1].

$$P = \begin{pmatrix} p_{openess} \\ p_{conscientiousness} \\ p_{extravert} \\ p_{agreeableness} \\ p_{neurotic} \end{pmatrix}$$

As an example, we represent a character that is very conscientious and extravert but not very open, quite agreeable and neurotic:

$$P = \begin{pmatrix} 0.2 \\ 0.8 \\ 0.7 \\ 0.5 \\ 0.5 \end{pmatrix}$$

3.3 Relationship between Personality and Emotion

The relationship between personality and emotion remains problematic and no unified model which could be directly implemented seems to exist (André et al., 1999).

Many approaches have therefore been developed, such as (Bates, 1994) who maps emotion to behaviours in a personality specific way, (Allen, 2000) who treats personality as a variable that determines the intensity of a certain emotion and (Egges et al., 2004) who define a relationship between every personality factor of the OCEAN model and the goals, standards and attitudes in the OCC model.

In our emotional model, we consider that every emotion is influenced by one or many dimension of the personality. The intensity of this influence is represented in the MPE (m x n) matrix by a value in the interval [0,1] as following:

$$MPE = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mn} \end{pmatrix}$$

$$\forall i \in [1,m], j \in [1,n] : \alpha_{ij} \in [0,1]$$

The following matrix shows an example of an MPE (4 x 5) matrix.

$$MPE = \begin{pmatrix} 0 & 1 & 0 & 0.3 & 0 \\ 0 & 1 & 0 & 0.3 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 1 \end{pmatrix}$$

In this example, the emotion of satisfaction is influenced by the conscientious and agreeableness dimension of the personality as showed by the first line of the MPE matrix.

$$\begin{pmatrix} 0 & 1 & 0 & 0.3 & 0 \end{pmatrix}$$

By these values we mean that an emotion of satisfaction is strongly influenced by the conscientious dimension of a personality (value = 1) and lightly (value = 0.3) by the agreeableness one.

Whether or not the chosen values for the influence of personality dimension on an emotion are justified from a psychological perspective is not in the scope of this paper. However, we provide a possibility of having multiple personality dimensions influences on an emotion.

From this MPE matrix, we consider now the personality vector to define an *m*-dimensional vector S who represents the sensitivity of a character to emotions.

Every sensitivity to an emotion is considered as the threshold of the emergence of the emotion and represented by a value in the interval [0,1].

$$S = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \cdots \\ \theta_m \end{pmatrix} \forall i \in [1,m] : \theta_i \in [0,1]$$

$$With\ \ \theta_i = \frac{\sum_{j=1}^{n} \alpha_{ij}.p_j}{\sum_{j=1}^{n} \alpha_{ij}},\ \ \forall j \in [1,n]\ \ and\ \ \forall i \in [1,m].$$

As explained, one or many dimensions of a personality influence every emotion. Having a character with personality P considered as non-conscientious and quite agreeable, we define θ_{sat} as its sensitivity to the emotion of satisfaction.

$$P = \begin{pmatrix} 0.5 \\ 0.2 \\ 0.7 \\ 0.4 \\ 0.8 \end{pmatrix}$$

The line corresponding to the satisfaction emotion in the MPE matrix shows the following values:

$$\begin{pmatrix} 0 & 1 & 0 & 0.3 & 0 \end{pmatrix}$$

The sensitivity of this character to the emotion of satisfaction will be calculated as:

$$\theta_{sat} = \frac{(0.2 * 1) + (0.4 * 0.3)}{1 + 0.3}$$

This sensitivity represents the threshold of the emergence of the emotion of satisfaction. In other terms, it represents the minimum value that the intensity of satisfaction should have to be considered.

3.4 Mood

Another characteristic of interpersonal interactions is mood which acts as a filter for the emergence of emotions and influences the interpretation of a situation. A person in a 'good' mood tends to have a positive interpretation of the situation which moderates the emotion he or she feels. Conversely, a person in a 'bad' mood has a negative interpretation of the situation, accentuating the felt emotion.

No generally accepted model of mood was found in the psychology literature. Therefore, we define the mood as an *k*-dimensional vector, where all *k* dimension of the mood are represented by a value in the interval [-1,1].

$$H_t = \begin{pmatrix} h_1 \\ h_2 \\ \cdots \\ h_k \end{pmatrix} \forall i \in [1,k] : h_i \in [-1,1]$$

In our model, we use the mood as a *1*-dimensional vector whose value are in the interval [-1,1].

$$H_t = h_t$$

Mood is affected by 2 factors: the emotion felt and the contagion induced by the perception of other characters emotional state.

In order to model the influence of the emotional state on

mood, we define a MEH (k x m) matrix which associates a weight of every emotion with every dimension of the mood.

We assume that every emotion affects differently the mood and the intensity of the mood is depending on the intensity of the emotion felt.

$$MEH = \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ \beta_{k1} & \beta_{k2} & \cdots & \beta_{km} \end{pmatrix}$$

$$\forall i \in [1,k]; j \in [1,m] : \beta_{ij} \in [-1,1].$$

β_{ij} represent the value of the *i*-dimension of the mood when a maximum intensity of the emotion *j* is felt.

The following matrix shows an example on an MEH (1 x 4) matrix.

$$MEH = \begin{pmatrix} 0.8 & -0.5 & -1 & -0.6 \end{pmatrix}$$

These values represent the weight of our model's emotions on the mood. When a maximum intensity of emotion of 'satisfaction'is felt, the value of the mood will be set to 0.8.
Otherwise, the value of the mood will be set proportionally to the weight of the emotion specified on the MEH matrix.

3.5 Overview of the emotional model architecture

The architecture of the emotional model, which is shown in Figure 1, is composed of 4 modules: the "PERCEPTION" module which detects the events perceived from the virtual reality environment, the "EMOTION" module which evaluates the importance of the perceived event depending on the personality and the mood, and decides on the emotional state felt. The output of these two modules is passed to the "BEHAVIOR" module that will select the adequate emotive counterpart for each of the three emotional reactions: affective reaction (facial expression), behavioural reaction (emotional gesture) and cognitive reaction (change in cognitive parameters). In the last step, the "ACTION" module executes the actions decided by the "BEHAVIOR" module.

In this paper, we will focus on how the EMOTION module evaluates emotion.

3.6 Process of computing emotional state

The felt emotions are caused by a perception of cognitive context (some events perceived as well as the validation of certain rules). However, the cognitive context does not have the necessary qualities for the emergence of emotion.

To simulate this mechanism of the emergence of emotion, the module "Emergence Emotions" starts by determining the emotional impact on the virtual actor (I_{t+1})

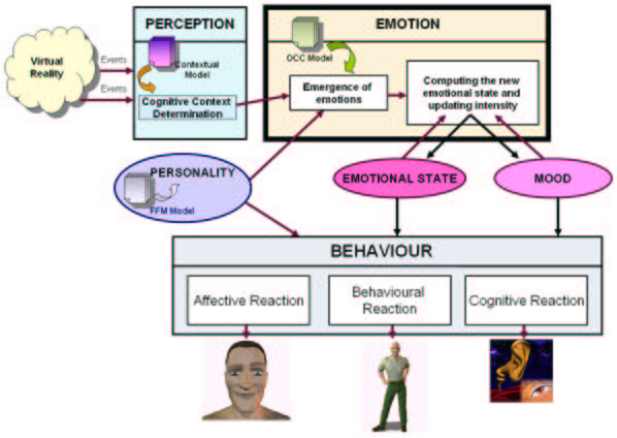


Figure 1: Architecture of the emotional model

based on the information received by the "perception module" and the "contextual model".

The "contextual model" represents a dynamic list of the contexts which could occur in a simulation and which are stored in an external data base. Every context is characterised by its desirability and its importance. This contextual model is managed and updated over time by the "PERCEPTION" module.

The emotional impact (I_t) is an *m*-dimensional vector which represent the value of the potential emotion intensity caused by the perceived event.

$$I_t = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \cdots \\ \lambda_m \end{pmatrix} \forall i \in [1,m] : \lambda_i \in [0,1] \cup \{-1\}.$$

When an event had no emotional impact on the character, the value of λ_i will be set to -1. Otherwise the intensity of the potential emotion will be in the interval [0,1] depending on the importance and desirability of the perceived event.

For example, the following vector shows the emotional impact of an event who is defined in the contextual model that may generate an emotion of FEAR on the character.

$$I_{t+1} = \begin{pmatrix} -1 \\ -1 \\ -1 \\ 0.8 \end{pmatrix}$$

This event had no emotional impact on the emotions of SATISFACTION, DISAPPOINTMENT and ANGER (value = -1) but an important impact on the emotion of FEAR (value = 0.8).

Every perceived event (I_{t+1}) is compared with the sensitivity (S) of the character to decide if an emotion may emerge or may not, in order to evaluate the future potential emotional state (we note EP_{t+1}).

The sensitivity (S) of the character represents the

threshold of the emergence of emotions (i.e. if an emotion is strong enough to emerge).

Once the condition for the emergence of the potential emotion is checked ($\lambda_i > \theta_i$), the module calculates the intensity of the potential emotion (ep_i) and uses this value (EP_{t+1}) to update the emotional state and the mood of the character.

If the event had no emotional impact on a dimension of the emotional state, the character continues to express its emotion.

The potential emotional state is defined as an *m*-dimensional vector, where all *m* dimension are represented by a value in the interval [0,1].

$$EP_t = \begin{pmatrix} ep_1 \\ ep_2 \\ \dots \\ ep_m \end{pmatrix} \forall i \in [1,m] : ep_i \in [0,1]$$

If ($\lambda_i < \theta_i$) then

(No impact on the emotion or the intensity of the emotion felt is lower than the threshold)

$ep_i = e_i - \epsilon$

(The character continues to express its emotion with some decay)

else

$ep_i = \frac{\lambda_i - \theta_i}{1 - \theta_i}$

(We calculate the intensity of the potential emotion by considering the importance and the desirability of the event and the sensitivity of the character to the emotion)

For example, a character with an emotional state E_t and a sensitivity S perceives an event with an emotional impact I_{t+1}.

$$E_t = \begin{pmatrix} 0 \\ 0 \\ 0.5 \\ 0 \end{pmatrix}; S = \begin{pmatrix} 0.2 \\ 0.5 \\ 0.3 \\ \mathbf{0.4} \end{pmatrix} and \ I_{t+1} = \begin{pmatrix} -1 \\ -1 \\ -1 \\ \mathbf{0.8} \end{pmatrix}$$

The emotional impact on FEAR (λ_{fear}) is checked with the sensitivity of the character to this emotion (θ_{fear}): $0.8 > 0.4$.

The emotion of FEAR can emerge and the potential emotion of FEAR will be:

$$ep_{fear} = \frac{0.8 - 0.4}{1 - 0.4}$$

The potential emotional state is calculated as:

$$EP_{t+1} = \begin{pmatrix} 0 \\ 0 \\ 0.5 - \epsilon \\ \frac{0.8-0.4}{1-0.4} \end{pmatrix}$$

Until now, we have calculated a potential emotional state considering only the importance and desirability of the perceived event and the sensitivity of the character according to its personality.

Computation of the new emotional state is undertaken by considering this potential emotional state and by moderating it considering the mood and the emotional state.

$$E_{t+1} = EP_{t+1} + \sigma(H_t, E_t, EP_{t+1})$$

With σ = moderation factor.

Mood can be seen to serve as a background affective filter through which both internal and external events are appraised.

The moderation factor depends on the actual mood (H_t), the actual emotional state (E_t) and the potential emotional state (EP_{t+1}).

In our emotional model, we consider the mood as a 1-dimensional vector whose value is in the interval [-1,1]. A character with a negative mood is considered in bad mood whenever a character with a positive mood is considered in a good mood.

A character in a good mood will tend to accentuate positive emotions and to moderate the negative ones. Conversely, a character in a bad mood will accentuate negative emotion and tend to moderate the positive ones.

The table 1 shows the different sign of the moderation factor (τ) according to mood and the emotion.

The moderation value of the intensity of the potential emotion is proportional to the value of the mood. The more the character is in a good mood, the more it can accentuate positive emotions and moderate the negative ones.

Sign of the moderation factor (τ=)	Bad Mood	Good Mood
Positive Emotion *(e.g. Satisfaction)*	-1	+1
Negative Emotion *(e.g. Disappointment, Fear, Anger)*	+1	-1

Table 1: Sign of the moderation factor

The new emotional state is given by the following formula:

$$E_{t+1} = EP_{t+1} + \tau . |h_t . (EP_{t+1} - E_t)|$$

With:

τ = the sign of the moderation factor : $\tau \in \{-1,1\}$

E_{t+1} = Emotional state at *t*.

EP_{t+1} = Potential emotional state at *t+1*.

h_t = Bad-Good dimension of the mood at *t*.

For example, a character with an emotional state E_t and mood h_t perceives an event that is evaluated to make it in a potential emotional state EP_{t+1}.

$$E_t = \begin{pmatrix} 0 \\ 0 \\ 0.8 \\ \mathbf{0.2} \end{pmatrix};\ h_t = -0.5\ \ and\ \ EP_{t+1} = \begin{pmatrix} 0 \\ 0 \\ 0.8 \\ \mathbf{0.7} \end{pmatrix}$$

The change on the potential emotion affects only the intensity of the 'FEAR' emotion that is evaluated in this example to pass from 0.2 to 0.7.
The value of the mood indicates that the character is in a bad mood.

According to table 1, the sign of the moderation factor is positive. Considering its bad mood, the character will accentuate the negative emotion felt.

The new emotional state is calculated:

$$E_{t+1} = \begin{pmatrix} 0 \\ 0 \\ 0.8 \\ 0.7 + 1 * |-0.5 * (0.7 - 0.2)| \end{pmatrix}$$

Mood influences the intensity of the emotion. It is then important to update the mood every time an emotion has occurred.

We suppose that every emotion affects differently the mood. The value of this influence is represented in the MEH matrix. From this matrix and the new emotional state (E_{t+1}), we calculate the new mood by the following formula:

$$h_t = \frac{\sum_{i=1}^{m} e_i.\beta_i}{\sum_{i=1}^{m} \frac{e_i}{e_i}}\ \ \forall e_i \neq 0\ \ and\ \ \forall i \in [1, m].$$

For example, we suppose we have a character with an emotional state E_t and with a MEH matrix which values are done by:

$$E_t = \begin{pmatrix} 0 \\ 0 \\ \mathbf{0.5} \\ \mathbf{0.8} \end{pmatrix};\ MEH = \begin{pmatrix} 0.8 & -0.5 & \mathbf{-1} & \mathbf{-0.6} \end{pmatrix}$$

The impact of each emotion felt on the mood will be calculated proportionally to its intensity according to its influence on mood as described in the MEH matrix.
We define $h_{emotion}$ as the impact of the emotion felt on the mood:
$h_{emotion}$ = intensity of the emotion * influence of the emotion on mood.

The emotion of ANGER will tend to set the mood to a value: h_{anger} = 0.5 * -1 = -0.5

The emotion of FEAR will tend to set the mood to a value: h_{fear} = 0.8 * -0.6 = -0.48

The final value of the mood is calculated as the middle value of h_{anger} and h_{fear}.

$$h_t = \frac{h_{anger} + h_{fear}}{2}$$

The emotional model takes into account the fact that the intensity of emotions decreases over time. The decay of emotion differs according to the type of emotion and the personality of the character (i.e. an individual tends to forget the positive emotions more quickly than the negative ones). As a result, in the absence of any emotional stimulus the emotional arousal will converge towards zero.

We define the emotional decay vector (D) as a static *m*-dimensional vector who represent the factor of decay of the emotion intensity over time.

$$D = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \cdots \\ \delta_m \end{pmatrix} \forall i \in [1, m] : \delta_i \in [0, 1]$$

In our model, we use the following vector (D):

$$D = \begin{pmatrix} 0.2 \\ 0.4 \\ 0.6 \\ 0.8 \end{pmatrix}$$

The intensity of the emotion felt will be maintained during a time proportional to the factor of decay ($\mu_1.\delta_i$) and begin to converge towards zero with a progression proportional to the factor of decay ($\mu_2.\delta_i$).

The figure 2 shows the variation of the intensity of the emotion of SATISFACTION and FEAR over time.

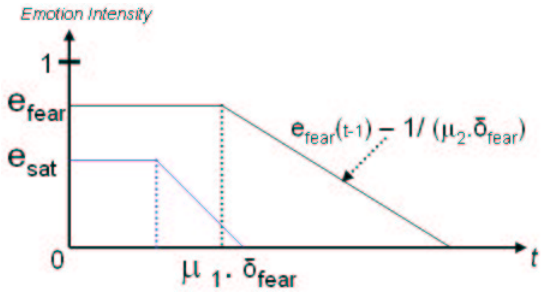


Figure 2: Process of decay of emotions

Finally, the figure 3 summarises the steps in computing the emotional state.

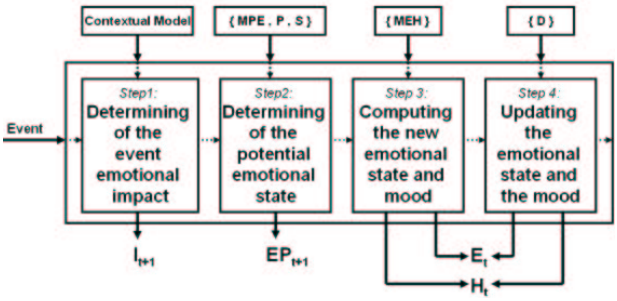


Figure 3: Process of computing emotional state

4 Emotional model and behavioural engine

In order for the character to exhibit comprehensive or at least credible non verbal and emotional behaviour, we have first to compute the most probable emotional state and then to dynamically produce gestures and facial expressions.

For this goal we have adopted a classical strategy which compares the real situation of the character with the expected one in terms of goals, expectations, possible issues, etc (Figure 4).

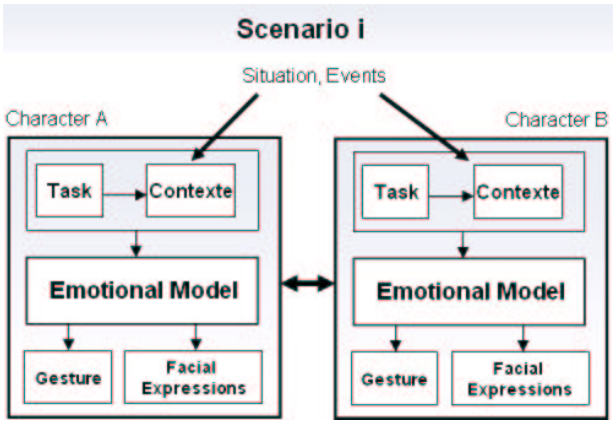


Figure 4: General architecture for the integration of emotion and animation

Depending on the discrepancy between goals and expectations; possible issues and resources and personality of the character, our model will determine the most expected emotion as well as the most suitable gestures or facial expression.

In doing this computation, our approach differs strongly from the traditional AI one (e.g. (Gratch and Marsella, 2001)) because instead of assessing all cognitive, emotional states for our character we start from a situation where it is the human itself that positions its character, animates it, etc. Thus, we compute the emotional state from a situated action (an action that results from human decisions based on all the user cognitive, social and cultural background).

Our model is based on a precise description of all possible contexts the user can be doing his (her) scenario. Depending on this context, his (her) actions, resources, we will generate the most probable emotion which will loop through the interaction with other characters or environment properties.

Example 1 : *How the walking pace can be modified depending of the environmental characteristics and how this body change can be mandatory for a good social regulation?*

This is one of the simplest examples we introduced in order to make the walking pace more credible for firemen in virtual situated actions (Figure 5).

Figure 5: Depending of the context, the same character exhibits different walking pace

In this case we modify the walking pace depending on the fire proximity. The closer the character is to a fire, the more we modify the walking pace. In case of very close proximity, even the position of the arms is modified. The amplitude of these modifications depends also on the personality of the character or his (her) operational status (fireman, civilian, etc.).

It is important to notice the possible cascade effects following this modification. Others characters can visually interpret this pace modification from their position in the virtual space. In case of professional people, they can also take new decisions depending on what they have observed. Again, it is a key point to notice that this decision processes are not IA based but a consequence of a situated decision where all the professional experience can play a role.

Example 2 : *Modeling cognitive, emotional and perceptive mechanisms*

In this example, we will show how we tried to model the well known "visual peripheral field restriction" when subjects are exposed to any kind of stress. It is established that peripheral vision is restricted when subjects are under stress. Unfortunately, our subjects are not fully immersed in their visual environment so they cannot experiment peripheral visual field effects. We tried to implement this mechanism from the output of our emotional model. In situations where the emotional model results in high level of stress, we artificially reduce the screen field of view thus implementing the emotional - cognitive effect (Figure 6).

5 Validation procedure

Our intent is not to validate the emotional model by itself (so many arbitrary parameters and difficulty meaningfulness of the parameter validation) but to assess if its output can generate not only credible behaviour but more

Figure 6: Restriction of the peripherical field of view in stressful situation

interestingly meaningful social interaction through verbal and non verbal interactions.

In order to design such a protocol, we intend to compare the same scenario in a real situation and a virtual one. In both cases we are working with professional agents (firemen). As we are dealing with a complex system (Pavard and Dugdale, 2002) we cannot expect to observe exactly the same behaviour but our intent is to conduct a cognitive post hoc analysis (from video tapes and virtual monitoring) in order to identify which environmental or social cues subjects have used in order to take their decisions. If they use the same class of cues (other subjects postures, emotional interpretation, deictics, etc..), we may conclude that interaction in the virtual space is potentially able to reproduce situated social and cultural interactions.

6 Conclusion

In this paper, we have adopted a different approach in modelling emotion, mood and personality. We adopt the now generally accepted OCC model of emotion (Ortony et al., 1988) and the Five Factor Model (FFM) of personality (McCrae and John, 1992).

Furthermore, in our emotional model we consider personality as a variable that influences the threshold of the emergence of emotions and the mood as an affective filter for moderating the intensity of the emotion felt.

The model so far developed is currently being integrated into a virtual reality training tool for fireman. The integration of this model enables a deeper feeling of immersion by involving the user in a emotionally rich environment that attempts to reproduce the emotional stress felt in a real emergency fire incident.

Our future work is along two axis: firstly, the development of a body behavioural model to include non-verbal communication skills such as gesture in our characters; secondly, to validate our model by adopting an ethnomethodological approach (Dugdale et al., 2003).

References

Steve Allen. A concern-centric society-of-mind approach to mind design. *Proceedings of the AISB'00 Symposium on How To Design A Functioning Mind., Birmingham, England*, 2000.

Elisabeth André, Martin Klesen, Patrick Gebhard, Steve Allen, and Thomas Rist. Integrating models of personality and emotions into lifelike characters. *A. Paiva (Ed.) Affect in Interactions Towards a New Generation of Interfaces.*, 1999.

Gene Ball and Jack Breese. *Emotion and personality in a conversational character*. In: Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (eds.): Embodied Conversational Agents, Cambridge, MA:MIT Press, 2000.

Joseph Bates. The role of emotions in believable agents. *Communications of the ACM*, 37(7):122–125, 1994.

Cynthia Breazeal. Toward sociable robots. *Robotics and Autonomous Systems*, 42(3), 2003.

Julie Dugdale, Nico Pallamin, Bernard Pavard, Marcus Sanchez-Svensson, and Christian Heath. A modelling framework for improving the interactivity in virtual worlds. Technical report, GRIC-IRIT, Toulouse and WIT, King's college, London., 2003. URL `http://www.irit.fr/COSI`.

Arjan Egges, Sumedha Kshirsagar, and Nadia Magnenat-Thalmann. Generic personality and emotion simulation for conversational agents. *Computer Animation and Virtual Worlds.*, 2004.

Jonathan Gratch. Emile: marshalling passions in training and education. *Proceedings of the Fourth International Conference on Intelligent Agents, Barcelona, Spain.*, 2000.

Jonathan Gratch and Stacy Marsella. Tears and fears: Modeling emotions and emotional behaviors in synthetic agents. *Proceedings of the 5th International Conference on Autonomous Agents, Montreal, Canada.*, 2001.

Sumedha Kshirsagar and Nadia Magnenat-Thalmann. A multilayer personality model. *Proceedings of 2nd International Symposium on Smart Graphics*, (1):107–115, 2002.

Stacy Marsella and Jonathan Gratch. Modeling the interplay of emotions and plans in multi-agent simulations. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society. Edinburgh,Scotland.*, 2001.

Stacy Marsella and Jonathan Gratch. A step towards irrationality: using emotions to change belief. *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-Agent Systems. Bologna, Italy.*, 2002.

Robert R. McCrae and Oliver P. John. An introduction to the five-factor model and its applications. *Special Issue: The five-factor model: Issues and applications. Journal of Personality:60*, pages 175–215, 1992.

Andrew Ortony, Gerald L. Clore, and Allan Collins. *The Cognitive Structure of Emotions*. Cambridge University Press., 1988.

Bernard Pavard and Julie Dugdale. From representational intelligence to contextual intelligence in the simulation of complex social system. *CASOS conference. Pittsburg.*, 2002.

Artistically Based Computer Generation of Expressive Motion

Michael Neff* Eugene Fiume*

*Department of Computer Science
University of Toronto
neff|elf@dgp.toronto.edu

Abstract

Understanding how to create the right movement for a specific character is a significant open problem in computer animation. This same problem, however, has been at the heart of the work of performance artists for hundreds of years. In this work, we try to learn from the lessons contained in the performing arts literature so that we can apply them to creating rich, engaging, animated characters. Three classes of movement properties are identified: those relating to shape, those relating to transitions and those relating to timing. Computational models of some of these properties have been developed and are briefly detailed. A software framework is also presented that allows these properties to be applied independently to a character's movement. The resulting animation system allows the key aesthetic aspects of movement to be quickly varied and allows motion to be easily customized for a particular character.

1 Introduction

A talented actor, dancer or animator can take a few abstract notions of a character, and from those build a rich, consistent, concrete piece of movement that captures the essence of that character. This ability to take an abstract idea and give it concrete form is the heart of a performer's art. In the research field of computer animation, we have barely begun to understand how this mapping works: how an often vague description of a few character properties can be mapped to a complete, concrete style of movement. We can, however, gain insight into the process by studying the arts literature. This can help us to understand what aspects of movement have the most significant impact on the expressive content of that movement and what aspects are most useful to vary in order to develop various types of characters. In a sense, this is like building a vocabulary of fundamental movement parameters that can then be used at any time to define the movements of a particular character.

Automatically generated motion has generally lacked the expressiveness required to create engaging characters. For this reason, character systems often rely on an animator to provide either key poses or animation sequences as the basis for their animation engines. For example, in the AlphaWolf system of Tomlinson et al. (2002), an interactive wolf simulation demonstrated at SIGGRAPH 2001 by the Synthetic Characters Group from Media Lab, an animator crafted animation sequences at the extremes of a dominance scale and the system then interpolated between them. In approaches such as this, the manner in which a motion is modified remains a part of the animator's craft and is not under control of the character modeling system. A long term goal of our work is to provide the tools and vocabulary necessary for the character system to directly control how a motion is modified in order to capture the nuances of a specific character's movements.

Our work learns and applies lessons from the arts literature. Through studying the performance literature from theatre, animation, mime, theatre anthropology and movement theory, we have built a list of key movement properties that significantly affect the aesthetic aspects of a character's movement. These have been divided into three main categories: those that affect the character's shape or pose; those that affect transitions, namely how a character moves from pose to pose and other transient effects; and those that relate to timing. We have built computational models for many of these movement properties and developed a software framework that allows this extensible set of properties to be combined in order to generate animated sequences.

The focus of this work is on how to vary the manner in which a given movement is performed, rather than on what movement is performed. In other words, we are interested in defining the expressive range over which an action can be performed, and in finding generic ways to parameterize these ranges.

One of the major desired outcomes of this research has been the creation of better authoring tools for character animators. We hope this work can also play an important role in research on autonomous or semi-autonomous expressive characters. In these fields, it is important to be able to customize the movement style of various characters in order to match their mood and personalities. Our research not only elucidates the movement properties that are likely to be useful to vary, but also provides computational models for many of these properties, and a framework in which these properties can be combined. A long term goal of our research is to examine how these movement properties can be combined in various ways to cre-

ate character sketches. Such sketches make use of the low-level movement properties to define a basic movement style for a character. Different requested actions will then be performed in a manner that is consistent with the character sketch.

2 Lessons from Computer Animation

Previous work in computer animation on expressive character motion can be divided into three categories: script based systems, model-based systems and data driven systems. These categories are distinguished by where the knowledge of what makes motion expressive lies. In script based systems, the user hand tunes expressive variations in motion using low-level constructs. The system does not encapsulate any knowledge of expressive movement. Model based methods try to capture some aspect of what makes motion expressive within a computational model, which can then be invoked by a user. The user is thus working at a higher or more abstract level. Data driven methods make use of motion capture data and the expressiveness is in the captured sequences. A user merely determines which data is used to create the final motion. Several model based approaches also use motion capture data, but the intent there is to build a model of a movement quality that will then be independent of the data.

A script based system was introduced by Perlin (1995) that allowed a user to interactively control an animated character. The actions of the character were specified ahead of time using script files that specified character poses and transition functions that could be used to move between these poses. The transition functions were either sinusoids or different frequency noise functions. Perlin and Goldberg (1996) extended this work to a general system for scripting interactive characters known as *Improv*. This system took a similar approach to generating animation, but also allowed behavioural scripts to be defined that determined what actions a character took.

Chi et al. (2000) use Laban's Effort-Shape movement analysis to define a fixed set of parameters that can be used to modify the style of a motion. This model-based approach is similar to our work in that their model is based on research in the arts literature, but we aim at a more open, extensible system and we target a larger set of movement properties.

Other model based approaches use captured data of human movement in order to develop a computational model. Work on expressive transforms (Unuma et al. (1995); Amaya et al. (1996)) has attempted to extract emotional content from a piece of captured motion. The extracted transforms can then be applied to other motions, in order to give them the same emotional 'feel'. Brand and Hertzmann (2000) provide for very high level editing of a motion's style. They learn a style, such as ballet or modern dance, from captured motion and can then apply this style to other movement sequences. Pullen and Bregler (2002) work at a similar high level, allowing an animator to specify key frames and then use a statistical model drawn from motion capture data to texture the key framed motion with the style of the captured data.

Rose et al. (1998) present a data driven system in which expressive motion is generated by interpolating within a large captured motion set. Actions, such as walking, are called verbs in their system. They capture the motion of each action of interest being performed in a large number of ways. These variations then define an "adverb" space. During playback, the user can specify an action and a location within the adverb space. The final motion is generated by interpolating between the captured motions. By modifying the location in adverb space, the user can change the style of the motion.

Bruderlin and Williams (1995) adjust captured motion by treating movement as a signal and adjusting the gain of various frequency bands of the signal. They suggest different bands capture different aesthetic qualities of the motion.

Our work is a hybrid between script based and model based systems. We build computational models of movement properties, but these are defined programmatically, potentially by the end user, and are highly configurable. Users can also build new properties which are composites of existing properties defined in the system. Unlike many previous approaches, the expressive properties represented in our system are all garnered from the arts literature and as such benefit from traditional practice. As well, a focus of our work is on how a varied and extensible set of movement properties can be combined. An additional advantage of our work is that we generally solve for expressive constraints at the same time as solving for hard constraints, such as having a character touch a certain location. Many of the above approaches operate by warping an existing piece of motion and can hence violate hard constraints in generating the final motion.

In previous work, we have presented a model for tension and relaxation in the human body and explored its expressive impact (Neff and Fiume (2002)). We have also presented a small set of simple yet powerful aesthetic edits that can be used to modify the feel of a motion sequence (Neff and Fiume (2003)).

3 Lessons from the Arts

One of the lessons taken from the arts literature is that performance movement is not the same as daily life movement. Eisenstein and Tretyakov (1996) argue that the purpose of stage movement is to infect the audience with emotion and according to Alberts (1997), it is generally when the attributes of an actor's movement are out of the ordinary that they will have the greatest significance for the audience.

Because of the need to communicate clearly with an audience, performance movement is based on two basic principles: *simplification* (Lawson (1957); Thomas and Johnston (1981); Lasseter (1987); Barba (1991b)) and *exaggeration* (Thomas and Johnston (1981); Lasseter (1987); Barba (1991a)). These two properties work together to clarify the meaning of a character's movement in the spectator's mind.

Simplification, also known as the "Virtue of Omission" (Barba (1991b)), works to bring focus to certain elements of a character's movement by eliminating extraneous movements. In mime, it is often necessary to slow down the pace of major actions to make them comprehensible (Lawson (1957)). In traditional animation, they preach the importance of having a character only do one thing at a time (Thomas and Johnston (1981); Lasseter (1987)). All this lends clarity to performance motion that is often lacking in the movements of daily life.

Once a movement has been simplified, it is exaggerated to ensure that its meaning is conveyed to the audience. Frank Thomas nicely summarizes the interplay of simplification and exaggeration as follows: "As artists, we need to find the essence of the emotion and the individual who is experiencing it. When these subtle differences have been found, we must emphasize them, build them up and at the same time, eliminate everything else that might appear contradictory or confusing." (Thomas, 1987, p.6). By so doing, an artist ensures that he/she is focusing the audience's attention on the key aspects of the movement and clear communication results.

We have organized our analysis of the arts literature into three main categories: shape, transition and timing. We will briefly examine a few of the most important properties from each category.

3.1 Shape

From the point of view of defining a character, one of the most crucial aspects of shape is stance. Stance is a combination of a character's posture and his/her balance point. Stance is one of the quickest indicators of both a character's overall personality and how he/she is feeling in a particular scene. One of the simplest views of posture, espoused by Alberts (1997), is as a measure of the overall level of tension in a character's body, or more specifically, to what degree the character expends energy to hold himself erect against the pull of gravity.

Barba (1991a) illustrate the importance of the curve adopted in the spine. The 'beauty line', seen perhaps most famously in the Venus de Milo and other ancient Greek statues, creates a large S-curve that snakes through a character's body and lends a very sensual appearance to the work. This provides a clear illustration of the role of spine curvature in the coronal plane. Spine curvature in the sagittal plane is most often related to a character's reaction to gravity, as suggested by Alberts. It can also be used to illustrate a character's reaction to an object by having a character either recoil away from an object or lean towards it.

Shawn (1963), in summarizing the work of the movement theorist, Delsarte, suggests that the part of the torso that a person habitually holds forward is a strong indicator of what kind of person they are. If they hold their chest high, this indicates self-respect and pride. If their abdomen is protruding, this indicates animality, sensuality and lack of bodily pride. A normal, balanced carriage will have the middle zone of the abdomen carried forward and the chest and abdomen withdrawn. This triad can be augmented by considering people who carry their head forward, normally indicating a mental or academic disposition.

Balance in stance refers to where a character's centre of mass is relative to his or her feet. Is the character's weight centred between the two feet or does the character have a lean? When we stand in daily life, we are never still, but rather are constantly making small adjustments, shifting our weight to the toes, heels, right side, left side, etc. Barba (1991b) argues that these movements should be modeled and amplified by the performer, underlying the expressive importance of balance adjustments. Indeed, Laban (1988) and Tarver and Bligh (1999) suggest it can be particularly powerful to situate a character near its balance limit. This makes the pose more precarious and acts to heighten the associated sense of tension for the audience.

Extent is another key shape property. Extent or extension refers to how far an action or gesture takes place from a character's body. It can be thought of as how much space a character is using while completing an action. If the arms are fully extended and straight out from the character's side, this would be maximal extension, whereas, if the hands were touching the torso and the elbows were held at the character's side, this would be minimal extension. Both Laban (1988) (also Tarver and Bligh (1999)) and Delsarte (Shawn (1963)) use the term *extension*. Alberts refers to this as *range*.

Laban refers to the area around a person's body as the kinesphere and defines three regions within it (Tarver and Bligh (1999)). The near region is anything within about ten inches of the character's body. This is the area for personal, intimate or perhaps nervous actions. The middle area is about two feet from the person's body and this is where daily activities take place, such as shaking hands. The area of far extent has the person extended to full reach. It is used for dramatic, extreme movements and in general is used more on stage than in daily life. Laban (1988) argues that people learn the space around their body and determine how it is most comfortable to perform an action.

The literature on body shape is rich and warrants considerably more investigation and, perhaps, critical scrutiny, than we are able to provide in this paper. Many writers deal with the meaning associated with different areas around the body. It is also common to associate

meaning with different parts of the body. The sequencing of body shapes, or put differently, how shape varies over time, is also very important expressively.

3.2 Transitions

Transitions deal with how a character moves from shape to shape. They also deal with transient aspects of movement, such as the interplay of tension and relaxation.

Laban (1988) suggests that the flux of movement can flow continuously, be intermittently interrupted, yielding a trembling kind of movement, or stopped, yielding a pose. It is worth distinguishing between a motion that is paused (Laban (1988); Tarver and Bligh (1999)) or suspended (Alberts (1997)), and one that is stopped. According to Tarver and Bligh (1999), when a motion is paused, the actor maintains focus on the final destination and can continue the movement at any time. There is no perceptible loss of intensity nor break in intention. When a motion is stopped, the energy of that motion has been lost and the actor's focus is no longer on the completion of a motion. It cannot be seamlessly continued with the same intensity. This may seem to be an esoteric point, but it is quite easy to distinguish between a paused and stopped motion when observing a performer.

According to Laban, motion can be either complete or incomplete(Tarver and Bligh (1999)). Many actions in daily life are incomplete – they are stopped before they reach their natural conclusion. Being able to stop an action midway can have a powerful effect as it can be a strong indicator of a character's internal mental process. Consider a character that reaches out to comfort a former lover and then stops the motion part way through. This is a clear indication of the internal conflict the character is experiencing.

Lasseter (1987) and Thomas and Johnston (1981) describe how Disney animators found it effective to have the bulk of footage near extreme poses and less footage in between in order to emphasize these poses. They referred to this as slow in, slow out, indicating that most of the time was spent on the main poses. Interpolating splines are used to generate this result in computer animation (Lasseter (1987)), often going beyond simple ease-in, ease-out curves using tool such as tension, continuity and bias splines(Kochanek and Bartels (1984)).

Modifications to the transition envelope of a movement are very important expressively. Some movements, such as a tired character dropping his head, start slowly and accelerate as they proceed. Other movements, like a feinted lunge, start very quickly and slow near the end as the character stops the motion in a controlled way. Other motions, such as those common in tai chi, move at a consistent pace throughout. Further to this, some characters will have a light touch while others have a heavy touch. This is not necessarily a reflection of the physical mass of the character, but is rather controlled by how the character moves his mass. Very heavy characters may still move in a light way, and vice versa.

Arcs are another Disney animation principle (Thomas and Johnston (1981); Lasseter (1987)). The *arcs principle* claims that natural movements are normally not straight, but follow some form of arc.

The interplay of tension and relaxation is another widely cited movement property. See for example: Dorcy (1961); Laban (1988); Shawn (1963); Lawson (1957); Barba (1991a). There is a flow between tension and relaxation. There must first be relaxation in order for there to be tension and tension is followed again by relaxation. Shawn (1963) describes the inflow and outflow of energy that accompanies tension modulation. There may be different tension levels in different areas of a character's body. Furthermore, as Lawson (1957) describes, tension changes can take place through the entire body or a tiny part. They can occur suddenly or gradually and there can be spasmodic changes back and forth. Laban (1988) suggests a rise in tension can serve to accent a movement. Physical or emotional pain can be shown by spasmodic contractions of muscles, followed by relaxation as the pain eases, according to Shawn (1963).

3.3 Timing

The two main components of timing are tempo and rhythm (Lawson (1957); Laban (1988); Alberts (1997)). *Rhythm* refers to the overall beat structure or pattern of a set of movements. For instance, the pattern could be long, long, short, repeat. *Tempo* refers to the speed at which motions are completed. Tempo is independent of rhythm. For instance, a given rhythm could be performed with a fast or slow tempo.

Other terms used to define timing include duration, which is the length of time an action takes (Alberts (1997)) and speed, which is the rate at which we let movements follow one another (Laban (1988)).

The timing or speed at which an object moves reflects the size and weight of the object (Lasseter (1987); Shawn (1963)). Larger objects have more momentum and thus accelerate and move more slowly than small, light objects. In his "Law of Velocity", Delsarte extends this idea to include the relationship between the inner gravity of feeling or meaning and the pace and size of its expression (Taylor (1999); Shawn (1963)). Thus the deepest of feelings may be expressed with complete stillness (Taylor (1999)). "Emotions of profound and deeply serious import, then, require slow and large movement patterns; emotions that are petty, light, trivial, nervous etc. take on small and quick movement patterns" (Shawn, 1963, p.64). Extending the idea to character types suggests that "[m]ajesty and nobility of emotion and character can only be conveyed by broad, slow movements while the petty, small, nervous, fearful, irritated, shallow person and emotion is revealed by quick, small movements." (Shawn, 1963, p.65) This same phenomenon is often at play in film making when superheroes gifted with great speed are

shown not moving more rapidly than those around them, but shown moving in slow motion to lend gravity to their deeds.

Lasseter (1987) suggests it is important to spend the correct amount of time on *anticipation* for an action, the action itself and then the reaction to the action. Timing can be used to indicate if a character is nervous, lethargic, excited or relaxed (Lasseter (1987)).

The use of *successions* is arguably the most important timing property. Successions deal with how a movement passes through the body. Rarely will every limb involved in a motion start and stop at the same time. As described by Shawn (1963), Delsarte defined two types of successions: true or normal successions and reverse succession. In a normal succession, a movement starts at the base of a character's torso and spreads out to the extremities. In a reverse succession, the movement starts at the extremities and moves in toward the centre of the character. Delsarte associated true successions with good, truth and beauty and reverse successions with evil, falsity and insincerity (Shawn (1963)). Shawn (1963), a pioneer in the medium, argues that the active use of successions was fundamental to the development of American modern dance.

The importance of successions was also known by traditional animators. Thomas and Johnson write: "Our most startling observation from films of people in motion was that almost all actions start with the hips; and, ordinarily, there is a drop – as if gravity were being used to get things going. From this move, there is usually a turn or tilt or a wind up, followed by a whiplash type of action as the rest of the body starts to follow through... . Any person starting to move from a still, standing position, whether to start walking or pick something up, always began the move with the hips."(Thomas and Johnston, 1981, p.72) This is consistent with a normal succession.

4 System Overview

Our animation system is focused on the creation of realistic, humanoid character motion for a standing figure. The actions we work with include gestures, posture changes and balance adjustments. Focussing on this limited set of motions allows us to explore the rich expressive range contained therein. The system can create motion either kinematically or using physics-based simulation for a 48 Degree of Freedom (DOF) skeleton. Dynamically simulated motion takes advantage of tension modelling and can produce higher quality results, but due to the computational cost of physical simulation, does not yet run as a real-time process. Kinematic generation is currently more appropriate for interactive applications.

To meet the goals of this work, it was necessary to develop a new software architecture. This architecture must achieve several ends. It must:

- allow motion input from multiple sources.
- allow different aspects of the motion to be varied separately.
- support the procedural definition of movement properties.
- provide a mechanism to smoothly combine all the movement properties that are active in the system.

The unifying concept underlying the architecture is the *movement property*, or *property* for short. Low-level properties vary a particular aspect of a movement, such as the amount of tension in a joint, the shape the torso takes during a reaching movement, or the transition function used to vary the envelope of a movement. Higher level properties may combine several low-level properties to vary several properties of a movement in a consistent way. Both high and low-level properties can be procedurally defined. Properties provide the handles by which an animator modifies a movement. They are used to both define and edit an action. Often a property will be bound to some underlying feature in the system that actually does the work of modifying the character's motion. For instance, the *Balance Shift* property allows an animator to adjust the desired x-z position of the character's centre of mass. This change is effected by the balance controller, which in conjunction with an IK algorithm, modifies the joint angles in the character's lower body to achieve the change. The property itself merely specifies the desired change.

The software architecture is shown in Figure 1. Animators, technical directors, or another computational process can interact with the system through three channels: the script, action descriptions, and direct animator edits. In the future, they will also be able to create character sketches that will automatically customize movement sequences for a given character.

The script contains a time-ordered list of actions. It defines what a character does. There are multiple tracks in the script allowing actions that affect different parts of the body to be superposed. The script can accept predefined actions or reactive controllers. Predefined actions are used for planned movements, such as reaching for an object. Reactive controllers actively change the character's behaviour in response to sensor input. Balance is the most significant reactive controller in the system. It actively modifies the angles in the lower body in order to achieve a desired balance point. This desired balance point can be moved over time to affect balance changes.

An action description is a small script file that defines a specific movement, such as a posture adjustment, a gesture or a reaching motion. The representation used for action descriptions is discussed in detail below. The basic philosophy of action descriptions is to attempt to find a minimal definition of a movement that then facilitates a wide range of expressive editing and customization.

Animator edits allow an animator or character controller to directly modify either an individual action or a series of actions that make up a movement sequence.

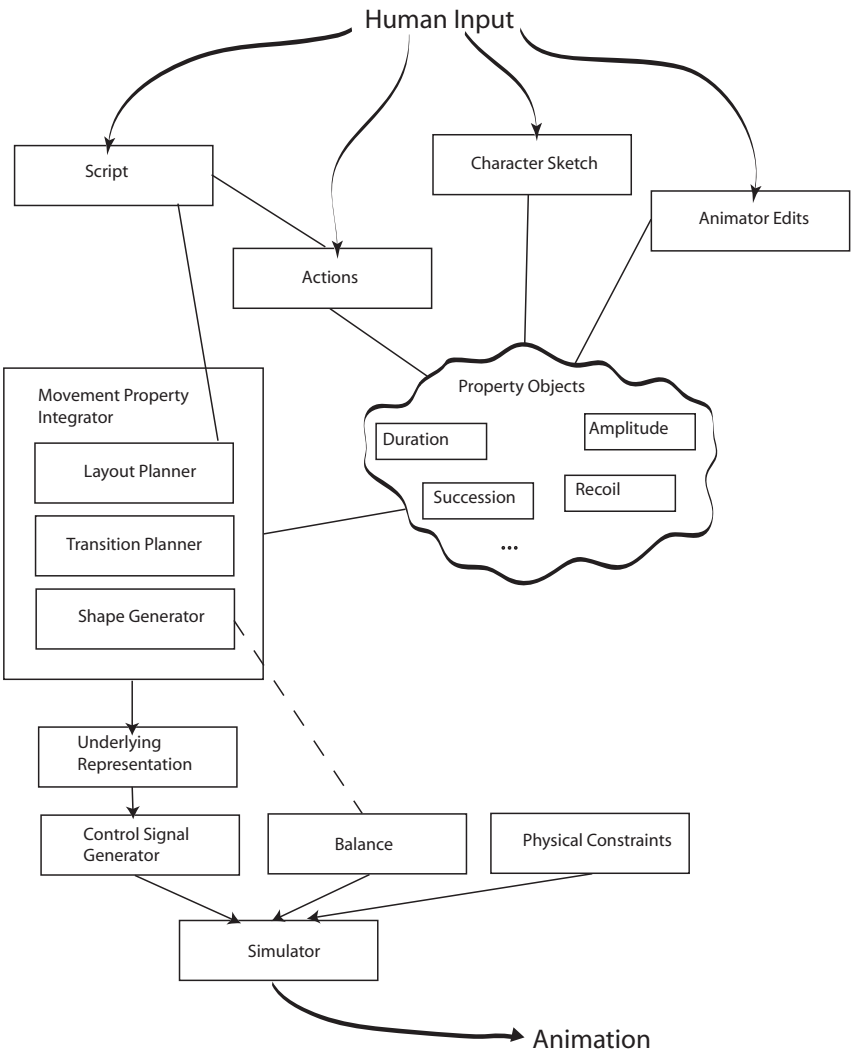


Figure 1: System architecture.

The edits work by either modifying the active movement properties or invoking additional movement properties.

The *character sketch* is used to specify a default set of tendencies for a given character. This functionality is currently being researched and will be added to future version of the system.

As much as possible, properties are defined to be orthogonal. For instance, modifying the shape of a character's torso is independent of the amount of tension in the torso and the timing used to move to this new shape. This orthogonality limits conflicts when various aspects of a character's movement are varied independently. Nonetheless, conflicts may occur. This can happen if two input sources, such as an animator edit and the character sketch, modify the same underlying property, or if composite properties modify a low-level property that is either directly modified or modified by a different composite property. It is the job of the *Movement Property Integrator* (MPI) to arbitrate between these conflicts in a consistent and predictable manner. Conflicts are currently dealt with by ordering the application of properties based on their source and time of application, but we are actively investigating more general conflict resolution schemes. The MPI is also responsible for mapping the movement properties to the low level system functions that are used to achieve them. Three planners are used to meet these goals, one for shape, one for timing and one for transitions. The MPI combines the movement properties in order to generate the *Underlying Representation* that will be used to drive the animation.

The Underlying Representation (URep) is used to generate the final either kinematic or dynamic animation. An-

imators do not need to be aware of this representation to create effective motion. All the movement properties active in the system are automatically transformed to the URep. The URep consists of a set of time-indexed tracks. There are two types of track. First, there is a track for each degree of freedom in the character's body which specifies the desired value for this DOF and other DOF related data. The second set of tracks contains control parameters that are not related directly to an individual DOF, such as the desired offset to the COM and the desired gaze direction. This data is used by processes that modify the character's movement as it is generated in order to meet these constraints.

Tracks are populated either with *Transition Elements* (TElements) or control parameters that are good for a single simulation time step. The attributes specified by TElements include a desired value of the DOF, the length of time it should take to reach that value, the transition function used to shape the envelope of the movement and desired tension changes. Planned actions are mapped down to a series of TElements by the Movement Property Integrator. Control params are generated at the beginning of each time step by reactive controllers, such as the balance controller. They are used to modify the character's movement in response to changes in the system state.

The *Control Signal Generator* (CSG) drives off the Underlying Representation to calculate whatever data is needed by the active simulator in order to generate the animation. With a kinematic simulator, the CSG simply determines the value of each DOF. With a dynamic simulator, the CSG must determine the two antagonistic joint gains and damping value required to achieve the desired DOF values. The details of the low-level muscle model are discussed in the Transition Properties section below.

The simulator is responsible for generating the final animation using all of the control data provided by the CSG. The Kinematic Simulator merely needs to update the transformation hierarchy based on the values of each DOF. The Dynamic Simulator must apply all forces acting on the character and then integrate the physical equations of motion in order to determine the new position of the character. A commercial package, SD/Fast (Hollars et al. (1994)), is used to generate the code that implements the equations of motion for the specific character.

4.1 Action Representation

The key role of the action representation is to facilitate editing of the motion. Action definitions attempt to capture the essence of a motion while specifying a minimal number of details. For instance, a reaching motion might be defined by a world space target the wrist must touch. A wave might require the arm to be in a particular area and the elbow angle to oscillate. Such minimal descriptions can be flushed out by other properties provided by the character sketch or animator edits. This allows a generic action to be customized for a particular character and situation.

Actions are essentially pose based, normally specifying particular poses at particular instants in time, although edits such as succession can break up this structure. Action descriptions are defined using both high and low-level properties. They include data on timing, some joint angles, external constraints such a world space point to touch and often composite movement properties with their parameters.

4.2 Timing Properties

Timing properties are applied directly to the time properties of the transition elements. Generally, they work to either shift the position of a transition element on its track or to scale it.

Succession is an example of a procedurally defined timing property. It operates by directly shifting the location of the transition elements in the underlying representation. A succession edit takes two parameters: whether the succession is normal or reverse and how much of a time offset (t) to use between the joints involved in the motion. The edit determines all of the transition elements it is being applied to and shifts their starting time based on where they are in the character's joint hierarchy. For instance, a normal succession would not modify the first joint in the spine, it would offset the next joint by t, the following joint by $2t$, and so on. The succession traces down all branches in parallel, for instance, modifying the start time of both collar bones, then both shoulders and then both elbows, etc.

4.3 Transition Properties

Two of the key transition properties in the system are transition functions and tension changes. Transition functions are used to warp the envelope of a motion. They are modelled using a cubic Hermite embedded in time and in space. This allows good control over the shape of the transition, including anticipatory and overshoot effects. A transition curve is associated with each transition element in the underlying representation. Each DOF can have its own transition curve or transition curves can be applied to joints, groups of joints, or an entire movement. Common transitions include: ease-in, ease-out; linear; fast start, slow end; slow start, fast end; backtrack before going forward (anticipation); and exceed final point and then return to it (overshoot).

Tension changes are used when the motion is dynamically simulated. Varying tension has a number of effects on a character's motion. Perhaps most obvious, the amount of tension in a character's body will effect how the character reacts to external forces, such as being shoved. One of the most important effects of tension changes is to warp the envelope of a motion. For instance, if a joint's tension is increased during a motion, the attached limb will move more quickly at the beginning of a motion and

more slowly at the end. Low tension effects include overshoot, where a character moves slightly beyond its desired target before moving back to it, and pendular motion, for instance, when a character brings his arm down to his side. Tension also effects how forces travel through the body and how precisely the transition function is tracked. In general, modifying tension provides an effective way to differentiate between a character that moves in a loose, relaxed style and a character that moves in a more tight and rigid style.

Human muscle is organized in agonist-antagonist pairs. This simply means that muscles act in opposite directions around a joint. The final position of the limb will be the equilibrium point of the forces generated by these muscles and any other forces, such as gravity, acting on the limb. One theory of movement, the *equilibrium point hypothesis* proposed by Feldman (1966), suggests that humans move by directly varying the equilibrium point of their muscles. We developed a simple muscle model that is based on the two ideas of equilibrium point control and antagonistic muscle pairs. It consists of two opposing angular springs and a damper for each DOF.

The control law for the antagonistic actuator is

$$\tau = k_L(\theta_L - \theta) + k_H(\theta_H - \theta) - k_d\dot{\theta}, \qquad (1)$$

where τ is the torque generated, θ is the current angle of the DOF and $\dot{\theta}$ is its current velocity. θ_L and θ_H are the spring set points which serve as endpoints for the motion range of a DOF, k_L and k_H are the spring gains, and k_d is the gain on the damping term. The stiffness or tension of the joint is taken as the sum of the two spring gains, $k_L + k_H$. The desired angle is not explicitly shown in the actuator, but is controlled by the value of the two gains. This can be seen as a reparameterization of Proportional-Derivative (PD) control that is more appropriate for directly modeling tension changes. The gains that will achieve any desired angle are simply a linear function of each other.

In previous work using PD controllers, the gains of a PD controller were normally either carefully hand tuned to achieve a given behaviour or they were kept high in order to minimize positional error. The second case generates stiff characters. The first case requires much painstaking hand-tuning. Neither case leads to characters that can easily modify their tension. We avoid this by using equilibrium point control that takes external forces acting on the limb, mainly gravity, into account and adjusts spring gains to compensate for them. This allows us to achieve any desired angle while also varying the tension. A more detailed account of our work on modeling tension is given in Neff and Fiume (2002).

4.4 Shape Properties

Shape properties such as posture changes act through a shape generation module called the *body shape solver*. This module solves for a pose that satisfies the various shape properties acting on the character. The shape generator accepts a combination of hard world-space constraints, such as touching a particular world space point or maintaining a foot position; soft expressive constraints, such as maintaining a certain curve in the character's spine or keeping the character's elbows out from his side; and balance adjustments.

The body shape solver is customized for the human skeleton. It contains a combination of analytic and optimization based Inverse Kinematics routines. Balance control is based on a feedback strategy that uses the error in the character's centre of mass position to adjust the angles of the dominant ankle.

Other edits, such as the *extent* and *amplitude* edits, work by directly modifying the desired joint angles stored in the TElements of the Underlying Representation. More detail on these can be found in Neff and Fiume (2003).

5 Results

All animations discussed here can be found online at http://www.dgp.toronto.edu/people/neff

The system has been used to generate a wide range of animations involving reaching motions, gestures, posture changes and balance adjustments. To illustrate how various movement properties are layered together, consider a simple example of a character reaching for an object, say an apple, on her right. We want the reaching motion to have a sensual, languid feel. The initial action description specifies the reach target and uses default values for timing and the transition function. The result is a rather plain piece of animation. To create the sensual look desired, a shape property is invoked that causes the body shape solver to seek a beauty-line posture and adjust the balance point. The character's gaze direction is also modified to look toward the apple. A succession edit is applied to increase the sense of flow associated with the motion and give it a languid feel. The duration of the motion is extended and the transition functions are adjusted to make the motion smoother.

Now consider a second character that has a nervous temperament and is afraid to touch the apple, but must do so anyway. Starting from the same basic animation, we apply a different set of properties. First, we invoke a shape property that has the character recoil away from the apple while reaching for it. This time, we use a slight reverse succession to heighten the sense of unease. Timing is shortened and the transition function is varied to have the character jerk away from the object. Starting with the same basic action, we have created two animations with completely different feel. The initial reach solution, a frame from the sensual reach and a frame from the recoiling reach are shown in Figure 2

An animation was created of a very loose, relaxed character giving the direction "go left". Two versions of the animation were generated: a dynamic version that makes

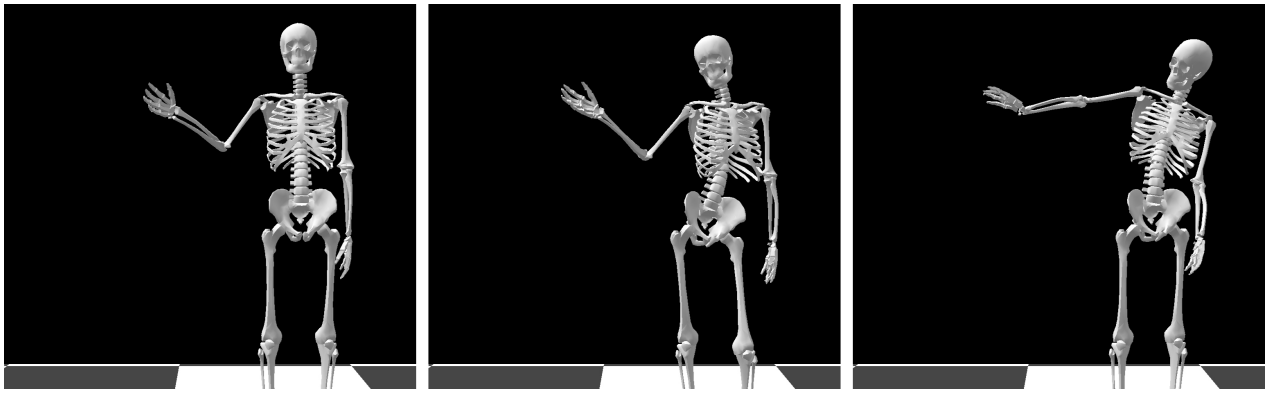


Figure 2: This figure shows a frame from each of three different reaching motions. The first is from the neutral, default motion. The second shows a more sensual character reaching for the object. The third shows a character who would rather avoid the object.

use of tension changes and a kinematic version for comparison purposes. Both animation sequences are based on the same set of actions and use the same poses. The dynamic animation with tension changes does a much better job of capturing the desired sense of looseness and flow. This is due to several effects resulting from tension modeling that only occur in the dynamic version: the loose wrists hang in an appropriate way and show the effect of momentum during movements, the arms have a slight bounce to them as they are brought to the character's side, and arm motions induce some secondary movement in the torso.

Figure 3 shows a 0.4 sec section from the animation. During this clip, the character has just brought his arms down to his sides such that his forearms are parallel to the ground. The top image shows five frames from the kinematic version of the animation. There is no movement during this period. The bottom image shows five frames from the dynamic version of the animation. The bounce in the spine, shoulders, elbows and wrists can be clearly seen. This additional motion greatly adds to both the sense of looseness capture by the animation and the realism of the clip.

The dynamic version of the animation is also smoother. This is because dynamic simulation acts much like a low-pass filter on the motion, leading to a smoother final output.

As a third example, we show several versions of an animation in which a character rises from a bow to make an emphatic two handed gesture. The intensity of the motion is increased by increasing the extent and adjusting the transition functions. A succession is again used to increase the sense of flow. Figure 4 shows three different versions of the motion with different extension.

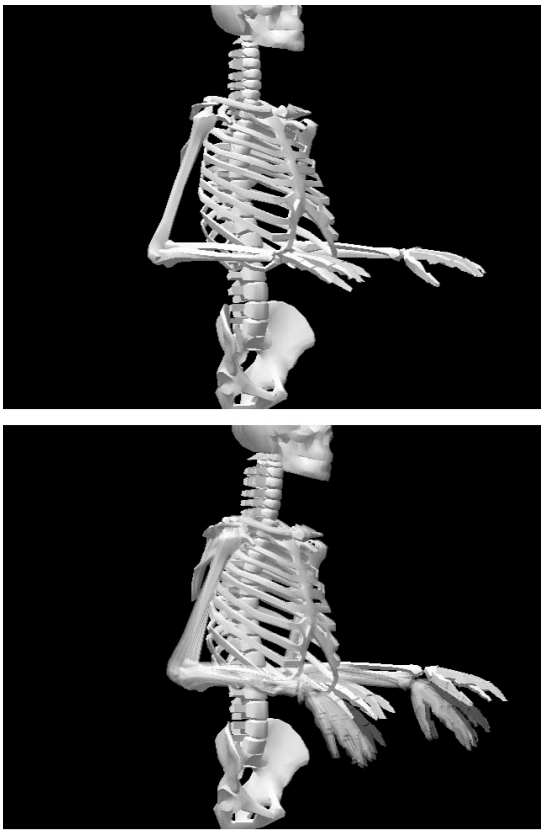


Figure 3: Comparison of kinematic and dynamic motion: The top image shows 0.4s from a kinematic version of the "go left" animation and the bottom image shows the same 0.4s from a dynamic version of the animation.

6 Conclusion and Discussion

Determining the correct movements for a specific character remains a challenging problem. The performing arts

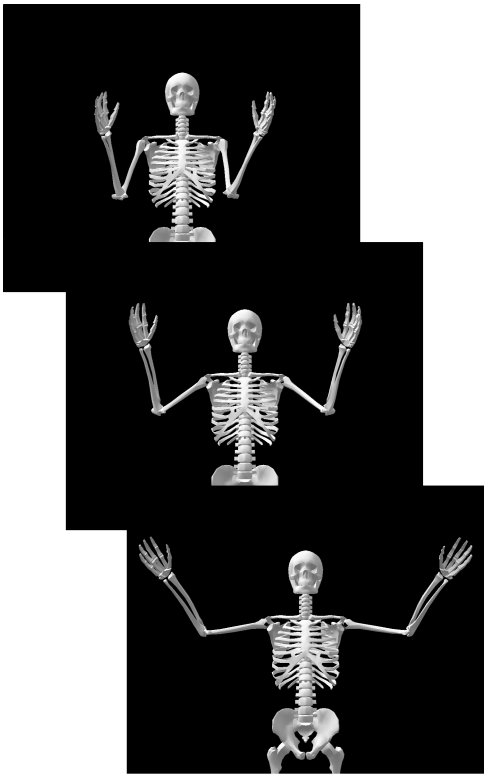


Figure 4: Three different extent versions of the same gesture.

literature, however, offers a rich source of information that can be applied to the task. We have tried to ground our computer animation tools in the lessons learned from the arts community. This helps to ensure that the problems we solve are the problems that need to be solved in order to effectively model expressive motion.

In this paper, we have summarized a number of key lessons from the arts community, providing a new organization of these ideas into the categories of *shape*, *transition* and *timing*. Computational models of a number of these properties were presented along with an architecture that allows various properties to be combined and new properties to be added to the system. The effectiveness of the approach was shown by demonstrating how motion could be quickly customized for a particular character.

Our work provides a vocabulary and set of computational models that can be used by high level character systems to modify the low-level properties of movement to match the personality and mood of a given character. Prescripted motion limits the actions a character can perform. Our hope is that the ideas contained here will be useful in developing character systems that can benefit from the flexibility of automatically generated motion while still offering expressive, engaging characters.

Given the difficulty of the problem, there is much more work to be done. Our research priorities are to develop the character sketch and provide a tighter set of rules for integrating the various movement properties. In addition, faster simulation techniques and more robust dynamic balance control would help make physics-based animation practical for real time applications.

Acknowledgements

Financial support for this work was provided by NSERC and the Department of Computer Science, University of Toronto. The software described here is built on top of DANCE (Dynamic Animation aNd Control Environment), developed by Victor Ng-Thow-Hing and Petros Faloutsos. DANCE is available at: http://www.cs.ucla.edu/magix/projects/dance/index.html Early on in this work, Prof. Stephen Johnson provided some very useful pointers into the arts literature.

References

David Alberts. *The Expressive Body: Physical Characterization for the Actor*. Heinemann, Portsmouth, N.H., 1997.

Kenji Amaya, Armin Bruderlin, and Tom Calvert. Emotion from motion. *Graphics Interface '96*, pages 222–229, May 1996. ISBN 0-9695338-5-3.

Eugenio Barba. Dilated body. In Eugenio Barba and Nicola Savarese, editors, *A Dictionary of Theatre Anthropology: The Secret Art of The Performer*. Routledge, London, 1991a.

Eugenio Barba. Theatre anthropolgoy. In Eugenio Barba and Nicola Savarese, editors, *A Dictionary of Theatre Anthropology: The Secret Art of The Performer*. Routledge, London, 1991b.

Matthew Brand and Aaron Hertzmann. Style machines. *Proceedings of SIGGRAPH 2000*, pages 183–192, July 2000. ISBN 1-58113-208-5.

Armin Bruderlin and Lance Williams. Motion signal processing. *Proceedings of SIGGRAPH 95*, pages 97–104, August 1995. ISBN 0-201-84776-0. Held in Los Angeles, California.

Diane M. Chi, Monica Costa, Liwei Zhao, and Norman I. Badler. The emote model for effort and shape. *Proceedings of SIGGRAPH 2000*, pages 173–182, July 2000. ISBN 1-58113-208-5.

Jean Dorcy. *The Mime*. Robert Speller and Sons, Publishers, Inc., 1961. Translated by Robert Speeler, Jr. and Pierre de Fontnouvelle.

Sergei Eisenstein and Sergei Tretyakov. Expressive movement. In Alma Law and Mel Gordon, editors, *Meyerhold, Eisenstein and Biomechanics: Actor Training in Revolutionary Russia*. McFarland and Company, Inc., Publishers, Jefferson, North Carolina, 1996.

A. G. Feldman. Functional tuning of the nervous system with control of movement or maintenance of a steady posture - ii. controllable parameters of the muscles. *Biophysics*, 11(3):565–578, 1966.

Michael G. Hollars, Dan E. Rosenthal, and Michael A. Sherman. *SD/FAST User's Manual*. Symbolic Dynamics Inc., 1994.

Doris H. U. Kochanek and Richard H. Bartels. Interpolating splines with local tension, continuity, and bias control. *Computer Graphics (Proceedings of SIGGRAPH 84)*, 18(3):33–41, July 1984. Held in Minneapolis, Minnesota.

Rudolf Laban. *The Mastery of Movement*. Northcote House, London, fourth edition, 1988. Revised by Lisa Ullman.

John Lasseter. Principles of traditional animation applied to 3d computer animation. *Computer Graphics (Proceedings of SIGGRAPH 87)*, 21(4):35–44, July 1987. Held in Anaheim, California.

Joan Lawson. *Mime: The Theory and Practice of Expressive Gesture With a Description of its Historical Development*. Sir Isaac Pitma and Sons Ltd., London, 1957. Drawings by Peter Revitt.

Michael Neff and Eugene Fiume. Modeling tension and relaxation for computer animation. In *ACM SIGGRAPH Symposium on Computer Animation*, pages 81–88, July 2002.

Michael Neff and Eugene Fiume. Aesthetic edits for character animation. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 239–244, July 2003.

Ken Perlin. Real time responsive animation with personality. *IEEE Transactions on Visualization and Computer Graphics*, 1(1):5–15, March 1995. ISSN 1077-2626.

Ken Perlin and Athomas Goldberg. Improv: A system for scripting interactive actors in virtual worlds. *Proceedings of SIGGRAPH 96*, pages 205–216, August 1996. ISBN 0-201-94800-1. Held in New Orleans, Louisiana.

Katherine Pullen and Christoph Bregler. Motion capture assisted animation: Texturing and synthesis. *ACM Transactions on Graphics*, 21(3):501–508, July 2002.

Charles Rose, Michael F. Cohen, and Bobby Bodenheimer. Verbs and adverbs: Multidimensional motion interpolation. *IEEE Computer Graphics and Applications*, 18(5):32–40, September - October 1998. ISSN 0272-1716.

Ted Shawn. *Every Little Movement: A Book about Francois Delsarte*. Dance Horizons, Inc., New York, second revised edition, 1963.

Jennifer Tarver and Kate Bligh. The physical art of the performer, November 1999. Workshop: A 30 hr. intensive introduction to Laban and Grotowski held at the Nightwood Theatre studio, Toronto.

George Taylor. Francois delsarte: A codification of nineteenth-century acting. *Theatre Research International*, 24(1):71–81, 1999.

Frank Thomas. The future of character animation by computer. In Bill Kroyer and Phillipe Bergeron, editors, *3D Character Animation By Computer*, volume 5 of *Siggraph 87 Course Notes*. ACM Siggraph, 1987.

Frank Thomas and Ollie Johnston. *The Illusion of Life: Disney Animation*. Abbeville Press, New York, 1981.

Bill Tomlinson, Marc Downie, Matt Berlin, Jesse Gray, Derek Lyons, Jennie Cochran, and Bruce Blumberg. Leashing the alphawolves: Mixing user direction with autonomous emotion in a pack of semi-autonomous virtual characters. In *ACM SIGGRAPH Symposium on Computer Animation*, pages 7–14, July 2002.

Munetoshi Unuma, Ken Anjyo, and Ryozo Takeuchi. Fourier principles for emotion-based human figure animation. *Proceedings of SIGGRAPH 95*, pages 91–96, August 1995. ISBN 0-201-84776-0. Held in Los Angeles, California.

Speaking and acting - interacting language and action for an expressive character

Sandy Louchart*
*Centre for Virtual Environments
University of Salford M5 4WT
S.Louchart@salford.ac.uk

Daniela Romano†
†Deparment of Computer Science
University of Sheffield S1 4DP
D.Romano@dcs.shef.ac.uk

Ruth Aylett*
‡Centre for Virtual Environments
University of Salford M5 4WT
R.S.Aylett@salford.ac.uk

Jonathan Pickering*
§Centre for Virtual Environments
University of Salford M5 4WT
J.H.Pickering@salford.ac.uk

Abstract

We discuss in this paper the FearNot! application demonstrator, currently being developed for the EU framework V project VICTEC. It details the language structure, content, interactions management, general architecture and design of the FearNot! Demonstrator, as well as presenting the VICTEC project and its motivations. This paper also focuses on the different sets of Speech Act inspired language action lists developed for the project and discusses their use for an interactive language and action system for the elaboration of expressive characters. The paper also presents early development and implementation work as well as system and speech evaluation planning.

1 Introduction

This paper discusses the language system being developed for the EU framework V project VICTEC - Virtual ICT (Information and Communication Technologies) with Empathic Agents. This seeks to use virtual dramas created by interaction between intelligent virtual agents as a means of dealing with education for children aged 8-12 in which attitudes and feelings are as important as knowledge. The project thus focuses on Personal and Social Education, which includes topics such as education against drugs, sex education, social behaviour and citizenship. The topic specifically addressed by VICTEC Victe (2004)is education against bullying. The project expects to contribute to an understanding of the role of empathy in creating social immersion, and to the evaluation of virtual environment ICT for child users. It also expects to contribute to a deeper understanding of empathy and its role in bullying, and to the relationship between Theory of Mind (TOM) Woods et al. (2003) and bullying behavior. The building of empathy between child and character is seen as a way of creating a novel educational experience.

An output of the project is the FearNot! demonstrator, currently under construction, figure 1. The overall interactional structure of this demonstrator alternates the enaction of virtual drama episodes in which victimisation may occur, and interaction between one of the characters in these dramas and the child user, who is asked to act as their 'invisible friend' and help them to deal with the problems observed in the dramatic episodes. The advice given by the child will modify the emotional state of the character and affect its behaviour in the next episode. The narrative approach undertaken by the VICTEC project is the one of Emergent Narrative Aylett (1999), Louchart and Aylett (2002), Aylett and Louchart (2004). The research on Emergent Narrative aims at finding and elaborating a narrative structure appropriate and suitable for an optimal use of Virtual Environments, combining the entertainment values of both storytelling and virtual experiencing.

The FearNot! Demonstrator represents an intuitive interface between the virtual world and the child user. The characters appearing in the demonstrator have been modelled to be believable rather than realistic, with the use of exaggerated cartoon-like facial expressions. Evaluation to date Woods et al. (2003) has shown that providing the narrative action is seen as believable, lack of naturalism is not perceived as a problem by prospective child users. FearNot! Draws upon feelings of immersion and suspension of disbelief, essential characteristics of Virtual Reality (VR) and Virtual Environments (VE), in order to build empathy between the child and the virtual character as the child explores different coping behaviours in bullying.

2 Integrating language and action

Unlike most dialogue systems or talking heads, VICTEC mixes language interaction with physical actions. Bullying can be categorised as verbal, physical, or relational (manipulating social relationships to victimise), so

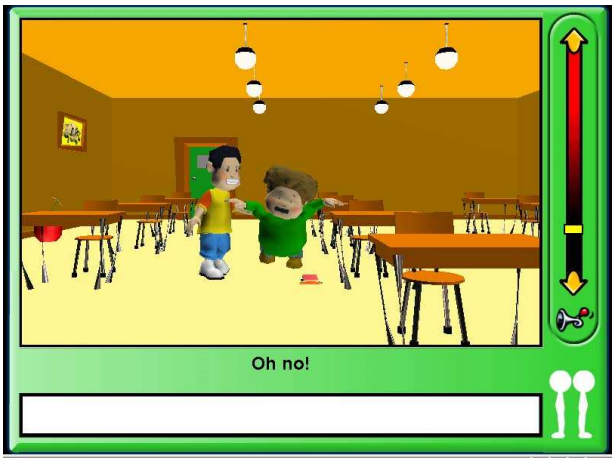


Figure 1: A screen shot of the FearNot! demonstator.

that actions such as pushing, taking possessions and hitting must be modelled. Each character displayed in the FearNot! Demonstrator is provided with its own autonomous action selection mechanism, and the overall architecture is shown in figure 2. An appraisal of events and the other characters is carried out, using the emotion-modelling system of Ortony, Clore and Collins Ortony et al. (1988) and the resulting emotional state is combined with the character's goals and motivations to select an appropriate action. Thus a common representation for both physical actions and language actions is needed so that both can be equally operated upon by the action-selection mechanism.

This representation is provided by the concept of a speech act Austin (1962), Searle (1969), defined as an action performed by means of language. Here, language is categorised by its illocutionary force, that is, the goal that the speaker is trying to achieve; the same view of action taken by an action-selection mechanism, and highly relevant to bullying scenarios. Speech Acts however work at a very high level of abstraction (e.g. assert, promise, threaten) and only a subset of those generally used are relevant to bullying scenarios. Moreover much of the subsequent work - such as that in Dialogue Acts Bunt (1981) - has taken place in language-only domains and does not address the close relationship between language and actions required for the VICTEC project. It was therefore decided to define a set of language actions in the spirit of speech acts, using a corpus of bullying scenarios constructed by school children using a story-boarding tool Kar2ouche Education (2004).

Of course a speech act does not uniquely specify the utterance in which it is expressed - its locutionary form. Moreover it was created as an analytic tool, while the language system being created here must function in a generative capacity (see Szilas (2003) for other work with this aim). In addition, language and other actions must form coherent sequences, accepted as such by the child users. The approach must also take account of cross-cultural language practices such as the specific language used in schools in the UK, Portugal and Germany, the countries of the project partners.

Finally, there are two different contexts in which the language system must work. The first is within dramatic episodes in which characters interact with each other. The second is between episodes in which the character must interact with the child user.

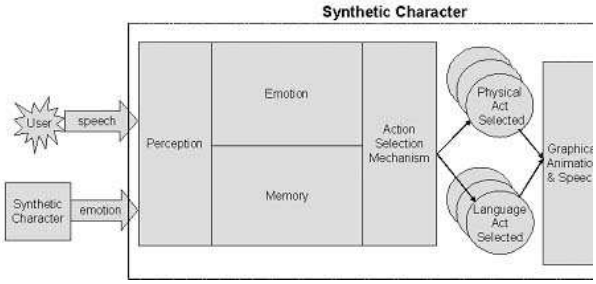


Figure 2: The architecture developed to support synthetic characters.

2.1 From action to utterance

An action can be described as a collection of the following instances: an object on which the action can be performed (an object in the environment or another character); the agent performing the action; the action priority (used to order and deal with conflicting actions); the context in which the action is performed (i.e. location, props, internal goal, history of previous actions, topics); the emotional status of the character at that time, and the utterance (relating to the language action) that should be played, and the animation of the body of the character involved and accompanying gestures. The emotional status of the character feeding into its action-selection mechanism will determine whether the action to be performed is implemented via language action, physical activity or both.

Assuming that the next action selected is physical, from a current pose of the character a series of animations are possible: however to reach the current selected one it might be necessary to introduce an intermediate pose (i.e. next action: walk to the door. Current pose: sitting. Intermediate pose necessary: stand up).

We can visualise this as a tree of behaviours, where, from a current state, the next animation is possible only when the correct status of the character is reached and that action can began, requiring the introduction of an intermediate pose. See figure 3.

In order to generate the utterance for a selected language action, it has been decided to use a shallow-processing approach, as originally used in ELIZA Weizenbaum (1966) and more recently in chat bots Mauldin (1994). The rationale for this approach is that it takes little processing resource compared to a deep approach based on parsing and semantics, thus allowing the

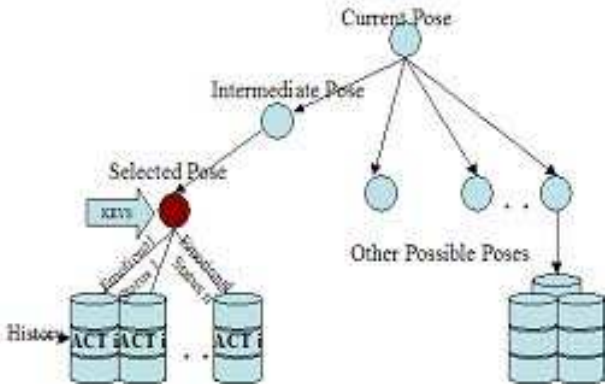


Figure 3: The search tree showing the space of possible behaviours.

graphics engine the resource it needs to run in real-time. In addition, such systems can show surprising resilience in limited domains such as that of FearNot!, in which the language to be used is specific to the bullying scenarios. To prevent problems experienced with such systems in dealing with unexpected inputs, the FearNot! demonstrator will specifically drive the conversation in child-agent mode by using leading questions with a limited range of options for answer. Wizard of Oz studies are in progress to determine in more detail what language coverage will be required.

The FearNot! demonstrator's natural language system is required to adopt and implement techniques and technologies inspired by research in conversational agents Braun (2002), Braun (2003), Rist et al. (2003), Prendinger and Ishizuka (2001), Prendinger and Ishizuka (2002). Similar approach as already successfully been implemented in FACADE Mateas and Stern (2003), where the agent has the possibility of chosing between actions and language when interacting with another agent or a user. However, FACADE's low level of abstraction approach would be hard to manage for VICTEC and would require more development than actual resources allow.

In agent-agent interaction, the language system starts with the language action generated by the action-selection system, which has the advantage of knowing exactly what action (language or otherwise) it is responding to. This indexes a group of utterance templates in which the previous utterance or physical action is used to fill in variable slots with an appropriate choice. For example, assume the utterance from the other agent was" I like flowers", the following group of utterance is selected: I like ... too, why do you like ...?, what do you find in ...?. The first unused utterance here is: "why do you like ...?" the dots are filled with the recognized object of the discourse in the user's input: flowers. The generated character utterance is "why do you like flowers?".

Child and character interaction is different. Here the previous language action is not known, but must be inferred. The incoming text is matched against a set of language templates, and the language and action index is then taken as the starting point for the language action with which the agent must respond as discussed below. Since an objective is to retain control of this dialogue by keeping the conversational initiative with the character, the Finite State Machine structures discussed below can also be used to generate expectations about what language actions the child has produced. In addition 'sentence starters' will be provided to help the child with the typing burden and these will provide clues to the language action the child has carried out.

3 The FearNot! Speech Act Knowledge-base

Since the FearNot! demonstrator developed for use in the VICTEC project includes in the same application both agent-to-agent and agent-to-user interactions, it is essential that such this feature is taken into account when designing and developing the language actions' articulation and content. For this reason, the FearNot! demonstrator must combine the use of a bullying themed language while operating on these two different and distinct levels.

In order for the FearNot! Demonstrator to successfully meet VICTEC's evaluation objectives, it is crucial that continuity and coherence is maintained during interactions (contextualisation) between agents while insuring that the communication is engaged and led by an agent when agents and users interact together. This not only fundamentally affects the design of the language system, it also requires the design of two distinct sets of actions, independent of each other as just discussed. For instance, in the case of an agent-to-agent communication, the process starts with the selection of a language action and ends with the selection of an utterance. The opposite occurs in the case of agent-to-user communication since the system needs to recognise an utterance via keywords and then select an appropriate language action or action in order to provide an answer to the user.

3.1 Action categorisation

A set of appropriate actions for bullying and victimization interactive scenarios has been identified. Those actions can be triggered and generate agent utterances according to their emotional states. As such a system is dealing with a number of actions and utterances, we have grouped the entire language content within three categories, Help, Confrontation and Socializing table 1.

Each category includes a variable number of appropriate language and other actions. For instance, the confrontation category contains a considerably larger number of actions than the help section since there is a very limited number of coping behaviours available in dealing with bullying Woods et al. (2003).

Table 1: Actions catagories and example listings

Catagorie	Listings
HELP	Ask for help / Offer help / Help question / Help advice / Help introduce to friend / Help talk to someone / Help invitation / Offer protection / Non assistance confirmation
CONFRONTATION	Order / Aggressive questioning / Do / Forbid / Defiance / Tease / accusations / Insult / Threat / Aggressive answer / Apology / Abandon action / Action / Hit / Lie / Steal / Obey / Deny / Ask why / Beg / Claim back / Leave / Struggle
SOCIALISING	Greeting start / Topic introduction / Exclusion topic introduction / Information topic / Information exclusion topic / Questions topic 2 / Question topic 3 / Exclusion question 2 / Exclusion question 3 / Exclusion invite / Invitation / Greeting end

The Help set articulates the actions needed to generate offering-help interactions between agents. It covers the interactions needed for the generation of enquiries from agent-to-agent with respect to emotional states and related goals. In addition, this function also generates advice and offers such as help, protection or assistance. As with the other categories, the Help language and action set category has been designed according to a potential sequential structure. This can be triggered either by an agent asking for the help of another or in response to an aggressive action carried out on a particular agent.

The Confrontation language and action set provides the necessary content for an altercation between two different agents. This category covers most of the physical bullying expressions and involves threats, insults, orders, aggressive behaviour that leads to aggressive actions and violent behaviour. Finally, the Socialising category includes language and actions that can be used in social discussion by pupils in schools (sports, homeworks, music, video games) and language and actions that can be used in generating relational bullying. Relational bullying is different from physical bullying, depending on social exclusion and should therefore be integrated into social interaction, as opposed to help or confrontational actions. Although the structure is simple in theory, its implementation requires a large number of utterances and topics.

Table 2: An example sequence of speech act utterances

Speech act	Utterance
DO	You, [order] now!
If speech act = DENY	You must be joking, [rejection] [insult]
If speechact = OBEY	Ok, but please don't hurt me!

3.2 Actions Finite State Machine (FSM)

Each action category also possesses its own organisation and consequently requires the design of its own Finite State Machine (FSM). A language action is coherent to both the system and the user if organised into structured speech sequences. While this has to be taken into account it is also essential that the speech system focuses on organising the possible sequences of utterances and ensure the transfer and communication of content without interfering with the agent action selection mechanism. Since, as with all speech system, there are issues of contextualisation, the utterances that constitute the content of the system are formed of templates that can be filled appropriately by the speech system, based on keyword recognition.

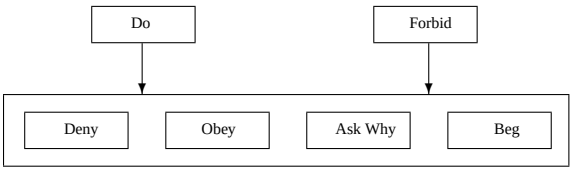


Figure 4: An example of a sequence of speech acts.

Each FSM integrates the language actions relative to the category itself but also potential elements of answers for discussion or interaction. For instance, the actions 'DO' or 'FORBID' in a confrontational situation will be followed by the actions 'DENY', 'OBEY', 'ASK WHY' or 'BEG' Figure 4, to retain conversational coherence.

The VICTEC language actions and utterances have been elaborated according to sequence of actions observed in the scenarios developed by school children mentioned above.

Speech acts are materialised on the FearNot! Demonstrator by utterances. The situation presented in Figure 4 would produce, in case of denial or obedience from the victim the following exchange shown in table 2.

3.3 User-to-agent language action design

Since, the language generated by the user is highly ambiguous and there are no means for the system to understand the meaning of a sentence, the user-to-agent interaction, as we mentioned previously, needs a different approach. As a sentence can only be "understood" by the

keywords included in it, it seems sensible to leave the initiative to the agent rather than the user. The fact that the system leads the conversation with the user presents an advantage in terms of believability for the speech system in the sense that, the system can be expectation driven and can expect a certain type of answer from the user and adjust and compare the answer to a set of pre-defined templates. Although the system could not understand its human interlocutor, it could generate a high level of believability and interact with its user by asking simple and adequate questions.

In order for the agent to keep the upper hand in terms of interaction with the child user, it must be the one asking for advice and the one who generally ask questions to which the child user is expected to answer.

It has been decided, due to the high possibility of misspelling from the children who are going to use the system, that the language system includes a keyword recognition feature that should allow it to recognize the intention of the user and make the association with existing categories of actions.

Although the speech system, in the case of a user-to-agent interaction mainly requires language actions expressed through utterances from the agent rather than the user, it is however important to predict and anticipate answers in regards to the different possibilities that are been offered to the user. Since the VICTEC project is mainly being tested by children aged between 8 and 12 years old, it has been thought that such approach would also have the advantage of helping them in formulating their answers to the agent. The speech system is, in the particular case of the VICTEC project divided into two distinct sections, the agent language actions and utterances and the user's answers.

Table 3: User and Agent Language actions lists

Catagorie	Listings
AGENT	Ask for advice / Ask again / Prompt / Cannot understand statement / Ask for reason / justification / Thank user for advise / Confirm advice with user / Express reproach to user / advice rejection / Express disappointment towards user / Report result of interaction / Beg for help
USER	Give advice / Refuse to give advice / Ignore the agent / No answer / No helpful comments / Advice confirmation / Justification

4 Implementing Language Acts

In this section we address the problem of implementing speech acts within the Victec system. The approach will be to take the small set of speech acts defined above in section 3 and find a structural similarities in the sentences used to represent them.

4.1 Speech and Dialog Acts

Speech acts are descriptions of utterances in terms of the function they perform Searl (1975). The simplest example is the sentence, "I name this ship Lady Day.", which can be classified as the act of naming.

More recently *dialog acts* have been proposed Bunt (1994), which extend the concept of speech acts to include analysis of the conversational use of an utterance. For example the utterance "I'm sorry, do you mean the first or second letter", would be classified as a repair action, as is clears up failure to communicate properly.

Formally speech act theory may be defined as a branch of pragmatics that classifies utterances by the role or use that they serve in a communication. The role played by an utterance is the function that it provides as distinct from the semantic meaning of the utterance. This is best clarified by an example, consider the two following sentences.

1. Give me the keys.
2. Please may I have the keys.

The meaning of both sentences is the same, the utterer is asking for the keys. However the pragmatics is very different, in the first case the utterer is giving an order while in the second a request is being made. It is important to note that speech acts are not unique classifiers of utterances. It is equally possible for instance, to classify both the previous utterances as communications of a desire. Some examples of speech acts follow.

Question to gain information the questioner needs some information, for example "Where is the milk?".

Exam question the questioner knows the answer but is testing the candidate, for example "Who is the president of Mexico?".

Continuer during a long monologue a speaker will pause, giving chance for the lister to indicate that they are still following the speaker, for example "Yes go on".

Each of the above examples can be satisfied by many utterances and it is impossible to tell from the language act alone how to construct or choose and utterance. The problems are: speech acts alone contain no semantic information; speech acts are not unique; and speech acts cannot in general be mapped to syntax. It is claimed Jurafsky and Martin (2000), that classifying utterances into

speech acts is an AI complete problem, meaning that a human being, or a computer system equivalent to a human being, would be required to correctly classify them.

4.2 Microgrammes

Although the general problem of classifying speech acts is currently unformalisable. It is possible to produce automatic classifiers that give partial coverage of common acts. The method for doing this exploits the fact that many speech acts correlate to structural features in a conversation. These structures, introduced by Goodwin (1996) have been called *microgrammes*. They comprise set of features which are classified into three different types.

Words and collocation certain words and particularly combinations of words (collocation) indicate some speech acts. For example the words 'who, when, where', indicate questions.

Prosody the tone of voice used in an utterance may indicate its intended act. In English questions,for example, can be indicated by a rising intonation at the end of a sentence.

Conversational Structure the current context and the immediate predecessor statements may give an indication of the speech act. A simple example of which is that the utterance after a question is probably a reply or a request for clarification.

In the case of a textural system such as Victec prosody will have no part to play. The burden of the work will have to be achieved using pattern matching to identify words and collocation. Although hopefully, some support can be provided through the use of context.

4.3 Language Acts

Because the Victec project is centred on the development of autonomous agents that interact in a virtual environment by the use of actions, it was natural to use speech acts to define the agent's speech system. This would allow the agent to remain in an action reception, action appraisal, action selection loop.

The problem is the lack of semantics and multiple definitions of speech acts. To allow for the first some semantic information has had to be added to the agents actions. We have called the combination of speech act plus semantic information *language acts*.

We intend to solve the second problem, of how to identify sentences with speech acts, by applying microgrammers to the very small set of sentences that have been classified in the knowledge base. A microgrammer can be written for each speech act. When parsed in conjunction with the semantic information and contextual knowledge of the source and sender of the speech act the microgramme will generate a sentence.

For example consider the act of greeting a person. The set of possible sentences is very small, consisting of a greeting word, possibly the name of the person being greeted, and possibly a greeting question.

Hello
Hello Sue
Hi
Hi Tom
Hi Jo, how are you?

We can immediately see a general form to these sentences and written down in Backus Naur Form (BNF) it is.

⟨ Greeting word ⟩ ⟨ ToName? ⟩ ⟨ status_question? ⟩
⟨ Greeting word ⟩ = ⟨ Hello ⟩
= ⟨ Hi ⟩

⟨ ToName? ⟩ = ⟨ `receiver` ⟩

⟨ status_question? ⟩ = ⟨ how are you? ⟩
= ⟨ are you all right? ⟩

The term `receiver` is a context variable that is set by the semantic information in the language act.

4.4 Implementing Agent to Agent Language Acts

The database of language acts will be specified in XML, which has the expressive power of a context free grammar and so can express BNF statements. Each language act would be implemented as a template consisting of rules. The rules are implemented separately and may be recursive.

A language act generated by an agent will contain the name of the sending and receiving agents, the type of act and some semantic information, that is act specific. When received the template for the requested act is found, context variables are set and then the rules are repeatedly applied until a response has been formed.

For example the following XML specifies a request for a greeting from agent Tom to agent Sue, with an optional question about Sue's current status.

⟨ Type ⟩ Greeting ⟨ /Type ⟩
⟨ Sender ⟩ Tom ⟨ /Sender ⟩
⟨ Receiver ⟩ Sue ⟨ /Receiver ⟩
⟨ SemanticValue name="true"
statusQuestion="random" /⟩

Fist the variables `sender` and `receiver` would be instantiated to the names Tom and Sue, then the template for the language act looked up. Using a greeting language act as specified above in section 4.3 the template requires a greeting word that can be 'Hello' or 'Hi', as there is no other information a random choice would be made.

Next the because the name attribute of the semantic information was set to `true` the 'ToName' must be

added. The rule for this evaluates to the context variable `receiver`, so the value of Sue would be added to the reply.

Finally the status question attribute of the semantic information was set to `random` so the the language system will chose with equal probability between the adding and not adding a status question. If a question is to be added the rule is evaluated which gives a random choice of two possible questions.

This results in one of the six following possible greetings being generated.

Hi Sue
Hello Sue
Hi Sue how are you?
Hello Sue how are you?
Hi Sue are you all right?
Hello Sue are you all right?

The rules can be recursive, allowing a rule to contain other rules, which will allow the case and gender agreements of German and Portuguese to be applied.

4.5 Implementing User Agent Dialog

In the case of user agent dialog the problem of classifying the user's speech acts and extracting the semantic information for appraisal by an agent must be addressed. As was stated in 4.1 this is in general a very difficult problem.

The problem may be simplified by noting three features that will apply in the case of FearNot.

- The dialog will be very short and focused only on the previous bullying episode.
- The users will be children of age 10 and so will only type simple sentences.
- In order to help the children buttons providing part formed sentences will be provided.

It is hoped that these features will so constrain the input domain as to allow the identification of speech acts using pattern matching to look for words and collocation supported by some conversational structure information.

5 Conclusion

This paper has described the interactional structure and articulation of the language system being developed for the VICTEC project and reported on progress made so far. It also detailed the different language actions and their categorisation in relation to the specific theme of bullying.

The language system and its content have been developed based on actual language currently in use amongst school children, however it requires iterative refinement and testing of both its efficiency and the coherence as well as evaluation of its capacity to suspend or limit the initial disbelief commonly generated by this type of system. A series of Wizard of Oz experiments Maulsby et al. (1993) along with psychological and usability evaluations Woods et al. (2003) are therefore planned. Further evaluation of the whole FearNot! Demonstrator is also planned: for example, a large sample of children (400) will take part in a psychological evaluation at the University of Hertfordshire in June 2004. However, while the agent architecture of the system and systems integration is still under development, language graphical content has already been produced for preliminary evaluation and the VICTEC team is working with the aim to present a first prototype of the system by April 2004.

References

J L Austin. *How to Do Things with Words*. Harvard University Press, 1962.

R Aylett. Narrative in virtual environments - towards emergent narrative. In *Proceedings of the AAAI Fall Symposium on Narrative Intelligence.*, 1999.

R Aylett and S Louchart. Towards a narrative theory of vr. *Virtual Reality Journal, special issue on storytelling*, 7, January 2004.

N Braun. Automated narration - the path to interactive storytelling. In *Workshop on Narrative and Interactive Learning Environments, Edinburgh, Scotland*, 2002.

N Braun. *Storytelling & conversation to improve the fun factor in software applications*, chapter 19. Kluwer Academic, 2003.

H Bunt. Rules for the interpretation, evaluation and generation of dialogue acts. IPO annual progress report 16, Tilburg University, 1981.

Harry C. Bunt. Context and dialogue control. *THINK Quarterly*, 3:19–31, 1994.

Immersive Education. Kart2ouche. www.kar2ouche.com, 2004.

C Goodwin. Transparent vision. In E Ochs, E Schlegloff, and S Thompson, editors, *Interaction and Grammer*. Cambridge University Press, 1996.

D Jurafsky and J Martin. *Speech and Langugage Processing*. Prentice Hall, 2000.

S Louchart and R Aylett. Narrative theories and emergent interactive narrative. In *Proceedings Narrative and Learning Environments Conference NILE02 Edinburgh, Scotland*, pages 1–8, 2002.

M Mateas and Andrew Stern. Integrating plot, character and natural language processing in the interactive drama faade. In *Proceedings Technologies for Interactive Digital Storytelling and Entertainment (TIDSE) conference, Darmstadt, Germany*, pages 139–152, March 2003.

M Mauldin. Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition. In *Proceedings AAAI 94*, 1994.

D Maulsby, S Greenberg, and R Mander. Prototyping an intelligent agent through wizard of oz. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 277–284. ACM Press, 1993.

A Ortony, G L Clore, and A Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, 1988.

H Prendinger and M Ishizuka. Let's talk! socially intelligent agents for language conversation training. *IEEE Transactions on Systems, Man,and Cybernetics - Part A: Systems and Humans*, 31(5):465–471, 2001. (Special Issue on Socially Intelligent Agents - The Human in the Loop, ed. K Dautenhahn).

H Prendinger and M Ishizuka. Evolving social relationships with animate characters. In *Proceedings Symposium of the AISB-02 Convention on Animating Expressive Characters for Social Interactions, London, UK*, pages 73–78, 2002.

T Rist, M Schmitt, C Pelachaud, and M Bilvi. Towards a simulation of conversations with expressive embodied speakers and listeners. In *16th International Conference on Computer Animation and Social Agents*, pages 5–10. IEEE Computer Society, 2003.

J Searl. A taxonomy of illocutionary acts. In P Cole and J Morgan, editors, *Speech Acts: Syntax and Semantics Volume 3*, pages 59–82. Academic Press, New York, 1975.

J Searle. *Speech Acts*. Cambridge University Press, 1969.

N Szilas. Idtension: A narrative engine for interactive drama. In *1st International Conference on Technologies for Interactive Digital Storytelling and Entertainment (TIDSE 2003)*, pages 24–26, 2003.

Victe. Victec home page. http://www.VICTEC.org, 2004.

J Weizenbaum. Eliza - a computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery*, 9:36–45, 1966.

S Woods, L Hall, D Sobral, K Dautenhahn, and D Wolke. Animated characters in bullying intervention. In *IVA 2003*, number 2972, pages 310–314. Springer-Verlag LNAI, 2003.

Expressive characters and a text chat interface

Marco Gillies*
*University College London, Adastral Park Campus
Orion Building GND pp13, Adastral Park,
Ipswich IP5 3RE, UK
m.gillies@ucl.ac.uk

Barry Crabtree†, Daniel Ballin†
†Radical Multimedia Lab, BT Exact
Ross Building pp 4, Adastral Park
Ipswich IP5 3RE, UK
Barry.Crabtree@bt.com, Daniel.Ballin@bt.com

Abstract

Avatars in 3D virtual worlds have the potential to be a highly expressive communication medium, reproducing many of the non-verbal cues available in face to face communication. However, up to now there has been little success in actually achieving this aim. One reason is that explicitly controlling the non-verbal behaviour is difficult and user normally fail to do it. Our solution is to have avatars autonomously generate non-verbal behaviour with some high level control from the user. We present the Demeanour framework for generating non-verbal behaviour and a user interface for influencing the behaviour. This user interface is based around a text chat system and aims to be considerably simpler and easier to use than explicitly controlling behaviour.

1 Introduction

User avatars in virtual worlds are now very graphically sophisticated, but often seem dull and lifeless through their lack of expressive body language or non-verbal communication. As Vilhjálmsson and Cassell (1998) point out this is often not due to the lack of capacity for expressive behaviour but with problems caused by the user having to explicitly control this behaviour. Directly requesting a gesture every time can be very time consuming and distracting from the main task, particularly if the interfaces are very different. For example during conversation going from a text interface to a GUI with buttons and sliders for body language requires the user to move their hands from the keyboard to the mouse. Though the overhead is rather small it can discourage users from performing non-verbal communication, particularly when engaged in a task such as a fast paced conversation, the result is that the avatars again lack expressive behaviour and seem lifeless. This problem is due to the fact that in real life we perform non-verbal communication in parallel with other activities, and do so often sub-consciously. The sub-conscious nature of expressive behaviour causes other problems, as people often do not consciously know what action is appropriate at a given time and so explicit control is difficult. For these reasons Vilhjálmsson and Cassell (1998) propose that user avatars should no longer be passive graphical bodies but should be able to produce non-verbal communicative behaviour autonomously. This is an important and interesting area for expressive character research, providing many new challenges.

However, autonomous behaviour is insufficient, it is no good if the avatar is constantly expressing emotions and attitudes to the world if these are unconnected to the feelings of the user. User interfaces are therefore vital in semi-autonomous avatars, the avatar should fulfil the wishes of the user without the user having to constantly control it. Vihljlmsson and Cassell do provide some user control but the range of options is small and their system will not scale well to more complex models of non-verbal behaviour.

In order to avoid the problems of explicitly controlling body language there are two main constraints on the user interface. Firstly the user should not have to constantly be controlling the avatar. It should produce reasonable body language without user intervention and the user should only have to take action occasionally to achieve a particular effect or to correct an error (it would be even better if the avatar could learn from these corrections). Secondly the user interface should be well integrated with the main interface of the virtual world so that controlling the avatar is not a major distraction. We have chosen in our demonstrator to use a text chat interface which is very common in 3D multi-user virtual worlds and more generally on the internet. Users have a form of conversation consisting of short messages which are sent to the other users on pressing the return key. We have implemented such an interface using the Jabber instant messaging system.

2 Related Work

The origin of our work comes from work on virtual environments and particularly providing embodiments of user, called avatars (see Benford et al. (1995) for an overview). In particular this work is motivated by studies that show that the problem of controlling non-verbal behaviour is a major limitation of collaborative virtual environments such as Tromp and Snowdon (1997)

Our work builds on a body of research on autonomous characters for virtual environments, for example, Blumberg and Galyean (1995); Badler et al. (1993); Tu and Ter-

Figure 1: The different levels of control within the Demeanour framework.

zopoulos (1994), and Rickel and Johnson (1999). There has been extensive research on autonomously producing expressive behaviour a number of types including facial expression (Pelachaud and Poggi, 2002), eye gaze (Cassell et al., 1999; Rickel and Johnson, 1999; Gillies and Dodgson, 2002), gesture (Cassell et al., 1999), style of motion (Chi et al., 2000) and posture (Cassell et al., 2001a; Bécheiraz and Thalmann, 1996).

This work builds on the research by Vilhjálmsson and Cassell (1998) described in the introduction. A number of other researchers have done work on integrating autonomy and user control for avatars. Sengers et al. (2000) discuss semi-autonomous avatars but their system is not concerned with expressive avatars. The TEATRIX system by Paiva et al. (2001) combines direct control of an autonomous character with a more reflective level of control which takes users (in this case school children) out of the virtual world allowing users to update the internal state of their character, while reasoning about their role in the story. However, this form of control is not suitable to all applications, so a method of control that is more integrated with the main interface is needed. Cuddihy and Walters (2000) simplify user interfaces by limiting the set of actions available to the user based on contextual information. Tromp and Snowdon (1997) propose a rule based system for providing automatic behaviour in avatars. Users can edit the rules, but this remains essentially a programming task and will therefore exclude many users.

Our work also builds on a longer tradition of methods for directing autonomous characters including multi-level control system proposed by Blumberg and Galyean (1995) or the Improv system by P. and Goldberg (1996). Similar issues are also important for tele-operated robotic systems (Wilson and Neal, 2001) where the requirements of an operator must be applied to an autonomous robot. In particular our three-layer control model is partially inspired by the work of Scerri et al. (2000).

3 Demeanour

Demeanour is a framework for the generation of non-verbal communication in avatars (figure 2 shows an overview of the system). It is able to generate non-verbal behaviour in real-time based on a user definable behaviour model (currently we are working with psychologically based models). Avatars react appropriately to each other's body language, making it unnecessary for the user to explicitly react to the actions of others. Non-verbal communication expresses itself in a number of ways, or modalities we are currently interested in three: posture, gesture and eye gaze.

3.1 Different levels of control

The aim of this research is to find ways in which end-users can influence the behaviour of semi-autonomous avatars without causing an excessive overhead on their other activities. One method would be to measure the affective state of the user, e.g. using computer vision based analysis of facial expression, and map it onto the avatar. Unfortunately such methods are currently difficult or unreliable. Also users often want to hide their real feelings. Therefore our aim is to develop user interfaces based on direct control but make them as unintrusive as possible. There are two types of control. Users can give commands to the avatar in real time while using the virtual world and interacting with other avatars, we call this real time control. It is also possible to perform customizations on the avatar's behaviour before using the virtual world or between sessions, we call this off-line customization In order to minimize overhead while using the virtual world it is desirable to have as much control as possible provided through off-line customization. We therefore propose a methodology for user control of avatars that divides control and customisation into three levels shown in figure 1:

Behaviour Language. The behaviour of the avatar should not be controlled by a fixed program but it should be entirely definable by the designers of a virtual world. This is important as the behaviour of avatars is likely to vary considerably between different worlds used for different purposes (e.g. between a business conferencing environment and an adventure game). This is capability is provided via a definition language.

Profiles. It has been shown that users of on-line worlds are very keen to customize the graphical appearance of their avatar (Cheng et al., 2002) and it is likely that they will also be happy to perform similar customization on the avatar's behaviour, if easy to use tools are provided. Demeanour therefore contains a set of such tools with which end users can define a profile for their avatar. This profile defines the unique aspects of its behaviour.

Real-time Control. It is also necessary for a user to provide some control of their avatar while interacting on-line. It is important that this is unobtrusive as possible and well integrated with the other tasks to be performed in the world. This interface is the primary focus of this paper.

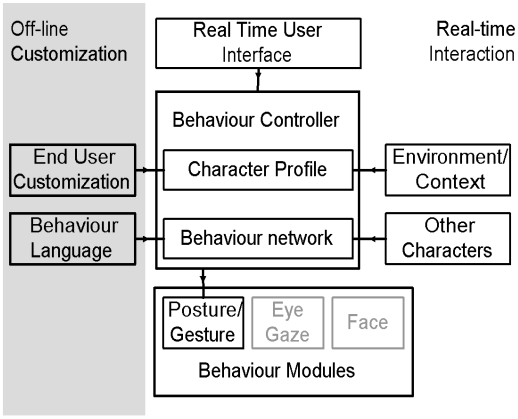


Figure 2: An overview of the Demeanour Framework

3.2 The Demeanour Architecture

We have developed a non-verbal communication framework, Demeanour, which embodies three-level control. Figure 2 show an overview. The main section of Demeanour is a behavioural controller that determines the type of behaviour to be performed at a given time. This contains a behaviour network which gives the structure of the controller and is defined using the behaviour language. The details of the behaviour can vary between avatar and is determined by an avatar profile. This profile is defined by end-users during an off-line customization step. Sections of profile can also be loaded and unloaded during real-time interaction, enabling different behaviour depending on context. Avatars react to the behaviour of other avatars and the end-user can also control the avatar's behaviour using a text-chat based interface. Finally the avatar is actually animated by behaviour modules which interface with the underlying graphics API[1]. Currently Demeanour supports posture, gesture and eye gaze and we are considering adding a facial animation module.

As shown in figure 2 the Demeanour framework combines a number of factors such as user input; context, and the behaviour of other avatars and from the result generates appropriate expressive behaviour. The outputs of the architecture are a number of parameters passed to behaviour generating modules. Input factors are mapped to outputs via of a number of terms which are intermediary values calculated from other terms, including input factors. For example, terms can represent the attitudes values of an avatar such as "affiliation" (see below). The attitude values are calculated from internal parameters of the avatar and input from other avatars and are themselves used to calculate the term values that are used to directly generate behaviour. When there are multiple avatars interacting then the input from each avatars will be different. Rather than requiring a different set of terms for each avatar, Demeanour maintains a single set of terms each of which can be evaluated differently for each avatar. Terms can have a number of types. The simplest are parameters which are single values. These can provide input from other avatars, or they can provide a means for users to control the behaviour of their avatar, either through off-line customisation or real-time control. Terms can also be the combination of a number of other terms, this combination can be done in a number of ways:

- Sum of product terms combine their inputs via multiplication and addition of their values.
- Switch terms choose one of their inputs based on the value of another term.
- Random group terms map their input to a number of outputs. The outputs values are each a proportion of the input, the proportions are chosen at random.

An important part of autonomous behaviour for avatars is the ability to react appropriately to the behaviour of other avatars. Behavioural controllers can therefore take into account the behaviour of other avatars. This is implemented by allowing an avatar access to terms in the behaviour controller of other avatars. Each network specifies a number of terms that are exported and therefore accessible to other avatars. It also defines a number of parameters as being imported, i.e. corresponding to a term belonging to the other avatar. When two avatars start to interact the imports of one are matched, by name, to the exports of the other. For the rest of the interaction the values of the exported terms are used as the values of the imported parameters.

4 Non-verbal behaviour

This section describes a behaviour network for non-verbal communication that we have developed. It models the way people relate to each other or their attitude to each other and is based on the work of Argyle (1975). In our model the attitude of one person to another is expressed through posture and, to a more limited degree, gesture. It is discussed in more detail in Gillies and Ballin (2003).

Though there is an enormous variety in the way that people can relate to each other Argyle identifies two fundamental dimensions that can account for a majority of non-verbal behaviour, affiliation and status. Affiliation can be broadly characterised as liking or wanting a close relationship. It is associated with high levels of eye gaze and close postures, either physically close such as leaning forward or other close interaction such as a direct orientation. Low affiliation or dislike is shown by reduced eye gaze and more distant postures, including postures that present some sort of barrier to interaction, such as crossed arms. Status is the social superiority (dominance) or inferiority (submission) of one person relative to another, we will not discuss it directly in our examples.

[1]The current implementation uses TARA, BT Exact's scene-graph based graphics API

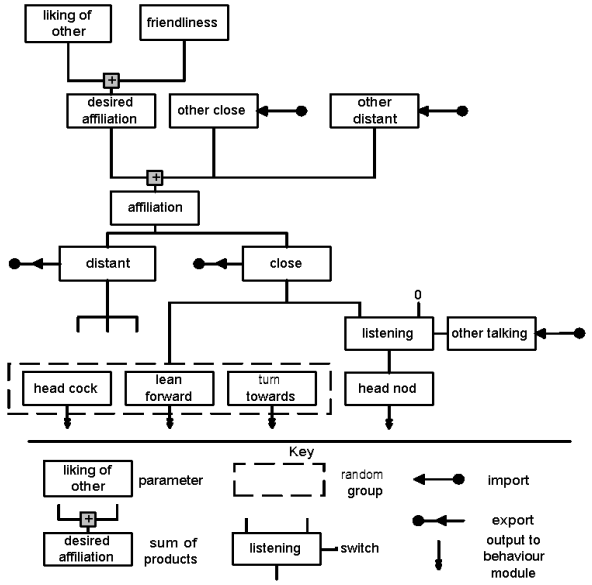


Figure 3: A section of a behavioural controller.

It is also very important that avatars are able to react to each others' behaviour. The relationship between the attitude behaviour of two avatars can take two forms, compensation and reciprocation. Argyle presents a model in which people have a comfortable level of affiliation with another person and will attempt to maintain it by compensating for the behaviour of the other, for example, if the other person adopts a closer posture they will reduce their levels of eye gaze. Conversely there are times where more affiliation generates liking and is therefore reciprocated, or where dominance is viewed as a challenge and so met with another dominant posture. Argyle suggests that reciprocation of affiliation occurs in early stages of a relationship.

Figure 3 shows in diagrammatic form a fragment of the attitude network that deals with affiliation (status is calculated in a similar way) and posture (eye gaze is discussed in section 4.2). At the top of the diagram the actual value for affiliation is calculated as a weighted sum of a number of factors (for the sake of clarity not all the factors used are actually shown). This is done in two stages, firstly the factors depending on the avatar itself are calculated. These factors are represented as parameters (here 'liking of other' and 'friendliness' are shown). Then factors depending on the other avatar's behaviour ('close' and 'distant') are added in. These are import terms and are therefore taken directly from the controller of the other avatar. As the behaviours associated with positive and negative affiliation are very different it is split into two terms, 'close' which is equal to the affiliation and 'distant' which is its negation. Both of these terms are constrained to be greater than 0. The 'close' term is then mapped into actual behaviour (as is 'distant' but it is not shown in the diagram). In order to vary the behaviour produced a random group is used. At semi-regular intervals a new combination of the various behaviours ('head cock', 'lean forward' and 'turn towards') is produced, this combination is always proportional to the value of 'close'. These behaviour types are output terms and are passed as parameters to the underlying animation system. Another affiliative behaviour is head-nodding, but this is only shown when the other person is talking. This behaviour is controlled by a switch node ('listening'), based on a boolean import term which specifies if the other avatar is talking. If 'other talking' is true then 'head nod' is proportional to 'close' otherwise it is zero.

Figure 4 shows examples of body language generated by the Demeanour framework.

4.1 Posture and Gesture

Human bodies are highly expressive; a casual observation of a group of people will reveal a large variety of postures. Some people stand straight, while others are slumped or hunched over; some people have very asymmetric postures; heads can be held at many different angles, and arms can adopt a huge variety of postures each with a different meaning: hands on hips or in pockets; arms crossed; scratching the head or neck, or fiddling with clothing. Computer animated characters often lack this variety of expression and can seem stiff and robotic; however, posture has been relatively little studied in the field of expressive virtual characters. It is a useful cue as it is very clearly visible and can be displayed well on even fairly graphically simple characters.

Research on posture generation has been limited relative to other modalities. Cassell et al. (2001a) have investigated shifts of postures and their relationship to speech, but not the meaning of the postures themselves. As such their work is complimentary to ours. Coulson (2002) uses an OCC model of emotion to generate postures. Bécheiraz and Thalmann (1996) use a one-dimensional model of attitude, analogous to our affiliation, to animate the postures of characters. Their model differs from ours in that it involves choosing one of a set of discrete postures rather than continuously blending postures. This means that it is less able to display varying degrees of attitude or combinations of different attitudes.

The generation of gestures has been studied by a number of researchers. For example, Cassell et al. (1999) have produced a character capable of extensive non-verbal behaviour including sophisticated gestures. Chi et al. (2000) present a way of generating expressive movements, similar to gestures using Laban notation. Gestures are closely related to speech and should be closely synchronised with it. Cassell et al. (2001b) present a system what parse text and suggests appropriate gestures to accompany it. Gestures are less closely related to attitude than posture, though some connection can be made, for example head nodding while listening is a generally affiliative gesture.

As described in the previous section the attitude model

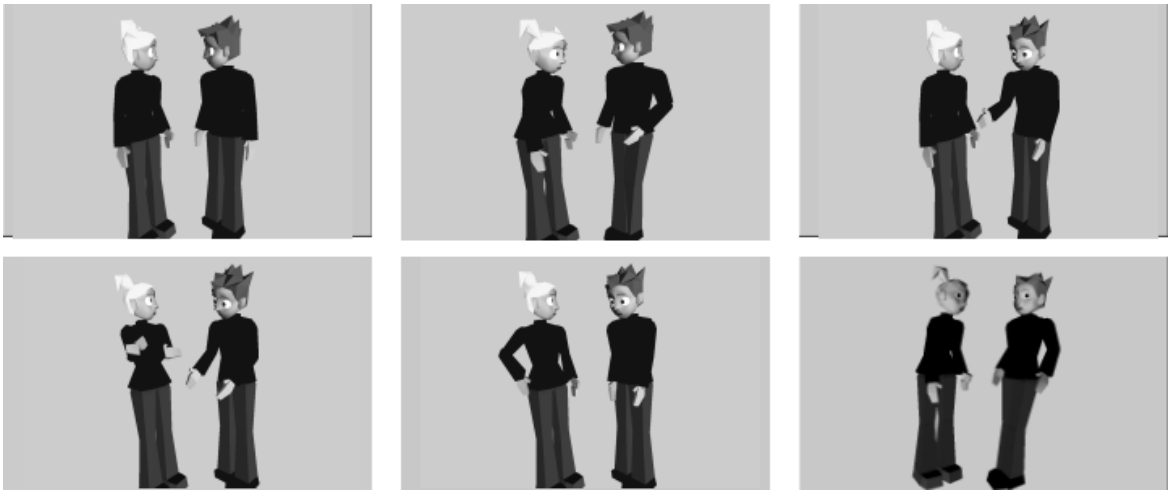


Figure 4: Examples of body language generated by the Demeanour framework due to different attitudes between the avatars. Clockwise from top left: mutual gaze; close and relaxed postures; the male character is gesturing while talking; the female character has a distant, hostile posture; the female character has a high status, space filling posture the male character has a low status, submissive posture; the male character is relaxed (a high status posture) and the female character has a close posture.

generates a high level description of the behaviour of the avatar in terms of a value of each of a number of behaviour types. The behaviour modules themselves must translate this description into concrete behaviour. Each behaviour type can be expressed as a posture in a number of different ways, for example space filling can involve raising to full height or putting hands on hips while closeness can be expressed as leaning forward or making a more direct orientation (or some combination). Actual postures are calculated as weighted sums over a set of basic postures each of which depends on a behaviour type.

The basic postures were designed based on the description in Argyle (1975) and Mehrabian (1972) combined with informal observations of people in social situations. The weights of each basic posture are the product of the value of its behaviour type and its own weight relative to the behaviour type. The weights of the basic postures are varied every so often so that the avatar changes its posture without changing its meaning, thus producing a realistic variation of posture over time. This is done using a random group terms as shown in figure 3. Each basic posture is represented as an orientation for each joint of the avatar and final posture is calculated as a weighted sum of these orientations.

Gesture is generated using the same body animation system as postures, the main difference being that gestures are multi-frame animations and so weighted sums must be performed over a number of frames. They are also no longer merely static poses that can be held for a period of time; they must be repeated at appropriate intervals. More importantly gestures are more closely integrated with the flow of conversation and so must be synchronised with conversation as described below in section 5.2. Of course as the conversation is textual the synchronisation does not have to be as exact as it would be with spoke language. We also do not attempt to parse text so dgestures are not strongly connected to the meaning of the text as in Cassell et al. (2001b). Our gesture model serves only to indicate when some one is talking and to express a degree of attitude. Figure 4 shows examples of postures and gestures.

4.2 Eye gaze

Natural eye gaze is critical to the realism and believability of an animated character. This is because eye gaze is fundamental in showing interest levels between characters and as means of anticipating events. Typically a person will look to another before exhibiting any behaviour, such as moving towards them or speaking to them. In conversation, a listener will typically spend a large proportion of their time looking at the speaker. A complete lack of gaze towards the speaker is a clear message of the lack of interest of the audience towards the speaker and will be picked up very quickly. Conversely, mutual gaze, in which two people are looking into each others eyes is a powerful mechanism that induces arousal in the individuals, so typically mutual gaze is short (of the order of a second).

From an early age children learn to first look at the eyes of a person to determine the intention of that person towards them. Are they looking at them? Are they looking at some other person? Are they looking at something that might be a threat? By first looking at the eyes of another the intention, and therefore an appropriate response, can immediately be judged, see Baron-Cohen (2001).

Mutual gaze, in which two people are looking into each other's eyes is a powerful mechanism that induces arousal in the individuals, so typically mutual gaze is short (of the order of a second). Patterns of mutual gaze much longer than that either induces negative arousal, for example when someone stares aggressively, or positive arousal in an intimate setting. It is clear then that if a avatar does not look at an individual at all, it is seen as strange because it is an inbuilt primitive defence mechanism to look at other people (and their eyes) to determine the intent of that person (interest, disinterest, threat etc.). It is also strange if there is eye contact for too long due to the increased arousal this produces.

Argyle and Cook (1976) have done extensive studies with pairs of individuals to understand levels of eye gaze, and mutual gaze, and has detailed results covering (among other things) conversations and the level to which individuals will look at the other while speaking (35%) and listening (75%) etc. We have used these results to influence our model of gaze and mutual gaze in-group settings. Eye gaze is also related to attitude. Higher affiliation results in higher levels of eye gaze. Argyle and Cook have demonstrated compensatory behaviour for eye gaze. People react to higher levels of eye gaze by reacting with more distant postures, and conversely people will look at each other less if they are placed close together.

Existing simulations of eye gaze fall into two broad categories. Gillies and Dodgson (2002) and Chopra-Khullar and Badler (1999) simulate the eye gaze of characters navigating and performing actions in an environment but do not handle social factors of gaze between people. Our work is closer to the other type of simulation that deals primarily with social gaze. Garau et al. (2001) and Colburn et al. (2000) simulate the patterns of eye gaze between pairs of characters based on frequencies of mutual gaze. Vilhjálmsson and Cassell (1998) use eye gaze to help regulate the flow of conversation by indicating when a speaker is about to finish talking, when someone wants to start or end a conversation and other similar information. Rickel and Johnson (1999), in their character based virtual reality tutoring system, use gaze primarily as a method of indicating to the user an area of interest in the environment. Thórisson (1998) simulates eye gaze in the context of more general work on multi-modal communicative behaviour during conversation.

Our eye gaze model, as part of the Demeanour framework separates out the interests of the individuals that is the people or objects that currently demand that persons attention, from the low level details which trigger particular behaviour of the individual such as eye or head movement to look between individuals and at other objects. Each avatar has a set of foci of interest, which are objects that it will look at. The level of interest is specified as the proportion of time spent looking at that object. So for example if the avatar is in conversation with another avatar, while talking the level of gaze will be set to (say) 35%, and whilst listening to about 75% to approximate the natural gaze levels in conversation between two people. The natural animation of the avatars comes from the conflicting constraints of needing to look at things of interest for a certain proportion of the time, tempered by the (social) need to not stare!

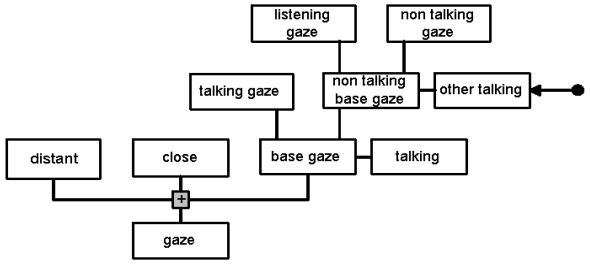


Figure 5: The section of behavioural controller dealing with eye gaze.

We have said that a avatar has a set of 'things' of interest at any one time, together with the proportion of time to look at that thing. There is also a maximum 'stare' time for each object, and, if that thing is a person, there will be a target for maximum mutual gaze. The animation framework for eye gaze continually monitors the gaze of each avatar and tracks how long has been spent gazing at a particular object, and the overall time gazing at that object. When these values reach the thresholds for an individual, it triggers the gaze control framework to take action to change the gaze of a avatar.

The threshold and mean gaze values are generated by the Demeanour framework. As described above levels of gaze are different depending whether a person is talking, listening to some one else talking or neither. These three conditions are distinguished based on the "talking" and "other talking" parameters shown in figure 5. The "talking" parameter is set when the avatar is speaking and the "other talking" parameter is an import parameter that allows access to the "talking" parameter of other characters. The behavioural controller has a separate base gaze value for each of the talking conditions, which are parameters of the framework as shown in figure 5. One of these values is chosen based on switch nodes based on the "talking" and "other talking" parameters. Thus the "non talking base gaze" is either equal to "listening gaze" or "non talking gaze" depending on whether the "other talking" input is true. Using two switch nodes ("non talking base gaze" and "base gaze" in the figure) we choose one of the gaze proportions depending on the talking condition.

However, this base value is also affected by the affiliation attitude between the avatar that is looking and the one that is being looked at. A close attitude increases proportion of gaze (up to a maximum of 100%) and distant behaviour reduces it (to a minimum of 0%). This scaling is achieved by combining the base gaze values with the "close" and "distant" terms using a Sum of Products term as shown in figure 5. The exact formula used to determine

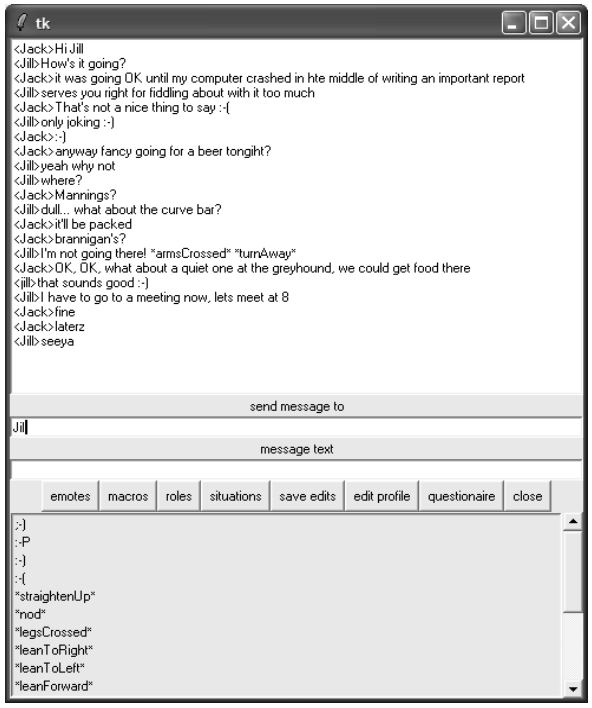


Figure 6: The text chat interface for Demeanour. (The 3D avatars are currently shown in a different window).

the actual eye gaze is:

$$g = g_{cond} - g_{cond}\frac{distant}{d_{max}} + (1 - g_{cond})\frac{close}{c_{max}}$$

where g is the proportion of time spent gazing at the target on average. g_{cond} is the gaze proportion due to the condition (talking, listening or neither). $distant$ and $close$ are the values for the close and distant attitudes and d_{max} and c_{max} are the values at which the gaze proportion is either 0 or 1. The output of the terms shown in figure 5 is the "gaze" term which is passed to the eye gaze module to actually generate gaze behviour using the calculated proportions.

In conversation between people a person will look at another then look away, usually by averting their gaze rather than moving their head, but they are not looking specifically at any other object, just averting their gaze. In our model we achieve this by having a number of 'halo' points around the head of a avatar that can be selected to look at if we need to look away, and have no other object that demands our attention.

5 Real time user control

The real time control interface to an avatar should be well integrated with the main user interface of the world. We have chosen to demonstrate Demeanour using a text chat interface that is very common in 3D multi-user virtual worlds and more generally on the internet. Users have a form of conversation consisting of short messages that are sent to the other users on pressing the return key. We have implemented such an interface, using the Jabber instant messaging system Adams (2001), the interface is shown in figure 6.

5.1 Direct control

All user control of the avatar's behaviour ultimately occurs through altering the values of the parameters of the behavioural controller, for example increasing the "friendliness" parameter will result in more affiliative behaviour. The obvious interface is therefore to allow users to directly control the values of the parameters. Demeanour uses a fairly traditional interface to do this, users are presented with a set of sliders that they can use to alter the values of a number of parameters. This kind of interface is common in many types of computer system and is a valuable way to give the user direct control of the avatar's internal state. However, we do not consider this to be sufficient for controlling an avatar in real time as it suffers from many of the problems discussed in section 1: the interface is not well integrated with text chat, users must constantly be moving from keyboard to mouse to control the behaviour. This required attention shift means that users are less likely to actually control the character effectively. The following sections describe interfaces that are more specifically designed for text chat systems.

5.2 Text chat based interface

The main interface directly uses the conventions of text chat. Textual communication on the internet already has its own vocabulary to express emotion and attitude, namely emoticons or smilies :-). In our example a smiling emoticon :-) will increase the "friendliness" parameter while a frownie :-(will reduce it. This provides a very natural interface that does not intrude on a conversation, and is already well understood by internet users.

Users can type in smilies and other commands as part of their messages using an identical interface. The user interface maintains a list of commands and each message sent is scanned to see if it contains any of them. The commands themselves can be arbitrary text strings and so can either be emoticons ;-) or words and phrases *bow*. We use the convention that textual commands are enclosed in asterisks to distinguish them from normal text. Figure 7 shows the result of a user typing a command (the later sections of this figure are described in the next section). If a command is detected it is passed to the behavioural controller. Each command has associated with it a number of actions which the controller executes when it receives the command. These actions can be of two types, firstly they can set the value of a particular parameter so, for example the smilie :-) can increase the value of the "friendliness" parameter. The second type of action is a direct request for a type of behaviour. For example, a particular posture

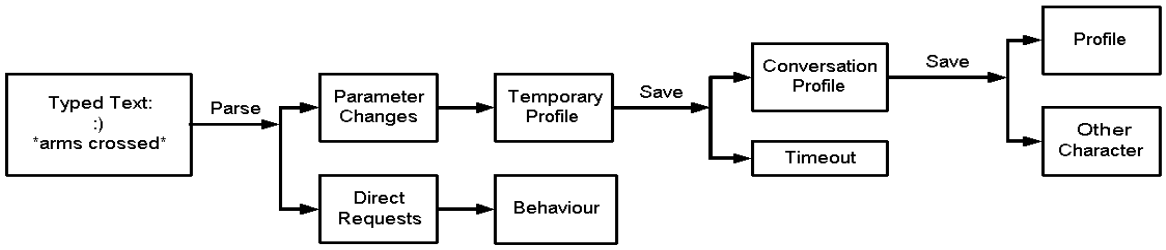


Figure 7: The sequence of steps by which typed commands can influence behaviour

or gesture such as *arms crossed*. This allows a more direct control over the body language that allows the user to create more exact effects. Internet conventions can also be used here, common abbreviations are often used to denote action, for example “lol” for “laugh out loud”. The two can be combined, if a posture such as crossed arms is known to be associated with low affiliation the command to request it can also reduce an associated parameter such as ‘friendliness’. This allows for a more implicit control of the state of the avatar. The user is able to control the affective state both directly and by demonstrating associated behaviour. This allows users to have a very precise affective control (both the attitude and its exact expression) and to be able to control an avatar without having to understand exactly how the internal parameters work. These three types of control, parameter, action and combined control, promise to provide a flexible, user friendly and unobtrusive user interface for a semi-autonomous avatar.

There are also two features that make it easier to use both frequent and rare commands. If a user frequently uses a command or sequence of commands (or even a piece of text) they can define a macro which allows them to reproduce the text with less typing. A macro is a backslash (‘\’) followed by a text sting. When it appears in a message it is replace with an arbitrary string defined by the user. This can greatly reduce the number of characters required to type a command. At the opposite end of the scale some commands are used rarely and so are difficult for users to remember. New users also need a simple way to find out the commands available. The Demeanour interface therefore also provides a menu for choosing commands (shown in the bottom panel in figure 6).

5.3 Temporal Scope

An important aspect of the Demeanour user interface is that commands do not merely result in instantaneous actions but in lasting changes of state of the avatar. For example, a direct request for a posture has a relatively short duration, however, an increase in friendliness may have a far longer temporal scope. Unfortunately it is not very clear from the context what this scope should be, a smilie :-) might in fact have a very short scope, just the length of the current utterance or it might indicate a permanent positive attitude to the person being talked to. This makes it difficult to choose a length for a change of state, and in these circumstances it is often better to leave the decision up to the user rather than risking a mistake. We therefore allow the user to choose between four scopes:

- A temporary change will last for a limited period, disappearing after a time out.
- Changes can also be made to last for the whole length of the current conversations
- They can be permanent changes to the attitude to the conversational partner
- Finally changes can be a permanent change to the avatar’s profile, allowing users to gradually update the behaviour of the avatar during use without extensive off-line customisation.

The user interface for controlling these changes is again aimed at being unintrusive. Figure 7 show the control flow which allows users to define different temporal scopes for their commands. Demeanour maintains a number of parameters that control the behaviour of avatars. These profiles contain values for subsets of the parameters of the controller. Thus when a parameter value is changed it can be saved to a profile. When a user types a command this is converted into a change to a parameter value as described in the previous section. This is initially stored in temporary profile. This temporary profile is deleted after a short period of time and all the edits it contains are deleted. Thus the default scope for edits is that they are temporary. However, when a temporary profile is present (i.e. after the user sends a command) a button appears in the text chat interface allowing the user to save the profile. If the user clicks this button the temporary profile is saved into a conversation profile, which has a longer scope lasting for the entire conversation.

At the end of the conversation the user can save the resulting edits to a permanent profile that is always used in future. This can be to the avatar’s main profile which controls its behaviour and is the chief method of customising an avatar. However, an avatar also maintains a profile for its interactions with a particular person, called a contact profile. Thus two people who are friends might each have

a higher value for the "friendliness" parameter in their contact profiles for each other than in their main profile. Users can save the conversation profile into a contact profile for their conversational partner.

The different temporal scopes allow commands to be used in a variety of different ways. They can allow users to adopt fleeting moods at an appropriate moment in a conversation and are therefore good real-time control methods. However, the permanent profiles allow them to be used simultaneously as a method of long term customisation. Rather than explicitly defining a profile for the avatar a user can gradually correct and shape the behaviour during real time interaction. This means that the user avoids the long and possibly tedious task of defining a profile. It also means that the customisation is situated in a coversation. It is likely to be easier for a user to know what behaviour is appropriate when actually engaged in a conversation than to think about it abstractly during an off-line customisation step. The first factor is likely to be particularly important when creating contact profiles. Users are unlikely to want to spend time create a new profile by hand for each new person they meet.

5.4 Flow of Conversation

Another features of our text chat interfaces is that non-verbal behaviour is used to display the flow of conversation. If we look at a group of people it is normally easy to tell which is talking from their non-verbal behaviour, even when out of earshot. Adding appropriate non-verbal behaviour to avatars aides and makes clearer the flow of conversation. Text chat does not normally have a clear concept of conversation as messages are sent instantaneously, rather than spoken over time. In Demeanour a single speaker is maintained at a given time. A person becomes the speaker when they send a message and remain the speaker until a time out is reached or some one else sends a message. When an avatar becomes the speaker the "talking" parameter is set in their behavioural controller (see figure 3), and their behaviour can be varied based on this parameter. Each avatar also has access to the other avatars' "talking" parameter via the "otherTalking" output. This allows behavioural control to determine which of three states it is in: the speaker, listening to a speaker or there is no-one talking. The controller can therefore produce different behaviour in each of the circumstances.

Results from Argyle and Cook (1976) show that eye gaze depends heavily on the flow of conversation. We have different values for the base proportion of time that an avatar looks at an other avatar depending on which if any of the two avatars is talking. Based on the results of Argyle and Cook (1976) we set looking while talking to be about 35%, looking at the speaker to be 70% and looking at a non-speaker to be 10%. These are base values for the proportion of gaze that are modified based on the affiliation of one avatar to another (see section 4.2 and figure 5).

Talking and listening also have an important effect on body motion. Generally gestures are closely linked to speech and only generally occur when talking. We therefore only activate gestures when the avatar is talking. This is done by a switch node that depends on the "talking" parameter. There are certain gestures, called back-channel gestures that occur when listening, and give feedback to the speaker. The most common is head nodding which shows approval or agreement, and which encourages the speaker. We therefore also have a nodding gesture which is activated based on the "other talking" parameter (as shown in figure 3).

6 Conclusion

We have described the Demeanour framework for generating non-verbal communicative behaviour and in particular described its user interface. This interface aims to integrate seemlessly with a text chat interface. We believe it will a simple user friendly way of controlling complex behaviour in an avatar and go some way to solving the problems of non-verbal communication in avatars. Of course it is not possible to really know this until we have conducted a user trial which we are currently planning.

Acknowledgements

We would like to thank BT Exact for sponsoring this research. We would also like to thank the UCL Department of Computer Science Virtual Environments and Computer Graphics group for their help and support, and Amanda Oldroyd for the use of her avatar models.

References

D. J. Adams. *Programming Jabber*. O'Reilly, Sebastapol, CA, USA, December 2001. ISBN 0-596-00202-5.

M. Argyle and M. Cook. *Gaze and Mutual Gaze*. Cambridge University Press, 1976.

M. Argyle. *Bodily Communication*. Routledge, 1975.

N. Badler, C. Philips, and B. Webber, editors. *Simulating Humans: Computer Graphics, Animation and Control*. Oxford University Press, 1993.

S. Baron-Cohen. *Mindblindness: An essay on autism and the theory of mind*. The MIT Press, 2001.

P. Bécheiraz and D. Thalmann. A model of nonverbal communication and interpersonal relationship between virtual actors. In *Proceedings of the Computer Animation '96*, pages 58–67. IEEE Computer Society Press, June 1996.

S. Benford, J. Bowers, L. E. Fahlén, C. Greenhalgh, and D. Snowdon. User embodiment in collaborative virtual

environments. In *SIGCHI conference on human factors in computing systems*, pages 242–249, 1995.

B. Blumberg and T. Galyean. Multi-level direction of autonomous creatures for real-time virtual environments. In *ACM SIGGRAPH*, pages 47–54, 1995.

J. Cassell, T. Bickmore, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. Embodiment in conversational interfaces: Rea. In *ACM SIGCHI*, pages 520–527. ACM Press, 1999.

J. Cassell, Y. Nakano, T. Bickmore, C. Sidner, and C. Rich. Non-verbal cues for discourse structure. In *41st Annual Meeting of the Association of Computational Linguistics*, pages 106–115, Toulouse, France, 2001a.

J. Cassell, H. H. Vilhjálmsson, and T. Bickmore. BEAT: the behavior expression animation toolkit. In *ACM SIGGRAPH*, pages 477–486, 2001b.

L. Cheng, S. Farnham, and L. Stone. Lessons learned: Building and deploying virtual environments. In Ralph Schroeder, editor, *The Social Life of Avatars, Presence and Interaction in Shared Virtual Worlds*, Computer Supported Cooperative work. Springer, 2002.

D. Chi, M. Costa, L. Zhao, and N. Badler. The emote model for effort and shape. In *ACM SIGGRAPH*, pages 173–182. ACM Press/Addison-Wesley Publishing Co., 2000.

S. Chopra-Khullar and N. Badler. Where to look? automating visual attending behaviors of virtual human characters. In *Autonomous Agents Conference*, 1999.

A. Colburn, M. Cohen, and S. Drucker. The role of eye gaze in avatar mediated conversational interfaces. Technical report, Microsoft Research, 2000.

M. Coulson. Expressing emotion through body movement: a component based approach. In Ruth Aylett and Lola Cañamero, editors, *AISB workshop on Animating Expressive Characters for Social Interactions*, Imperial College London, April 2002.

E. Cuddihy and D. Walters. Embodied interaction in social virtual environments. In *Proceedings of the third international conference on Collaborative virtual environments*, pages 181–188, September 2000.

M. Garau, M. Slater, S. Bee, and M. A. Sasse. The impact of eye gaze on communication using humaniod avatars. In *ACM SIGCHI*, pages 309–316, 2001.

M. Gillies and D. Ballin. A model of interpersonal attitude and posture generation. In Thomas Rist, Ruth Aylett, Daniel Ballin, and Jeff Rickel, editors, *Fourth Workshop on Intelligent Virtual Agents*, Kloster Irsee, Germany, September 2003.

M. Gillies and N. Dodgson. Eye movements and attention for behavioural animation. *Journal of Visualization and Computer Animation*, 13:287–300, 2002.

A. Mehrabian, editor. *Nonverbal Communication*. Aldine-Atherton, 1972.

A. Paiva, I. Machado, and R. Prada. The child behind the character. *IEEE Transactions on systems, man and cybernetics: Part A: Systems and Humans*, 31(5):361–368, Sept 2001.

C. Pelachaud and I. Poggi. Subtleties of facial expressions in embodied agents. *Journal of Visualization and Computer Animation.*, 13:287–300, 2002.

Ken P. and A. Goldberg. Improv: A system for scripting interactive actors in virtual worlds. In *Proceedings of SIGGRAPH 96*, Computer Graphics Proceedings, Annual Conference Series, pages 205–216, New Orleans, Louisiana, August 1996. ACM SIGGRAPH / Addison Wesley.

J. Rickel and W. L. Johnson. Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied Artificial Intelligence*, 13: 343–382, 1999.

P. Scerri, J. Ydrén, and N. E. Reed. Layered specification of intelligent agents. In *PRICAI 2000*, August 2000.

P. Sengers, S. Perry, and J. Smith. Traces: Semi-autonomous avatars. current available at http://www-2.cs.cmu.edu/ phoebe/work/publications.html, 2000.

K. Thórisson. Real-time decision making in multimodal face-to-face communication. In *second ACM international conference on autonomous agents*, pages 16–23, 1998.

J. Tromp and D. Snowdon. Virtual body language: Providing appropriate user interfaces in collaborative virtual environments. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST),*, pages 37–44, 1997.

X. Tu and D. Terzopoulos. Artificial fishes: Physics, locomotion, perception, behavior. In *ACM SIGGRAPH*, pages 43–49, 1994.

H. H. Vilhjálmsson and J. Cassell. Bodychat: Autonomous communicative behaviors in avatars. In *second ACM international conference on autonomous agents*, 1998.

M. Wilson and M. Neal. Diminishing returns of engineering effort in telerobotic systems. *IEEE Transactions on systems, man and cybernetics: Part A: Systems and Humans*, 31(5):459–465, Sept 2001.

Speaking with Emotions

E. Bevacqua

Department of Computer and System Science
University of Rome "La Sapienza"
elisabetta.bevacqua@libero.it

M. Mancini and C. Pelachaud

LINC - Paragraphe
IUT of Montreuil - University of Paris 8
m.mancini@iut.univ-paris8.fr
c.pelachaud@iut.univ-paris8.fr

Abstract

We aim at the realization of an Embodied Conversational Agent able to interact naturally and emotionally with user(s). In previous work [23], we have elaborated a model that computes the nonverbal behaviors associated to a given set of communicative functions. Specifying for a given emotion, its corresponding facial expression will not produce the sensation of expressivity. To do so, one needs to specify parameters such as intensity, tension, movement property. Moreover, emotion affects also lip shapes during speech. Simply adding the facial expression of emotion to the lip shape does not produce lip readable movement. In this paper we present a model that adds expressivity to the animation of an agent at the level of facial expression as well as of the lip shapes.

1. Introduction

With the development of 3D graphics, it is now possible to create Embodied Agents that can talk simulating that kind of communication that people know since they are born. Moreover, nonverbal communication is as important as verbal one. Facial expressions can provide a lot of information. In particular they are good window on our emotional state [9]. Emotions are a fundamental aspect of human life influencing how we think and behave and how we interact with others. Facial expressions do improve communication [13]; they can make clear what we are thinking, even without speaking. For example wrinkling our nose in front of something that we dislike communicates very clearly our disgust. Therefore, believable synthetic characters make the interaction between users and machine easier and more fulfilling, providing a more human-like communication. Experiments have shown that interface with facial displays reduces the mental gap between users and computer systems [25].

Most of the ECA systems developed so far have been concentrated in defining the appropriate nonverbal behavior linked to speech. But nonverbal behaviors are characterized not only by the signals of the expression itself but also by temporal parameters (e.g. time of appearance and disappearance of an expression) and by muscular activities quality (such as tense movement). In the aim of creating believable embodied conversational agent (ECA) able to serve as bridge in the communication between humans and the machine, ECA ought to be empowered with human-like qualities. In this paper we present a computational model of expressivity for facial expression and for the lip movements. The work presented in this paper is part of the Greta system that have been developed within the EU project MagiCster [1] . The system takes as input a text tagged with communicative functions information. The language of tags is called Affective Presentation Markup Language, APML [6]. APML is used as a script language to control the agent's animation. To endow agent with expressivity quality we have extended APML.

In the next section we prevent an overview of the state of the art. We then present our extension of APML. We follow by describing the lip shape model.

2. State of the art

There is not a single way to approach the issue of emotions in ECAs. S. Kshirsagar and her colleagues have developed a 3D virtual characters with emotions and personality [8]. The agent can maintain a basic dialogue with users. Through a personality and emotion simulators the agent responds naturally to, both, the speech and the facial expressions of the user that are tracked in real time.

M. Seif El-Nasr, J. Yen and T. Ioerger [11] implemented a new computational model of emotions that uses a fuzzy-logic representation to map events and observations to emotional states. They based their work on the research on emotions that shows that memory and experience have a major influence on the emotional process.

A. Paiva and her research team approached the problem of modelling emotional states of synthetic agents

[1] IST project IST-1999-29078, partners: University of Edinburgh, Division of Informatics; DFKI, Intelligent User Interfaces Department; Swedish Institute of Computer Science; University of Bari, Dipartimento di Informatica; University of Rome, Dipartimento di Informatica e Sistemistica; AvartarME.

when implementing a computer game (*FantasyA*) where two opponents fight each other [14]. The battle is played in a virtual arena by two characters, one controlled by the user (through the doll *SenToy*), and the other by the system. They made their own decisions but are influenced by the emotional state induced by the user. So the agent's actions depends on its emotions, on its opponent's emotions, and on its personality. Carmen's Bright IDEAS [19] is an interactive drama where characters exhibit gestures based on their emotional states and personality traits. Through a feedback mechanism a gesture made by of a character may modulate its affective state. A model of coping behaviors has been developed by Marsella and Gratch [20]. The authors propose a model that embeds information such as the personality of the agent and its social role.

All those works aim to define models being able to decide what kind of emotion synthetic agents should "feel" in a particular situation.

Another area of the research on emotions in virtual characters is concerned with the expressivity of emotions.

M. Byun and N. I. Badler [3] proposed a method for facial animation synthesis varying preexistent expressions by setting a small number of high level parameters (defined based on the Laban notation [15] that drive low level facial animation parameters (FAPs from MPEG-4 standard [7]). The system is called FacEMOTE [3] and derives from EMOTE [4] which has been originally developed for arm gestures and postures.

Very few computational models of lip shape take into account emotion. M. M. Cohen and D. W. Massaro developed a coarticulation model [5], based on Löfqvist's gestural theory of speech production [16]. To model lip shapes during emotional speech, they add the corresponding lip shapes of emotion to viseme definition [21].

Our work is somewhat related to facEMOTE but we do not modified directly the animation stream; rather we create a new FAP stream to animate our 3D agent from an input text where tags specifying communicative functions are embedded [24, 6]. Our approach differs from other ones as we have added qualifiers parameters to simulate expressivity in facial expressions as well as in lip movements. Moreover we drive our computational model of emotional lip shapes from real data [17].

3. Temporal characteristics of facial expressions

A facial expression is not only identified by a configuration of facial muscles but is also important how this configuration will be temporally activated.
For example we can consider, starting from an initial neutral state, the time (or, identically, the speed) needed for the facial muscles to contract and reach the final state corresponding to the final expression; we call it *onset* time. Similarly, the corresponding time needed for the muscles to change from their current state back to the rest position is called *offset* time.
So a single expression (in the sense of "muscular configuration") can assume different expressivity depending on the manner it appears (onset), the time it remains on the face (called the *apex* time) and finally the speed it disappears (offset). We have chosen this specification to represent the temporal characteristics of facial expressions [22]. One example of this representation is shown in Figure 1. The time is on the *x* axis while on the *y* axis there is the intensity of the expression; where 0 means "muscles in rest position" and 1 represents "muscles in final position".

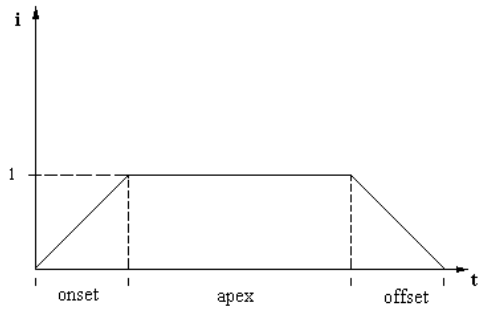


Figure 1: Temporal course of a facial expression.

Although these three parameters are extremely important in delivering expressivity of an expression, few studies assigning values to them exist; but vision systems could be used to extract them [12, 28].
It has been shown that, for example, expressions of sadness usually disappear slowly from the face. Thus, they have a long offset time [9]. On the other hand, expressions of joy appear very rapidly; this is characterized by a short onset time. Sometimes expressions may also differ in duration (apex time). Finally, fake expressions generally appear either too late or too early (e.g., polite smile versus real happiness smile); thqt is the value of delay varies.

4. Intensity of facial expressions

Facial movements corresponding to an expression can be produced with different intensity. An eyebrow can raise a little or a lot. A mild happiness will produce a small smile, while a great happiness will be shown by a large smile. Besides, P. Ekman found that if the emotion is felt very lightly, not every facial movement corresponding to the emotion will be visibly displayed; in the sense that changes expressed in the face may not be perceiv-

able [10]. For example, in the case of mild fear like apprehension, only slight expression around the mouth may be shown; while for extreme fear both areas, the muscles around the eyes and the mouth, are very tense [9].
Thus, the intensity of an emotion controls not only the amount of movements (strong or light) but also the appearance of some movements.

5. Giving expressivity to facial expressions

Until now we have been concentrating in elaborating computational models that define the most appropriate non verbal behaviors from an input text. The nonverbal behaviors as specified within APML correspond to frozen facial expressions; a facial expression is linked to an intensity; its temporal course is given by the XML span... Aiming at adding life characteristics to the agent, we have realized some modifications to the APML language in order to allow the Greta agent to communicate a wider variety of facial expressions of emotion as well as to allow for a more flexible definition of facial expressions and their corresponding parameters. Expressivity may be expressed differently depending on the considered modality: a face has different variables (timing variations, muscular intensity, as we have seen in sections 3 and 4) compared to gaze (length of mutual gaze, ratio of gaze aversion and looking at, ...). These modifications refer mainly to facial expressions timings as well as to their intensity.
An APML tag defines the meaning of a given communicative act [23]; the Greta engine looks up in a library of expression to instantiate this meaning into the corresponding facial expression. A facial expression has 3 temporal parameters as defined in Section 3: onset, offset and apex. In the previous version of the Greta engine, the value of the onset and of the offset were set as constants. An expression was set to start at the beginning of the tag and to finish at its end. That is the apex of an expression was set to be the total time length of a tag (computed as the duration of the speech embedded in the tag; this duration being provided by the speech synthesizer) minus the onset and offset times. APML provides a scheme to specify the mapping between meaning and signals for a given communicative act. We have extended APML to allow one to alter the expressivity of a communicative act. We have introduced a new set of 5 attributes that act both on expressions timings and intensities.
For each APML tag it is possible to specify one or more of the following attributes:

Delay : specifies the percentage of delay before an expression arises; it forces the Greta engine to delay the start of an expression for a certain time. This time is specified by a percentage of the total default animation time (that is the time of the speech embedded in the XML tag). If not specified, the default delay value is 0% (that is there is "no delay"); this corresponds to the previous version of APML.

Duration : specifies the total duration of an expression. It is specified as a percentage of the default expression duration. The Greta engine will set the duration of an expression (the apex of an expression) to last for this new value. The default value is 100% (that is "normal duration"); this corresponds to the previous version of APML.

Onset : specifies a value for the onset. This value is given as a number of animation frames that the engine have to use to render the "onset" phase of an expression. The default value is 0 and this tells to the engine to set the onset value to constant as defined in the previous version of APML and explained before.

Offset : specifies a value for the offset. This value is given as a number of animation frames that the engine have to use to render the "offset" phase of an expression. The default value is 0 and this tells to the engine to set the onset value to constant as defined in the previous version of APML and explained before.

Intensity : corresponds to a factor that multiplies the quantity of movement of each FAP involved in the facial expression. Until now the corresponding facial expression to a meaning for a given communicative act was explicitly defined. It was not possible to modify on the fly such a mapping. In order to have a facial expression with lower or greater intensity, one had to create a new entry in the library of expressions. This was quite cumbersome. To remedy this lack of flexibility, we introduce an intensity factor that can modify automatically any defined expression. Our facial model is MPEG-4 compliant [7]. An expression is defined by a set of FAPs (Facial Animation Parameter). The variation of their intensity corresponds to the modification of the value of each FAP. The default intensity value is 1 meaning that values of the FAPs defining an expression in the library of expressions are not changed.

Let us consider the following example:

```
<theme belief-relation="gen-spec"
    affect="anger">
some text </theme>
```

The timings of the expression as evaluated in the tag are given in the Figure 2.

Let us consider now the same example with the introduction of the new tags. For example we may have:

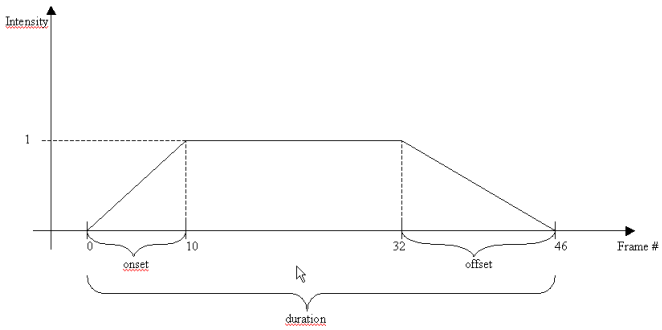


Figure 2: Temporal course of a facial expression.

```
<theme belief-relation="gen-spec"
    affect="anger" delay="40%"
    duration="40%" onset="4"
    offset="4" intensity="1.5" >
some text </theme>
```

The type of expression evaluated in the tag remains unchanged; what change are the temporal and intensity parameters of the expression. Figure 3 illustrates these changes. Figure 4 shows the resulting expressions as they can be seen on the agent's face.

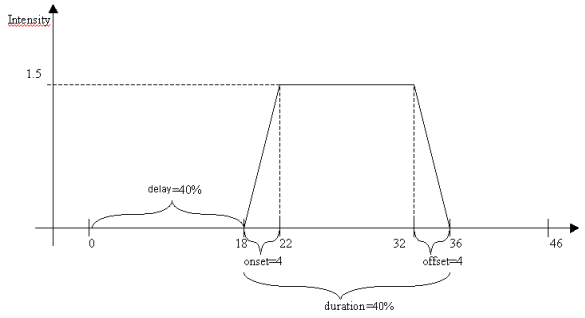
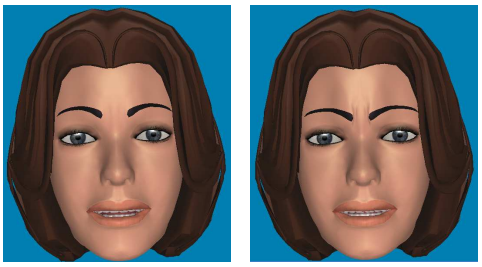


Figure 3: Temporal course of a facial expression when taking into account the new modifiers of APML.

Another example of how intensity can modify the appearance of an expression is given in Figure 5, in which an expression of joy is computed with an intensity of 1 in the first image and an intensity of 1.5 in the second.

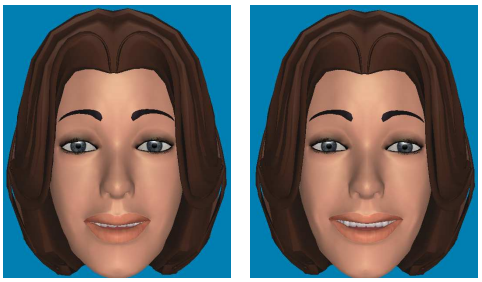
6. Computational lip model

Our lip model is based on captured data of triphones of the type $'VCV$ where the first vowel is stressed whereas the second is unstressed. The data has been collected with the optical-electronic system ELITE that applies passive markers on the speaker's face [17]. The data covers not only several vowels and consonants for the neutral expression but also different emotions, namely joy, anger, sadness, surprise, disgust and fear [17]. The original



(a) intensity=1 (b) intensity=1.5

Figure 4: Anger expression: in figure (b) the frown is more intense, the lips are more tense and the teeth are clenched.



(a) intensity=1 (b) intensity=1.5

Figure 5: Joy expression: in figure (b) the cheeks are more raised, the lips are more widely open and the smile is larger.

curves from the real speech data represent the variation over time of the 7 phonetically and phonologically relevant parameters that define lip shapes: upper lip height (ULH), lower lip height (LLH), lip width (LW), upper lip protrusion (UP), lower lip protrusion (LP), jaw (JAW) and lip corners (LC). On such curves we have selected the maximum or the minimum (target point) to characterize each viseme. To get a good representation of the characteristics of a vowel and of the breadth of its original curve, we choose two more points; one between the onset of the phoneme and its target and the other between its target and the offset of the phoneme. Instead, consonants are well represented considering just the target point. Since consonants, and at a slightest degree the vowels, are influenced by the context, we collect their targets from every $'VCV$ contexts (for instance, for the consonant /m/, the targets points are extracted from the contexts $/'ama/$, $/'eme/$, $/'imi/$, $/'omo/$ and $/'umu/$).

Targets data of vowels and consonants have been stored in a database. Besides targets values, other infor-

mation have been collected like the vowel or the consonant that defines the surrounding context, the duration of the phoneme and the time of the targets in this interval. A similar database for each of the six fundamental emotions and for the neutral expression have been created.

6.1. Lip shape algorithm

We have developed an algorithm that determines for each viseme the appropriate values of the 7 labial parameters by applying coarticulation and correlation rules in order to consider the vocalic and the consonantal contexts as well as muscular phenomena such as lip compression and lip stretching [2].

Our system works as follow. It takes as input a text (tagged with APML) an agent should say. The text file is decomposed by Festival [26] into a list of phonemes with their duration. The first step of our algorithm consists in defining fundamental values, called *target points*, for every parameter of each viseme associated with the vowels and consonants that appear in this phoneme list. A target point corresponds to the target position the lips would assume at the apex of the viseme (which may not always correspond to the apex of the phoneme production [17]). These targets have been collected analyzing the real data described above and are stored in a database, one for each emotion.

Afterwards, the algorithm modifies the targets according to the emotion in which the phonemes are uttered, to the coarticulation and correlation rules and to the speech rate.

Finally, the lip animation is computed on those targets.

	ULH	LLH	JAW	LW	UP	LP	CR
Neutral	0	0	0	0.2	0.1	0.1	0.2
Joy	1	1	1	0.8	0.9	0.9	0.8
Surprise	0	0	0	0	0	0	0
Fear	0	0	0	0	0	0	0
Anger	0	0	0	0	0	0	0
Disgust	0	0	0	0	0	0	0
Sadness	0	0	0	0	0	0	0

Table 1: Matrix of the emotion **Mild-Joy**.

7. Emotion model

We describe each emotion through a 7x7 matrix. The rows correspond to the seven recorded emotions, whereas the columns are the lip parameters. A value in the matrix represents the percentage of dependence that the corresponding lip parameter has on the corresponding emotion. Therefore, the value of the targets for each labial parameter will be an interpolation among the targets in the emotions that have a value on the column different from zero. Obviously the seven emotions have all 1 in the row corresponding to the emotion itself.

Let us consider the consonant /b/ in the triphone $/'aba/$ uttered with the emotion 'mild-joy'. The matrix of this emotion is shown in Table 1.
Now, let $N_a b_a$ be the target of the lip width parameter of /b/ uttered in a neutral emotion and let $J_a b_a$ be the target of the same lip parameter of /b/ uttered in the 'joy' emotion. The new value of this target $MJ_a b_a$ in the mild-joy emotion will be:

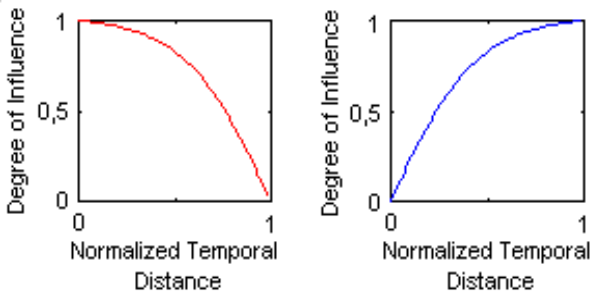
$$MJ_a b_a = 0.2 * N_a b_a + 0.8 * J_a b_a$$

As consequences, the lip width will be less wide in the 'mild-joy' emotion than in the 'joy' emotion.

7.1. Expressivity qualifiers

We have also defined two qualifiers to modulate the expressivity of a lip movement. The first one, *Tension Degree*, can be *Strong*, *Normal* and *Light*. It allows one to set different intensities of muscular strain. Such a tension may appear for the expressions of emotions like fear and anger. Such a tension can also appear, for example, when a bilabial consonant (as /b/) is uttered and lips compress against each other, or when labial width increases and lips stretch them getting thinner. The second qualifier, *Articulation Degree*, can take the values *Hyper*, *Medium* and *Hypo*. During hypo articulation, it may happen that lip targets may not reach their apex.

LOGISTIC FUNCTION



(a) Anticipatory coarticulation.

(b) Carry-Over coarticulation.

Figure 6: Logistic Function - Strong Degree of influence.

8. Coarticulation and correlation rules

Once all the necessary visemes have been loaded from the database and modified according to the emotions and

the expressiveness qualifiers, coarticulation and correlation rules are applied. In fact, to be able to represent visemes associated to vowels and consonants, we need to consider the context surrounding them [2]. Firstly, let us consider consonants. Since vowels are linked by a hierarchical relation for their degree of influence over consonants ($u > o > i > e > a$) [18, 27], we first determine which vowels will affect the consonants in $V_1C_1 \ldots C_nV_2$ sequence and which labial parameters will be modified by such an influence. Vowels act mainly over those lip parameters that characterize them. The new targets of the consonants for each lip parameter are computed through a linear interpolation between the consonantal targets ${}_{V_1}C_{iV_1}$ in the context deriving from the vowel V_1 and the consonantal targets ${}_{V_2}C_{iV_2}$ in the context of the vowel V_2:

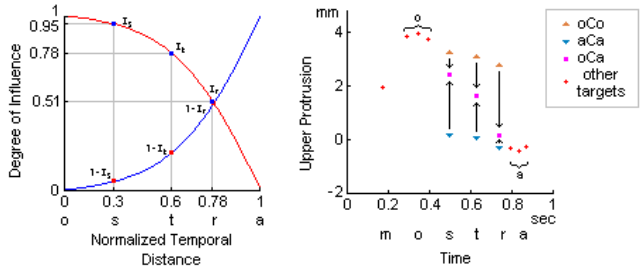
$$ {}_{V_1}C_{iV_2} = k * {}_{V_1}C_{iV_1} + (1-k) * {}_{V_2}C_{iV_2} \quad \textbf{(1)} $$

The interpolation coefficient k is determined through a mathematical function, called *logistic function*, whose analytic equation is:

$$ f(t) = \frac{a}{1+e^{-bt}} + c $$

This function represents the influence of a vowel over adjacent consonants, on the labial parameters that characterize it, and allows one to obtain carry-over coarticulation and anticipatory coarticulation (see Figure 7.1). The constants a and c force the function to be defined between 0 and 1 both on the abscissa and on the ordinate simplifying the computation. The constant b defines the slope of the curve that represents different degrees of influence. Time t=0 corresponds to the occurrence of V_1, and time t=1 corresponds to V_2. The consonants C_i are placed on the abscissa depending on their temporal normalized distance from the vowels.

INF E MOSTRA



(a) influence of /'o/ on /s/, /t/ and /r/.

(b) Alteration of UP targets for the Italian word /mostra/.

Figure 7: Coarticulation effects on consonants.

For example, let us consider the sequence /'ostra/ taken from the Italian word 'mostra' ('exhibition') and the lip parameter UP. To obtain the curve representing the upper protrusion, the consonantal targets of /s/, /t/ and /r/ in the context oCa must be calculated. The vowel /'o/ exerts a strong influence over the following three consonants and the algorithm chooses the steepest influence function. Figure 7(a) shows how the logistic function is applied to define the interpolation coefficients and Figure 7(b) shows how the targets are modified through equation (1).

Once all the necessary visemes have been calculated, correlation rules are applied modifying the value of the targets to simulate muscular tension. For example, when a bilabial consonant (as /b/) is uttered, lip compression must appear. Thus when a strong lip closing occurs the FAPs on the external boundary of the lips must be further lowered down.

Vocalic targets are modified in a very similar way, according to the consonantal context in which appear. Consonants are grouped on the base of which labial parameters they influence. For example, /b/, /m/ and /p/ will affect vowels on ULH and LLH parameters.

Finally, lip movement over time is obtained interpolating the computed visemes through Hermite Interpolation.

8.1. Speech rate

Speech rate strongly influences labial animation. At a fast speech rate, lip movement amplitude is reduced while at a slow rate it is well pronounced (lip height is wider when an open vowel occurs or lip compression is stronger when a bilabial consonant is uttered).

To simulate this effect, at a fast speech rate the value of targets points is diminished to be closer to the rest position; while at a slow speech rate, lips fully reach their targets.

9. Evaluation tests

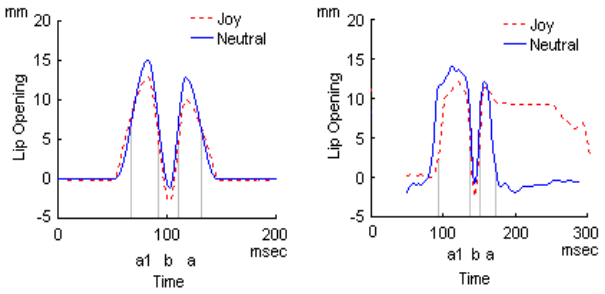
As evaluation tests, we compare the original curves with those computed by our algorithm. As first example let us consider the triphone /'aba/ uttered either in joy and in neutral expression. Figure 8 shows the curves that represent Lip Opening. We have differentiated the movement of the lower lip from the movement of the upper lip (to get a finer movement description). So the Lip Opening curves are obtained as a sum of ULH and LLH curves. The generated curves are shown in Figure 8(a) whereas the original ones in Figure 8(b). In both figures the red dotted line represents the lip movement in joy emotion, while the blue solid line describes the lip opening in neutral expression. As one can see the joy emotion causes a reduction of the lip opening (8(b)). The same behavior is also shown in generated curves (see Figure 8(a)).

The second example is quite similar to the first one but the word uttered is /mamma/ and the emotions considered are disgust and neutral. Lip Opening curves are shown in Figure 9. The red dotted line represents the lip movement in disgust emotion and the blue solid one is the lip opening in neutral expression. Figure 9(a) shows the computed curves whereas Figure 9(b) the original ones. Like joy, disgust emotion causes a diminution of lip opening. Both, the original and the computed curves display the same behavior.

In both examples, phonemes segmentation is identified by vertical lines.

For the generated curves, times are given by Festival, while for the original ones times come from the analysis of real speech. Moreover the generated curves always start at the neutral position. This is not necessarily true in the original data. Those differences can make the calculated curves slightly different from the original ones.

ABA



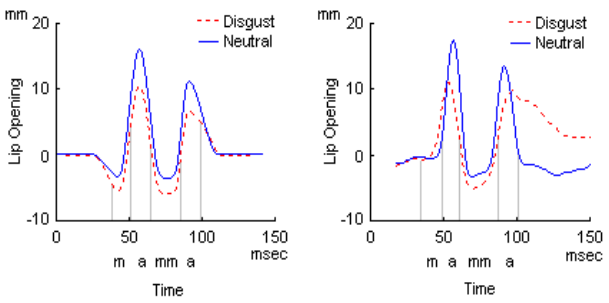
(a) Generated curves. (b) Original curves.

Figure 8: Lip Opening for the triphone /'aba/ for the joy and neutral expression.

10. Conclusion and Future Development

We have presented a model that adds expressivity to the animation of an agent. Expressivity is related not only to the specification of a facial expression but also how this expression is modulated through time and depending on the context. In this paper we have presented parameters to characterize expressivity along the face modality. We have also described a computation model of emotional lip movement. Our next step is to perform some perceptual evaluation tests to further check the feasibility of our models. Movies illustrating our method may be seen at the URL: http://www.iut.univ-paris8.fr/˜pelachaud/AISB04.

MAMMA



(a) Generated curves. (b) Original curves.

Figure 9: Lip Opening for the Italian word /mamma/ for the disgust and neutral expression.

11. References

[1] Benguerel, A. P. and Cowan, H. A., "Coarticulation of upper lip protrusion in french," *Phonetica*, vol. 30, pp. 40–51, 1974.

[2] E. Bevacqua and C. Pelachaud. Modelling an italian talking head. In *Auditory-Visual Speech Processing AVSP'03*, Saint-Jorioz, France, 2003.

[3] M. Byun and N. Badler, "FacEMOTE: Qualitative parametric modifiers for facial animations," *Symposium on Computer Animation*, San Antonio, TX, July 2002.

[4] D.M. Chi, M. Costa, L. Zhao, and N.I. Badler. The EMOTE model for effort and shape. In Kurt Akeley, editor, *Siggraph 2000, Computer Graphics Proceedings*, pages 173–182. ACM Press / ACM SIGGRAPH / Addison Wesley Longman, 2000.

[5] M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. In M. Magnenat-Thalmann and D. Thalmann, editors, *Models and Techniques in Computer Animation*, pages 139–156, Tokyo, 1993. Springer-Verlag.

[6] B. DeCarolis, C. Pelachaud, I. Poggi, and M. Steedman. APML, a mark-up language for believable behavior generation. In H. Prendinger and M. Ishizuka, editors, *Life-like Characters. Tools, Affective Functions and Applications*, pages 65–85. Springer, 2004.

[7] P. Doenges, T.K Capin, F. Lavagetto, J. Ostermann, I.S. Pandzic, and E. Petajan. MPEG-4: Audio/video

and synthetic graphics/audio for real-time, interactive media delivery, signal processing. *Image Communications Journal*, 9(4):433–463, 1997.

[8] A. Egges, S. Kshirsagar and N. Magnenat-Thalmann, "A Model for Personality and Emotion Simulation," *KES 2003*: 453-461.

[9] P. Ekman and W. Friesen. *Unmasking the Face: A guide to recognizing emotions from facial clues*. Prentice-Hall, Inc., 1975.

[10] P. Ekman and W. Friesen. Felt, false, miserable smiles. *Journal of Nonverbal Behavior*, 6(4):238–251, 1982.

[11] M. S. El-Nasr, J. Yen and T. Loerger, "FLAME - Fuzzy Logic Adaptive Model of Emotions," *International Journal of Autonomous Agents and Multi-Agent Systems*, 3(3):1-39.

[12] I. A. Essa. *Analysis, Interpretation, and Synthesis of Facial Expressions*. PhD thesis, MIT, Media Laboratory, Cambridge, MA, 1994.

[13] A. J. Fridlund and A. N. Gilbert, "Emotions and facial expression," *Science*, 230 (1985), pp. 607-608.

[14] K. Höök, A. Bullock, A. Paiva, M. Vala and R. Prada, " FantasyA - The Duel of Emotions," in *Proceedings of the 4th International Working Conference on Intelligent Virtual Agents - IVA 2003*.

[15] R. Laban and F.C. Lawrence. *Effort: Economy in body movement*. Plays, Inc, Boston, 1974.

[16] A. Löfqvist's, "Speech as audible gestures," *Speech Production and Speech Modeling*, pp. 289–322, 1990.

[17] E. Magno-Caldognetto, C. Zmarich, and P. Cosi. Coproduction of speech and emotion. In *ISCA Tutorial and Research Workshop on Audio Visual Speech Processing, AVSP'03*, St Jorioz, France, September 4th-7th 2003.

[18] E. Magno-Caldognetto, C. Zmarich and P. Cosi, "Statistical definition of visual information for Italian vowels and consonants," in *International Conference on Auditory-Visual Speech Processing AVSP'98*, D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson, Eds., Terrigal, Australia, 1998, pp. 135–140.

[19] S. Marsella, W.L. Johnson, and K. LaBore. Interactive pedagogical drama. In *Proceedings of the 4th International Conference on Autonomous Agents*, Barcelona, Spain, June 2000.

[20] S. Marsella and J. Gratch. Modeling coping behavior in virtual humans: Don't worry, be happy. In *proceedings of the /2nd International Conference on Autonomous Agents and Multiagent Systems*, Melbourne, Australia, 2003.

[21] D. Massaro. *Perceiving Talking Faces : From Speech Perception to a Behavioral Principle*. Bradford Books Series in Cognitive Psychology. MIT Press, 1997.

[22] C. Pelachaud, N.I. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(1):1–46, January-March 1996.

[23] C. Pelachaud, V. Carofiglio, B. de Carolis, F. de Rosis and I. Poggi, "Embodied Contextual Agent in Information Delivering Agent," in *Proceedings of AAMAS*, 2002, vol. 2.

[24] I. Poggi. Mind markers. In C.Mueller and R.Posner, editors, *The Semantics and Pragmatics of Everyday Gestures*. Berlin Verlag Arno Spitz, Berlin, 2001.

[25] A. Takeuchi and K. Nagao, "Communicative facial displays as a new conversational modality," *ACM/IFIP INTERCHI '93*, Amsterdam, 1993.

[26] P. Taylor, A. Black, and R. Caley. The architecture of the Festival Speech Synthesis System. In *Proceedings of the Third ESCA Workshop on Speech Synthesis*, pages 147–151, 1998.

[27] E.F. Walther, *Lipreading*, Nelson-Hall, Chicago, 1982.

[28] Y. Yacoob and L. Davis. *Computer Vision and Pattern Recognition Conference*, chapter Computing spatio-temporal representations of human faces, pages 70–75.

Reusing Motion Data to Animate Visual Speech

James D. Edge Manuel A. Sánchez Lorenzo Steve Maddock

*University of Sheffield
Regent Court
211 Portobello st
Sheffield S1 4DP
j.edge@dcs.shef.ac.uk
m.sanchez@dcs.shef.ac.uk
s.maddock@dcs.shef.ac.uk

Abstract

In this paper we describe a technique for animating visual speech by concatenating small fragments of speech movements. The technique is analogous to the most effective audio synthesis techniques which form utterances by blending small fragments of speech waveforms. Motion and audio data is initially captured to cover the target synthesis domain; this data is subsequently phonetically labelled and segmented to provide basis units for synthesis. Sentence, word and syllable level units are used by the system to synthesize novel speech utterances. The final synthesized utterances consist of the motion of points on the surface of the skin, these trajectories are retargetted and interpolated across the surface of a target mesh for animation.

1 Introduction

Audio and visual stimuli are both involved in speech communication as a part of natural discourse. Not only does this involve emotional information, such as smiling or scowling, but also the movement of the lips, which is an important cue with regards the disambiguation of meaning. What we see and hear during speech gives complimentary information, which is backed up by perceptual studies that report as much as a +15dB improvement in signal-to-noise ratio [Sumby and Pollack (1954)] and a corresponding increase in the intelligibility of speech with visual information. This, along with interest in talking heads as a part of a more natural human-computer interface, has provoked a great deal of interest in synthesizing visual speech.

Visual speech synthesis is a sub-field within general modelling and animation of facial expression. In this context expression is the grouping of gestural (e.g. conversational signals), emotional (e.g. scowling) and physical (e.g. blinking) actions required to communicate between individuals. Speech is inherently a physical process of shaping the vocal tract such that a meaningful sequence of sounds is created, according to the words and grammar of a particular language. This is particularly important because the physical relationship between the visual and audio modalities necessitates that facial movements we observe are 'correct'; where this is not the case perceptual difficulties may arise [McGurk and MacDonald (1976)] or at least there will be an impediment to the realism of the animation.

In this paper we discuss the synthesis of visual speech by concatenating short fragments from a library of motion-captured data. This idea is the analogue of the concatenative techniques used commonly in audio speech synthesis, allowing us to conceptually unify the models which deal with the audio and the visual streams. It is our assertion that speech animation in this manner is more natural and realistic than current popular techniques based upon the interpolation of visual phonemes.

2 Background

Much research into facial animation considers the difficulties in modelling static facial expression [Parke (1974); Waters (1987); Lee et al. (1995)]. Here we review the more specialized field of speech animation, for a detailed overview of the entire field see [Parke and Waters (1996)]. For a detailed discussion of visual speech synthesis and relating perceptual issues see [Massaro (1998)].

The animation of visible speech movements has lagged behind the corresponding techniques in audio speech synthesis. Much research in the topic of speech animation relies upon the simple interpolation between elementary speech units, often referred to as visual-phonemes or visemes. The problem in animating speech lies in the synchronicity of the visual movements with the audio and the naturalness of those movements. The naturalness of speech movements are judged against the experiences of the audience in real life, making the task much more difficult to solve than many areas in the field of animation.

One of the major difficulties in animating speech is the physical phenomenon called coarticulation [Öhman (1967); Löfqvist (1990)]. This term refers to the obscuration of boundaries between neighbouring atomic visual units. Whilst we may correctly be able to identify the lip

shapes for each of the distinguishable sounds in a word, it is quite possible that none of these ideal targets will be met in natural discourse. Some of the targets are more important than others, and will be met to a greater or lesser extent accordingly. Furthermore, the degree to which each unit is met is coloured by its context, for example the articulation of the final /t/ in 'boot' versus 'beet'. With this knowledge the most naïve methods to animate speech by direct interpolation of visemes are incorrect, and can result in visually disturbing movements.

Visible speech animation as a field focusses upon the recreation of coarticulation phenomena, and to this end several methods have been attempted: direct coarticulation modelling; mapping audio to visual parameters; and concatenation of visual units. The direct coarticulation models attempt to impose coarticulation upon the interpolation of atomic visual units. The second group of methods attempt to determine a direct relationship between audio and visual signals and exploit this in the synthesis of speech. Finally, concatenative methods take small chunks of real speech movements and paste them together to create novel utterances.

The most commonly used method for animating visual speech is to use dominance functions to represent the temporal extent of each atomic speech unit. This method, first proposed by Cohen and Massaro [Cohen and Massaro (1993)], has become the *de facto* standard for modelling coarticulation [Goff and Benoît (1996); King et al. (2000); Revéret et al. (2000); Albrecht et al. (2002)]. Unfortunately, such systems require a high degree of tuning to create visually correct speech movement. For example in [Cohen and Massaro (1993)] they require three parameters per function/parameter pair plus a global parameter controlling the shape of the dominance functions. Hidden Markov Models (HMMs) have been proposed [Brooke and Scott (1998),Brand (1999)] as a method for mapping between audio and visual parameters and thus to drive the animation directly from the audio with no intermediate annotation of the speech. Also, highly complex models of the skin/muscle structure have been used in an inverse-dynamics approach to speech animation [Pitermann and Munhall (2001)].

This paper describes a method for generating novel utterances using combinations of smaller speech fragments. The method is directly analogous to the most commonly used, and natural, audio synthesis methods which work by blending speech waveforms. Because the basis units come from real speech, coarticulation is implicitly catered for within each unit. The challenges lie in correctly selecting and blending the units together to produce the appropriate visual movements in synchrony with the audio. Examples of concatenative visual synthesis include Video-Rewrite [Bregler et al. (1997)] which blends triphone video sequences, and more recent work by Kshirsagar [Kshirsagar and Magnenat-Thalmann (2003)] on using visual-syllables (visyllables) for synthesis.

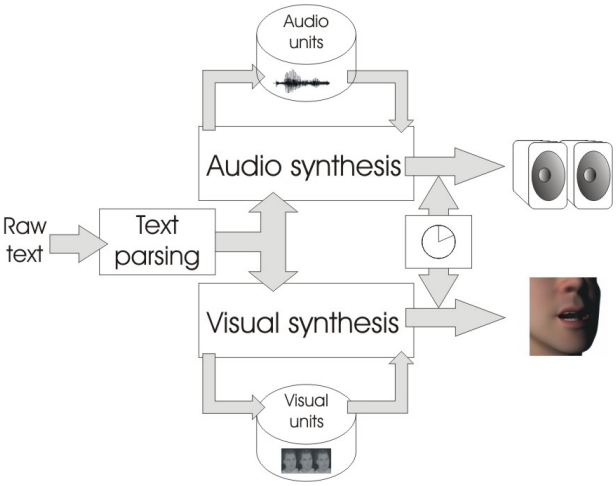


Figure 1: Overview of the synthesis system.

3 Our Approach

In this paper we introduce a method for animating speech by concatenating small segments of natural speech movements. The original speech data is in the form of motion captured sentences, segmented into units of varying sizes according to phonetic structure (e.g. phones, syllables, sentences, words etc.). By using a combination of natural motion fragments good quality speech animation can be achieved, without the necessity for complicated models of speech coarticulation.

Small fragments of speech are used to animate visual speech movements using a combination of motion warping, resampling and blending. Initially the fragments are warped such that they are phonetically aligned with the target utterance. Having stretched/squashed the fragments, each must be resampled to allow a consistent frame-rate throughout the animation. Finally, overlapping regions in the speech fragments are blended to provide smooth transitions.

Motion data is captured using a commercial Vicon mocap system. High-speed cameras, operating at 120Hz, capture the movement of markers placed on an actors face. The resulting data is a sparse sampling of the surface motion of the skin during speech production.

In order to animate a high resolution model of the skin from the motion of a few sparsely sampled points an intermediate deformer surface is used to map the motion to individual vertices. This controlling structure is composed of a set of Bézier triangles spanning the motion captured points. The deformation technique provides smooth natural animation of a face mesh, even when only provided with a sparse sampling of the original facial motion.

The process can be summarised into the following stages:

- **Data Capture** - A corpus of natural human speech motion (visual component) and sounds (audio component) is captured.

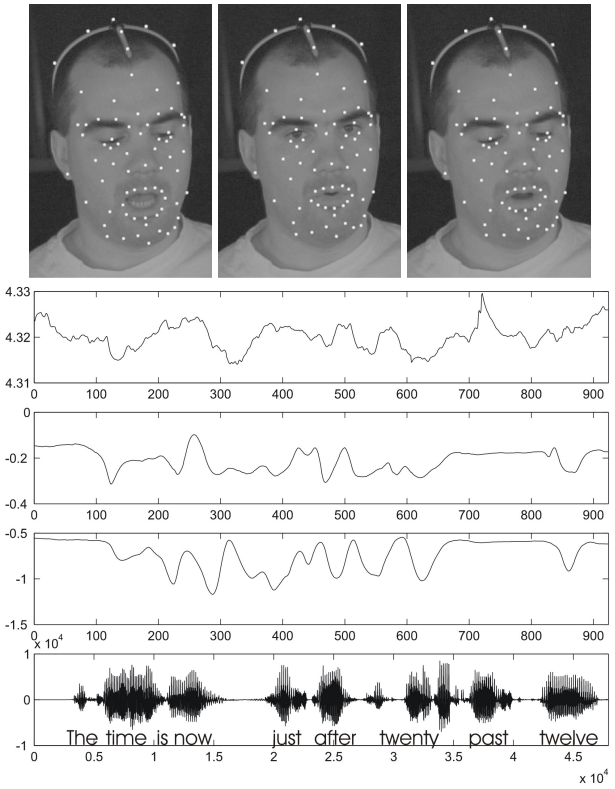


Figure 2: Example frames, motion trajectories and waveform from the captured database.

- **Preprocessing** - Rigid motion (e.g. head motion) and noise is removed from the motion-captured data.
- **Unit Generation** - Motion data is split into fragments representing the visual aspects of sentences, words and diphones (phone-to-phone transitions).
- **Motion Synthesis** - Combinations of motion fragments are used to generate visual information for novel speech utterances.
- **Retargetting** - Synthesized motions are transformed into the space of the target mesh according to the algorithms described in [Sánchez et al. (2003)]. This allows the captured motion data to be used in the animation of meshes which vary in both shape and scale from the original actor.
- **Animation** - Synchronized animation is produced using the BIDs (Bézier triangle Induced Deformations) [Edge et al. (2004)] technique to animate the visual component.

The implemented system performs Text-to-Audio-Visual-Speech (TTAVS) synthesis; an overview of the system structure is shown in Figure 1. Synthesis is split directly into separate audio and visual components: Festival [Black et al. (1999)] is used to synthesize the audio; visual synthesis is the focus of this paper and consists of unit selection, temporal alignment, and blending to generate the final motion. Both audio and visual synthesis components hold databases of speech data, synchronously captured from the same actor so that the synthesized motions will be correct for the synthesized audio. As input to the system raw text is taken from the user, this is phonetically annotated to act as input to the process. The same input is used for both audio and visual synthesis, however, each system is fundamentally independant and there is no assertion that the corresponding units will be used in both modalities.

4 Data Description

The data used in this paper consists of motion data from a commercial Vicon capture system. High speed cameras, operating at 120Hz, capture the movement of 66 markers on the surface of an actors face plus 7 more on a head mounted jig to capture rigid motion. Audio data was captured simultaneously and has been synchronized with the motion data. Figure 2 shows several frames from the captured data alongside captured motion parameters and audio.

Fifty-five sentences were captured from a limited domain time corpus, the sentences take the following form:

prompt	:=	{*prolog*} / {*time-info*} / {*day-info*}.
time-info	:=	{*exactness*} {*minutes*} {*hours*}
prolog	:=	'the time is now'
exactness	:=	'exactly' **or** 'just after' **or** 'a little after' **or** 'almost'
minutes	:=	'five past' **or** 'ten past' **or** 'quarter past' **or** 'twenty past' **or** 'twenty-five past' **or** 'half past' **or** 'twenty-five to' **or** 'twenty to' **or** 'quarter to' **or** 'ten to' **or** 'five to'
hours	:=	'one' **or** 'two' **or** . . . **or** 'twelve'
day-info	:=	'in the morning' **or** 'afternoon' **or** 'am' **or** 'pm'

This corpus can be used to generate simple time sentences such as:

'the time is now / exactly one / in the afternoon.'
or *'the time is now / quarter to ten / in the morning.'*

The data is specific to the time domain, and thus the implemented system presented in this paper is limited in generality. However, the techniques described are equally applicable to larger corpora or general synthesis using, for example, diphones as the lowest level speech unit. A simple corpus has been used to demonstrate the general technique and to ensure consistency in the dataset.

The captured motions require some processing in order to both remove noise and reconstruct missing data. Kalman filtering is used to remove noise from the data,

whilst resampling of the DCT is used to reconstruct the missing data segments, typically caused by marker occlusion. The rigid head motion is also removed at this stage using a combination of the estimate from the head mounted jig and a least-squares approach. This last step has the added benefit that the motion samples are initially spatially aligned enabling simpler concatenation during synthesis.

5 Speech Synthesis

The vast majority of techniques for the synthesis of visual speech movement rely upon the interpolation of a set of atomic phonetic units. The more successful paradigm, certainly in the case of audio synthesis, relies upon the concatenation of natural segments of speech. The analogous methods in the visual domain are becoming more popular [Bregler et al. (1997); Kshirsagar and Magnenat-Thalmann (2003)]. In this paper fragments of visual speech are concatenated to provide speech animation.

5.1 Visual Speech Fragments

As previously mentioned animating speech from small motion fragments provides the advantage that coarticulation need not be modelled, and the naturalness of the movements is implicit. However, there are also problems with this data-driven approach. Primarily, a database covering the entirity of the target domain must be captured. This impinges upon the size of fragments captured, for example if diphone (phone-to-phone) transitions are used there will be approximately 1500 units for British English. Larger units, such as syllables and words, will require an even greater (possibly unmanagable) database for synthesis. This choice of synthesis unit is a matter of balance, as it is also the case that larger fragments produce more natural animation.

Here, for the purposes of demonstration, sentences from the time domain are used. From these sentences diphone, syllable, word, and sentence fragments are extracted for synthesis. Together these fragments can be used to resynthesize any sentence from the time domain described in Section 4. In order to construct novel utterances from these fragments the following stages must be conducted:

- **Unit Selection** - Appropriate units must be selected from the database to generate the utterance.
- **Phonetic Alignment** - Each of the selected units must be phonetically aligned such that the movements appear in synchrony with the speech.
- **Resampling** - As a consequence of alignment speech fragments must be resampled to a consistent frame-rate for animation.
- **Blending** - Having aligned and resampled the motions, overlapping sections are blended to achieve a consistent trajectory over the synthesized utterance.
- **Retargetting and Animation** - A target face model is animated from the synthesized speech movements using the techniques in Section 6.

5.1.1 Unit Selection

The technique for unit selection used is dependent upon the underlying speech units. In this case units of varying duration are available, and thus a method must be defined to select the most appropriate selection to synthesize a target utterance. As input to the process the phonetic labels and timing of the target utterance are required, which can be directly recovered from the audio synthesis procedure (in this case the Festival synthesis system [Black et al. (1999)]). Pseudocode for the basic algorithm is shown below.

Fragment Selection Algorithm
Input: List of $phones$
Output: List of $fragments$

$frags \leftarrow []$
$i \leftarrow 1$
$j \leftarrow numPhones$
while $i < numPhones$ **do**
    while not FIND-UNIT$(phones, i, j)$ **do**
        $j \leftarrow j - 1$
    end while
    APPEND-UNIT$(frags, phones, i, j)$
    $i \leftarrow j$
    $j \leftarrow numPhones$
end while

In this code FIND-UNIT is a subprocedure which searches for a speech fragment which spans several phones in the target utterance, e.g. the closed sequence ['c','a','t']. APPEND-UNIT appends the found unit to the output list of fragments. Primarily this algorithm chooses fragments of longer duration, which is beneficial to the naturalness of the output speech. However, disambiguation is required where more than one speech fragment is available within the database for a given sequence. In this case, the factors which are taken into account when selecting units are: similarity in the phonetic timing to the target utterance, and similarity of context. Each of these conditions biases towards using fragments as similar as possible to the target utterance, and thus the synthesized trajectories should maintain the naturalness in movement of the captured data.

5.1.2 Alignment and Resampling of Speech Fragments

Given an appropriate selection of units, the next stage is to adapt these fragments so that in combination they can

be used to synthesize the target utterance. Essentially, this requires that the units are temporally aligned with the target utterance. Each speech fragment, whether it be diphone or a sentence, has a phonetic labelling, and must be variously stretched/squashed so that the labels are correctly aligned with the phonetic structure of the synthesized audio.

Simply, this can be achieved by evenly distributing motion samples between repositioned phonetic labels. However, this will lead to an uneven distribution in the sampling of the speech fragments, which will give an inconsistent frame-rate for animation. For this reason, having adapted the fragments so that they are aligned with the target utterance, the fragments must be further resampled to achieve a consistent frame-rate before blending.

This is the scattered-data interpolation problem, i.e. given a scattered sampling of data form a continuous curve/surface passing through the points. Many methods, such as B-spline interpolation, could be used to resample the data, here radial-basis functions (RBFs) are used.

The RBF method forms an interpolant as a linear combination of basis functions (1).

$$f(x) = p_m(x) + \sum_{i=1}^{n} \alpha_i \phi(|x - c_i|) \qquad (1)$$

In (1) the interpolated point, $f(x)$, is a linear combination of n basis functions, $\phi(x)$, and a polynomial term, $p_m(x)$. Each basis function is termed *radial* because its scalar value depends only upon the distance from its centre, c_i. The basis function used here is the inverse multiquadric, which has the advantage of being continuous in all derivatives, i.e. C^∞. The key step in using this form of interpolation is to determine the weights, α_i, which ensure that all of the basis centres are exactly interpolated. The weights can simply be determined by placing the basis centres back into (1), and solving the resulting system of linear equations. For a more thorough discussion of RBF interpolation refer to [Ruprecht and Muller (1995)].

To use RBFs for the purposes of resampling motion fragments, a basis centre is placed at each sampled point, ensuring that the interpolating curve will exactly fit the known data. The interpolated motions are in fact a mapping from the time-domain onto the spatial domain, and thus to finally resample the data requires only querying the interpolated motion at uniform temporal intervals.

5.1.3 Blending Motions

The final stage of synthesis, given appropriate aligned speech fragments from the previous stages, is to blend the fragments such that continuous motion is exhibited in the resulting animation. This involves only the overlapping regions of motions at the joints, a small degree of context is required in the fragments to facilitate this. Within the overlapping section, $t \in [t_0, t_1]$, a weighted blend of the two motion fragments to be concatenated is used (2).

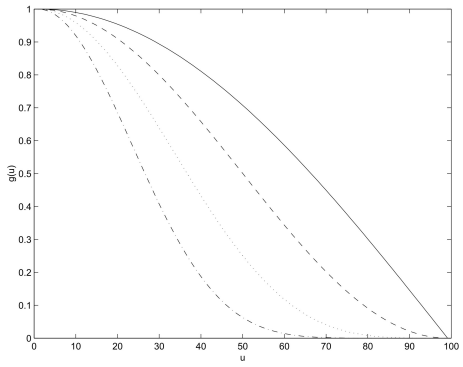


Figure 3: Example weighting functions.

$$\theta_{blend}(t) = g(u)\theta_0(t) + (1 - g(u))\theta_1(t) \qquad (2)$$

$$where \quad u = \left(\frac{t - t_0}{t_1 - t_0}\right)$$

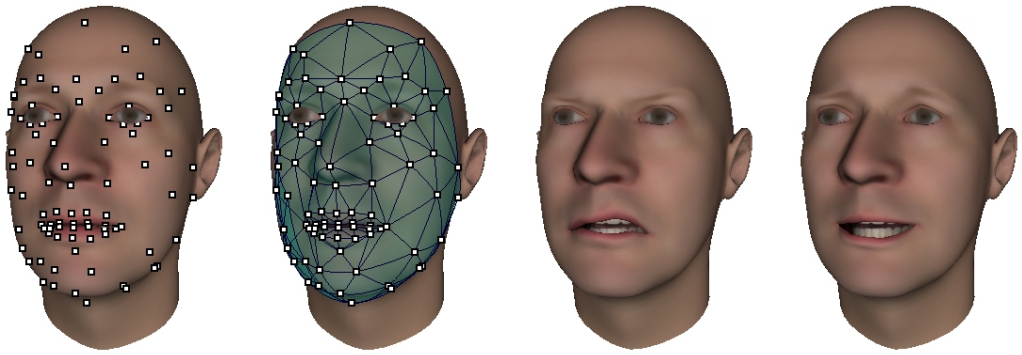
In (2), $g(u)$ is a weighting function (see fig. 3) which returns a value in the interval $[0, 1]$. The weighting function facilitates the blend and ensures a smooth transition between the fragments, which are represented here as functions of time ($\theta_x(t)$). The speed of decay in g will determine how fast the second fragment is faded in.

The use of blending relies upon the alignment of the motion fragments which is ensured in a preprocessing stage along with the removal of extraneous noise in the signals. The size of the overlapping regions depends upon the frame-rate of the fragments themselves, however, they should always be a fraction of the smallest phone-to-phone interval to prevent large fragments dominating over the target utterance. In practice, for animation frame-rates of 30 fps, there will never be more than a couple of frames overlap at each join, and for this reason high speed capture is advantageous as it allows larger blend intervals.

6 Retargetting and Animation

The result of synthesis by the techniques described in this paper leads to motion trajectories for a sparse sampling of points on the source actors face. This data is limited without further processing both to retarget the motions to a particular target mesh and to embed the motion in that mesh by interpolating the motion of points across its surface.

In order to retarget the motion to a target mesh we use the method described in [Sánchez et al. (2003)]. Because a mesh may vary in shape, scale and orientation this method consists of a volume warping method which provides a continuous mapping from the space of the original motion captured data to the space of the target mesh, i.e. $f : \mathbb{R}^3 \rightarrow \mathbb{R}^3$.



(a) (b) (c)

Figure 4: Face control structure: (a) motion-captured points; (b) triangulated Bézier control surface; (c) modelled facial expressions.

The mapping is determined using RBFs, described earlier in Section 5.1.2, to define a mapping between a neutral facial pose in the source motion and the target mesh. By reapplication of the warping function subsequent frames in the motion can be retargetted relative to the neutral pose. The fundamental technique is relatively simple, however, its application requires several technical issues to be addressed. For a full discussion of the retargetting task for facial animation and details of the technique in particular refer to the original paper [Sánchez et al. (2003)].

The result of the retargetting process is a sparsely-sampled motion embedded in a target mesh. Given that the mesh itself is often far more densely sampled than the motion it is important to interpolate the motion across the surface in a manner that preserves both the motion itself, and the characteristic geometric structure of the mesh (for example, the discontinuity between the lips).

In the system a surface oriented free-form deformation technique is used for this purpose. Free-form deformation tools provide control over high resolution meshes using a small number of controlling structures, usually lattices of control points [Sederberg and Parry (1986); Coquillart (1990); Singh and Fiume (1998)]. Here a deformer surface is defined as a triangulation of the motion captured points (fig. 4 (a) and (b)). This controlling structure is a Bézier triangle surface with continuity conditions at patch boundaries. Vertices in the target mesh are parameterized according to the parametric coordinates, $[u_V, v_V]$, of their projected image on the closest controller element (Bézier triangle), along with a normal offset, d_V, from the surface (3).

$$V_{def} = B_i(u_V, v_V) + d_V n_i(u_V, v_V) \qquad (3)$$

In (3), B_i is the parametric definition of the i^{th} triangular Bézier patch, and n_i its unitary normal map. As control points in the deformer surface are manipulated the target mesh will deform accordingly, maintaining its geometric relationship with the deformer. Figure 4 (c) shows example modelled facial expressions created using this technique.

Further constraints can be placed upon the attachment process to maintain discontinuities in the target mesh. This consists of thresholding the maximum angle allowed between the surface normals of the vertex and its image in the parametric domain of the closest Bézier element. Such constraints assert a similarity condition for the attachment of the target to the deformer. This is particularly important in controlling the movement of the lips which must be able to move entirely independently. Also, this technique does not require any form of explicit masking [Sánchez et al. (2003)] or other manual labour to be applied to an entirely different mesh.

The deformation technique is used to interpolate the movement of motion-captured points across the surface of a target mesh. Because the deformer surface approximates the mesh we achieve realistic and physically plausible movement from only a sparse sampling of an actors face. A more detailed description of the Bézier Induced Deformations (BIDs) technique can be found in [Edge et al. (2004)].

7 Results

Several frames and motion trajectories from an example animation are shown in Figure 5. Animations generated by the system demonstrate physically plausible motions, as would be expected given that the basis-units for synthesis are directly captured movements. This is particularly evident in the movement of the skin in the cheeks which is not often accounted for in morphable models of vocal articulation. The skin visibly stretches and bulges as you would expect from a physical model of the human skin, e.g. [Lee et al. (1995)].

The described system is only capable of deriving skin movement, which is the only movement evident in the initial motion captured samples. In order to animate the tongue and lower jaw either a more complex database needs to be captured (e.g. using electropalatography to determine tongue movement) or, as has been done here, a backoff technique can be used. The movement of the jaw is determined directly from the motion of captured points on the skin covering the jaw using. The tongue uses a simple morph-based model which is adequate given that it is often occluded both by the teeth and the lips.

All animation techniques described in this paper can be implemented in real-time on current PC hardware. The synthesis of individual phrases is also not particularly computationally intensive, however, the processing time will necesarily depend upon the length of the target utterance and the number of motion fragments required. Preprocessing and data preparation tasks can be labour intensive (for example, phonetically labelling the captured audio), but only need to be performed once per database.

8 Conclusions

There are three key advantages to using motion fragments in visual speech synthesis:

- Motion-captured data implicitly encapsulates dynamic coarticulation effects.
- It allows the unification of audio and visual synthesis by using the correct motions for the audio units concatenated during synthesis.
- An improvement in the naturalness of speech movements is attained, especially in comparison to interpolation techniques such as [Cohen and Massaro (1993); Goff and Benoît (1996); Albrecht et al. (2002)].

Furthermore, due to the use of retargetting and generic animation techniques, the implemented system is capable of driving any reasonable facial mesh. The parameters used are not model-specific, nor are we constrained to use particular point-sets (e.g. MPEG-4), or mesh topologies (e.g. [Parke (1974)]) making the system both generic and scalable.

One improvement over the currently implemented system would be to link unit selection in the audio and visual modalities. For each motion sequence we also retain the audio data, which is used by Festival for synthesis. Currently there is no link between unit selection, and so visual units may be selected which were not captured at the same time as the selected audio units. Using Festival to select both audio *and* visual units may remedy this and produce a slight benefit in audio-visual synchronicity. Furthermore, there may be some benefit in expanding upon selection criteria for visual units to bias towards units with given boundary conditions, i.e. similarity in the overlapping blend period.

The disadvantages to using motion capture for these purposes lies in the size of databases required to perform general synthesis (as opposed to limited-domain, e.g. rail announcements or time - as in this paper). Larger motion units will lead to an increase in the quality of synthesized speech, however, it also leads to larger-scale initial data capture. For example, using triphones as the lowest level speech unit will require approximately 1500 units in British English. Capturing this amount of audio-visual data is highly complex, particularly with regards to maintaining consistency in the database. In order to tackle these problems we use multiple scale units (phones, syllables, words etc.) to allow the greatest quality synthesis for the captured data. By the aforementioned use of retargetting techniques we are also able to capture once and use the data multiple times, i.e. to animate several characters.

One direction for future research is the use of backoff techniques to merge concatenative techniques with morph based models. This is analogous to the letter-to-sound rules used in audio synthesis and would allow motion-data to be used even without restricting synthesis to domain-specific problems or requiring large-scale data capture. Currently no systems have attempted to mix synthesis techniques in this manner, and the most commonly used coarticulation technique [Cohen and Massaro (1993)] is inappropriate for the task because continuity considerations at utterance boundaries are not taken into account.

The future of visual speech synthesis lies in the use of larger dynamic units, as has been long recognised by the audio synthesis community. The remaining problems focus upon concurent signals such as emotional and gestural signals. This would require that algorithms are developed to blend sampled motions without destroying the link between speech movements and audio. The authors believe that the wealth of research in full-body motion capture [Bruderlin and Williams (1995)] as well as recent developments in decomposing motions [Cao et al. (2003)] could be exploited to tackle this problem.

Acknowledgements

The authors would like to thank the EPSRC and the Pedro Barrié de la Maza Foundation for providing funding for this research, as well as the Advanced Computing Center for the Arts and Design Motion Capture Lab for the collection of the motion data. We would also like to thank Dr Scott King for his help in producing this research.

References

I. Albrecht, J. Haber, and H-P Seidel. Speech synchronization for physics-based facial animation. In *Proceedings WSCG'02*, pages 9–16, 2002.

A. Black, P. Taylor, and R. Caley. The festival speech synthesis system. (`http://www.cstr.ed.ac.uk/projects/festival/manual/festival_toc.html`), June 1999.

M. Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 21–28, 1999.

C. Bregler, M. Covell, and M. Slaney. Video rewrite: driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360, 1997.

N.M. Brooke and S.D. Scott. Two and three-dimensional audio visual speech synthesis. *Proc. AVSP'98*, 1998.

A. Bruderlin and L. Williams. Motion signal processing. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 97–104, 1995.

Y. Cao, P. Faloutsos, and F. Pighin. Unsupervised learning for speech motion editing. In *Proceedings of Symposium on Computer Animation*, pages 225–231, 2003.

M.M. Cohen and D.W. Massaro. Modeling coarticulation in synthetic visual speech. In *Proceedings Computer Animation '93*, pages 139–156, 1993.

S. Coquillart. Extended free-form deformation: a sculpturing tool for 3d geometric modeling. In *Proceedings of the 17th annual conference on Computer graphics and interactive techniques*, pages 187–196, 1990.

J.D. Edge, M.A. Sánchez, and S. Maddock. Animating speech from motion fragments. Technical Report CS-04-02, Department of Computer Science, University of Sheffield, 2004.

B. Le Goff and C. Benoît. A text-to-audiovisual-speech synthesizer for french. In *Proc. ICSLP'96*, volume 4, pages 2163–2166, Philadelphia, PA, 1996.

S. King, R.E. Parent, and B.L. Olafsky. An anotomically-based 3d parametric lip model to support facial animation and synchronized speech. *Proc. Deform 2000*, pages 7–9, 2000.

S. Kshirsagar and N. Magnenat-Thalmann. Visyllable based speech animation. In *Proceedings Eurographics 2003*, 2003.

Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 55–62, 1995.

A. Löfqvist. Speech as audible gestures. *Speech Production and Speech Modelling*, pages 289–322, 1990.

D. Massaro. *Perceiving Talking Faces : From Speech Perception to a Behavioral Principle.* Bradford Books Series in Cognitive Psychology. MIT Press, 1998.

H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

S.E.G. Öhman. Numerical model of coarticulation. *Journal of the Acoustical Society of America*, 41:310–320, 1967.

F.I. Parke. *A parametric model for human faces*. PhD thesis, University of Utah, 1974.

F.I. Parke and K. Waters. *Computer Facial Animation*. A. K. Peters, Ltd., 1996.

M. Pitermann and K.G. Munhall. An inverse dynamics approach to face animation. *Journal of the Acoustical Society of America*, 110:1570–1580, 2001.

Lionel Revéret, Gérard Bailly, and Pierre Badin. Mother : A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. *6th Int. Conference of Spoken Language Processing, ICSLP'2000*, 2000.

D. Ruprecht and H. Muller. Image warping with scattered data interpolation. *IEEE Computer Graphics and Applications*, 3:37–43, 1995.

M. Sánchez, J.D. Edge, S.A. King, and S. Maddock. Use and re-use of facial motion capture data. In *Proceedings Vision, Video and Graphics*, pages 135–142, 2003.

T. W. Sederberg and S. R. Parry. Free-form deformation of solid geometric models. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 151–160, 1986.

K. Singh and E. Fiume. Wires: a geometric deformation technique. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 405–414, 1998.

W.H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, 1954.

K. Waters. A muscle model for animation three-dimensional facial expression. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pages 17–24, 1987.

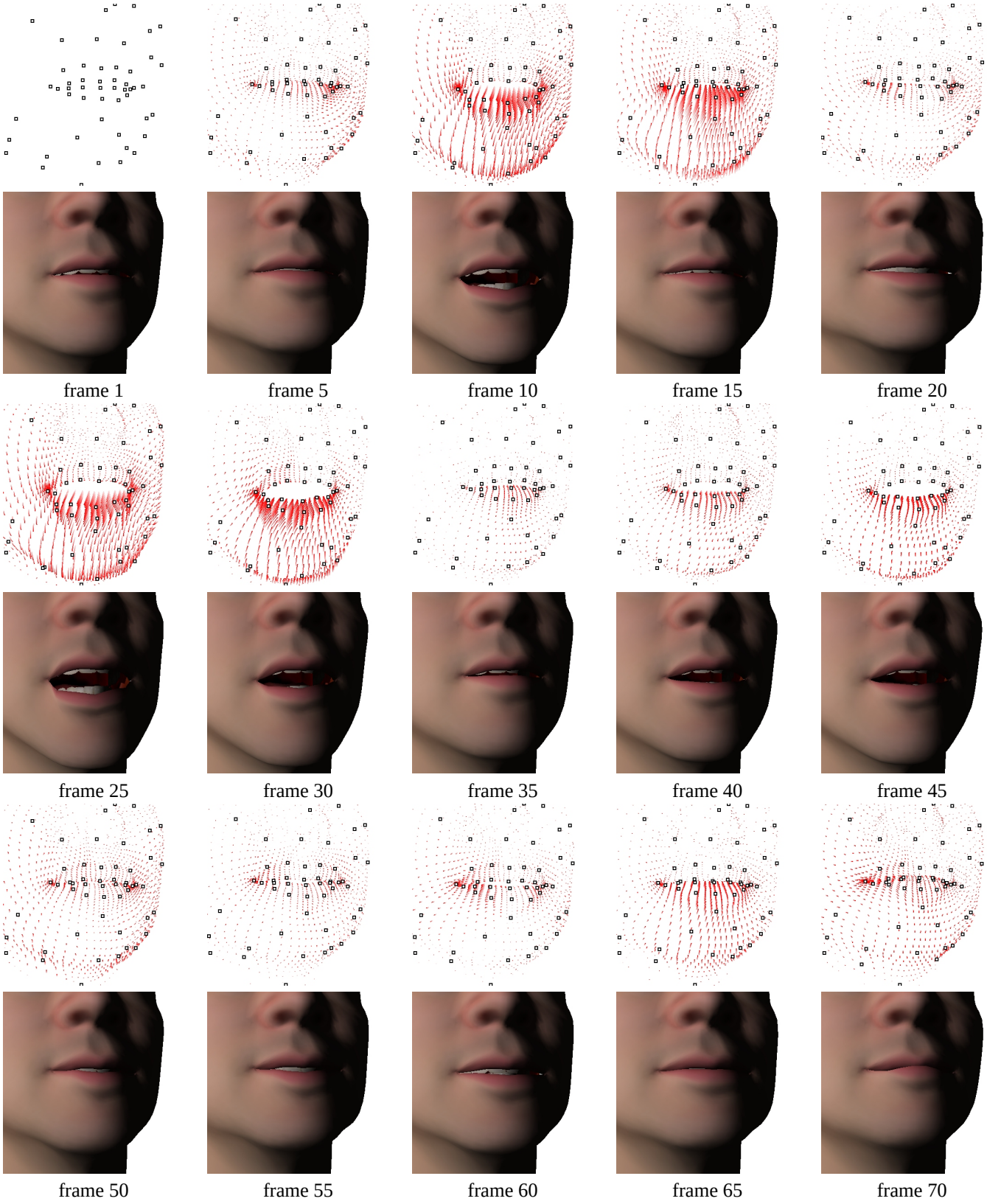


Figure 5: Example frames and vertex trajectories from a speech animation.

Influences on Embodied Conversational Agent's Expressivity: Toward an Individualization of the ECAs

Vincent Maya
LINC
IUT of Montreuil
University Paris VIII
maya@iut-univ.paris8.fr

Myriam Lamolle
LINC
IUT of Montreuil
University Paris VIII
m.lamolle@iut-univ.paris8.fr

Catherine Pelachaud
LINC
IUT of Montreuil
University Paris VIII
c.pelachaud@iut-univ.paris8.fr

Abstract

We aim at creating not a generic Embodied Conversational Agents (ECAs) but an agent with a specific individuality. Our approach is based on different expressivities: the agent's expressivity, the communicative of behavioral expressivity. Contextual factors as well as factors such as culture and personality shape the expressivity of an agent. We call such factors "influences". Expressivity is described in terms of signals (e.g. smile, hand gesture, look at) and their temporal course. In this paper, we are interesting in modelling the effects of influences may have in the determination of signals. We propose a computational model of these influences and of the agent's expressivity. We have developed a taxonomy of signals according to their modality (i.e. face, posture, gesture, or gaze), to their related meaning and to their correspondence to expressivity domains (the range of expressivity than they may express). This model takes also into account the signals dynamic instantiation, i.e. the modification of signals to alter their expressivity (without modify the corresponding meaning).

1 Introduction

We aim at creating an **E**mbodied **C**onversational **A**gent (ECA) that would exhibit not only a consistent behavior with her personality and contextual environment factors but also that would be defined as an individual and not as a generic agent. Most of the agents that have been created so far are very generic in their behavior type. We want to simulate that two different agents may behave differently in a same context and may express the same felt emotion differently, even if they belong to the same social-cultural sphere. Given a goal in mind, people differ in their manner of expressing themselves.

Several studies have shown the importance to consider complex information such as cultural factors, personality, environment setting when designing an agent (Brislin (1993), H. Morton and Jack (To appear), Isbister and Nass (Forthcoming) and Lee and Nass (1998)). These factors affect the interaction a user may have with the agent. Personality also is an important aspect that makes people look and act very differently. Gender, age (child vs. teenager), our social role (e.g., mother, doctor), our experience and memory intervene in the manner we interact with others, we talk about things. These differences arise at different levels: the formulation of our thought as well as their expression. They have influences on the surface generation and realization level during the dialog phase as well as on the selection of the non-verbal behaviors and of their expressivity (Ruttkay et al. (2003)).

At first we propose a taxonomy of the influences types. These influences act on what to say and when as well on how to say it and to express it. In this paper, we concentrate on the effects of the influences on the facial, gaze, gesture and body behaviors but we do not consider the modification of the speech (choice of word variation of paralinguistic values, of intonation tones). However we assume that the input text embeds these effect. Expressivity acts not only on the selection of a non-verbal behavior to convey a meaning but also on its *expressivity*, i.e. the *strength* of this non-verbal behavior. For example, in order to express the surprise, the agent may raise her eyelids, and the more the surprise is strong, the more the eyelids are up.

We model expressive effects within our framework based on **APML**, (Affective Presentation Markup Language, see DeCarolis et al. (2004)), based on a taxonomy of communicative functions proposed by I. Poggi (see Poggi et al. (2000)), and the ECA **GRETA**, showed in figure 1 and figure 7 (see Pelachaud and Poggi (2002)), we model their effects. From a tag that indicate only the meaning the agent has to express and its expressivity, we want to obtain a tag that indicate the signal to use, among all that are stored in libraries, and the technical value allowing the system to modulate this signal, that we call the dynamic instantiation coefficient.

In this paper, after presenting a state of art in section 2, we describe a taxonomy of influences in section 3. We then define what we mean by expressivity in section 4 and

Figure 1: Greta

the agent's model in section 5. In section 6, we describe our normalization tools allowing one to associate expressivity and dynamic instantiation information to signals, in order to use the model proposed in this paper. Finally we provide an overview of our APML translator.

2 State of the art

Agents exhibiting emotional behaviors have received quite some interest. In Ball and Breese (2000), the authors developed a model in which the emotion an agent is undergoing may affect her verbal and non-verbal behavior. They built a Belief Network that links emotion with verbal and non-verbal manifestation. Fiorella de Rosis and her colleagues have developped a computational model of emotion triggering using a dynamic Belief Network (Carofiglio et al. (in press)). Their model is able to determinate not only which emotion is triggered after a certain event for a given agent but also it is able to compute the variation of this emotion over time: this emotion may increase or decrease in intensity or it may also evolve in another emotion. The computational model uses a Belief Desire Intention (BDI) model of the agent's mental state. Fuzzy logic has been used to either model the triggering of emotions due to events (El-Nasr et al. (2003)) or to map facial expressions of an emotion with a given intensity (Bui et al. (2003)).

Emotions have also been considered during the interaction of a user with a system. Within the EU project Safira (Höök (To appear)), a new interaction device have been developed to interact with actors of a video game. This device is SenToy, a teddy bear with sensors attached to its joints (Paiva et al. (2003)). The user moves around the toy using a set of pre-defined moves with a given expressivity. These emotional behaviors are detected. When recognized by the system they are used to drive the behavior of one of the agents in the video game.

Several talking heads able to show emotions have been developed. In particular, in Kshirsagar et al. (2001), the authors have developed an agent that is able to react to the user's emotion detected through his facial expressions. This reaction is based on a computational model of emotion behavior that integrates a personality model. Carmen's Bright IDEAS (Marsella et al. (2000)) is an interactive drama where characters exhibit gestures based on their emotional states and personality traits. Through a feedback mechanism a gesture made by of a character may modulate her affective state. A model of coping behaviors have been developed by Marsella and Gratch (2003). The authors propose a model that embeds information such as the personality of the agent, his social role.

In an attempt to model cultural behavior for a talking head, King et al. (2003) have proposed a simple model using a table of correspondence between a given meaning and its associated behaviors. Scott King built such a table for each culture he considered (English and Maori). We are aware of very few other attempts.

The role of social context in an agent's behavior have been considered. DeCarolis et al. (2001) propose a model that decides whether an agent will display or not her emotion depending on several contextual and personality factors. Prendinger et al. (2002) integrate contextual variables, such as social distance, social power and thread, in their computation of the verbal and non-verbal behavior of an agent. They propose a statistical model to compute the intensity of each behavior. Rist and Schmitt (2003) modelled how social relationship and attitudes toward others affect the dynamism of an interaction between several agents.

To control the behavior of ECAs, several representation languages has been developed. Theses languages specify the agent's behaviors. They serve as interface between the different modules of the architecture of an agent system. The languages may embed various levels of abstraction: ranging from the description of the signals (a smile, a head nod) in VHML (VHML) to semantic information (rheme/theme, iconic) in Piwek et al. (2002), going through communicative function (performative, emotion) (DeCarolis et al. (2004)). Of particular interest to our work is the language SCRipting Emotion-

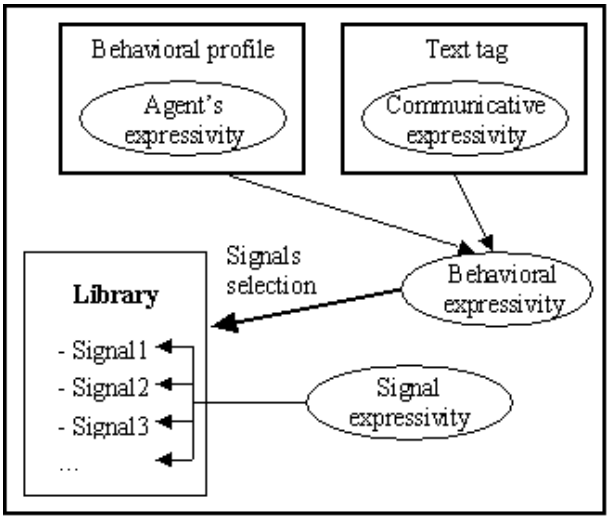


Figure 2: Expressivity Types

based Agent Minds (SCREAM). SCREAM has been designed to create emotionally and socially appropriate responses of animated agents placed in an interactive environment (Prendinger et al. (2002)).

The work that is more related to our is Ruttkay et al. (2003). The authors aim at creating agents with style. The authors aim at creating agents with style. They developed a very complex representation language based on several dictionaries. Each dictionary reflects an aspect of the style (e.g. cultural or professional characteristics or personality). They also defines the association meaning to signals. In this language the authors embed notions such as culture, personality, gender but also physical information such as gesturing manner or tiredness. To create an agent with style one needs to select a set of values (e.g. an Italian extrovert professor). The proper set of mappings between meanings and signals is then instantiated. The authors modelled explicitly how factors such as culture and personality affect behaviors. We distinguish our work from them in the sense that we do not modelled such factors, rather we modelled the different types of influences that may occur and how they may modulate an agent's behaviors.

3 Taxonomy of Influences

We call "influence" different factors: contextual factors as well as factors such as culture and personality. These factors shape the expressivity of an agent. Influences may act on the selection of a non-verbal behavior to convey a meaning (i.e. on the choice of the signals), on the expressivity of this behavior (e.g. on their intensity level), in order to lessen it or to accentuate it and on the communication strategies.

We differentiate three types of influences. The first type contains the *intrinsic* influences. We consider that each human being has a set of the conscious and unconscious habits that are reflected in the content of her discourse and that define her attitude and her behavior when she talks. These habits derive, amongst others, from her personality, her age, her sex, her nationality, her culture, her education and her experiences (Brown and Nichols-English (1999)). For example, some non-verbal behaviors are very culturally dependent, such as the emblems, gestures that may be directly translated into words. This might be the case with some iconic or metaphoric gestures that may take their origins in the culture (the action of eating will not be represented identically if one is used to eat with a fork or with sticks). However, not all gestures are culturally dependent. The type of gestures one makes while conversing might not differ over various cultures as much as one would have thought at first. The main discrepancy is more in the quantity of gestures made rather than on the type of gestures itself (Cassell (2000)).

The second type contains the *external* influences, that refer to the environment setting, such as the light conditions, the sound intensity, the spatial layout or the function of the conversation site. These factors may affect some speech and behavior characteristics. For instance, in a crowed room, some wide gestures are simply impossible. Likewise, in a noisy room, a person has to increase the volume of her voice due to pragmatic considerations. She will also be inclined to amplify her gestures. In the opposite, in a religious building or in a museum, a person ought to remain quiet and move silently due to social considerations. Another factor of influences is how the agent is placed in the environment; a person sitting in an armchair will not gesticulate as a person standing in a hallway.

The third type contains the *mental* and *emotional* influences. The mental state of the agent affects greatly the way the agent will behave: it modifies the prosody of speech, the amplitude of a facial expression, the movement tempo. A person does not talk and does not behave the same way whether she is angry or not. Her relationships with her interlocutor modulate also her behavior: she does not behave in the same way with a friend, and unknown person, an employee, a child or a doctor (DeCarolis et al. (2002)). The agent's mental state evolves all along the conversation. Her emotion varies through time, her goals and beliefs get modified as the conversation evolved.

In this paper, we oppose the *intrinsic* influences to the other ones, which we group in the *contextual* influences. The intrinsic factors are constant during a dialog session, whereas the contextual ones may vary. The contextual factors increase or decrease the effects of the intrinsic factors, or even cancel them.

```
 1. <librarySignal
 2.  format = "FAP">
 3.  <signal name = "smile"
 4.     fap4 = "-100" fap8 = "-100"
 5.     fap9 = "-100" fap51= "-100"
 6.     fap55= "-100" fap56= "-100"
 7.     fap12=  "150" fap13=  "150"
 8.     fap59=  "150" fap60=  "150"
 9.     fap6 =   "50" fap7 =   "50"
10.     fap53=   "50" fap54 =  "50"
11.  />
12.  <signal name = "joy_eyelids"
13.     fap19=  "128" fap20=  "128"
14.     fap21=   "64" fap22=   "64"
15.  />
16.  <signal name = "joy_1"
17.     combination = "smile"
18.     combination = "joy_eyelids"
19.  />
20.
... </librarySignal>
```

Figure 3: Example of signals library

4 Expressivity

We call expressivity the value that allows the system to relate *strength* to the communication act.

We do not aim at modelling what culture or personality mean, nor do we aim at simulating expressive animations. We limit our scope at representing influences that would modify the set of behaviors and the quality motions a particular agent will display to communicate a given meaning within a specific context.

According to the concepts they are applied to, we differentiate several expressivities, schematized in figure 2.

4.1 Communicative expressivity

The input text is marked with tags specifying the communicative function the agent aims at displaying. Each tag may have an attribute corresponding to the degree of expressivity attached to a given meaning for an agent. Is the agent puzzled or completely confused, slightly angry or madly angry? In figure 9, at line [10], the value of *communicative expressivity* related to the joy of the agent is *0.8*.

4.2 Agent's expressivity

Agent's expressivity is related to the qualitative property of behavior. Does an agent has the tendency to play down (acts so as not to be noticed) or on the opposite wants to catch all looks by acting wild? This value is given in the agent's definition. In figure 6, this expressivity is described by the lines [10] to [16].

```
 1. <library
 2.  modality = "face">
 3.  <expression meaning = "joy"
 4.    name = "smile"
 5.    ref = "0.4"
 6.    min = "0.3" max = "0.6"
 7.    minCoef = "0.9" maxCoef ="1.5"
 8.    dynInstType = "duration"/>
 9.  <expression meaning = "joy"
10.    name = "eyelids_joy"
11.    ref = "0.2"
12.    min = "0.1" max = "0.4"
13.    minCoef = "1.0" maxCoef ="1.0"
14.    dynInstType = "amplitude"/>
15.  <expression meaning = "joy"
16.    name = "joy_1"
17.    ref = "0.7"
18.    min = "0.7" max = "0.9"
19.    minCoef = "0.8" maxCoef ="1"
20.    dynInstType = "amplitude"/>
21.  <expression meaning="anger"
22.    name ="anger_mouth"
...    ...
... </library>
```

Figure 4: Example of expression library

4.3 Behavioral expressivity

The behavioral expressivity represents the way that the considered agent expresses the tag meaning, taking into account her characteristics and the contextual factors that may modify her expressivity. It is the result of the computation from the *communicative expressivity*, the *agent's expressivity* and the contextual factors that influence the way that she expresses the communication act. It intervenes during the selection of the signal and during the computation of its *dynamic instantiation* (see section 6): it modifies the quantity of movements related to these signals, their amplitude, their duration, their dynamism and/or their repetitiveness.

4.4 Signals expressivity

In order to choose the appropriate signals that best corresponds to a given meaning, the system has to know the expressivity related to each signal. For example, it has to know that mild smile is less expressive than a large smile. *Signal libraries* (see figure 3) contain *basic signals*(*smile*, from the line [3] to the line [11]), and *hight level* signal (i.e. defined as a combination of *basic* signals such as the signal *joy_1* define from the line [16] to the line [19]).

To be able to instantiate the *behavioral expressivity* into a set of expressive signals, the animation engine has to know the signals that are potentially available and to compute the appropriate *signal expressivity*. In order to integrate this *expressivity* type, we define *expression libraries*

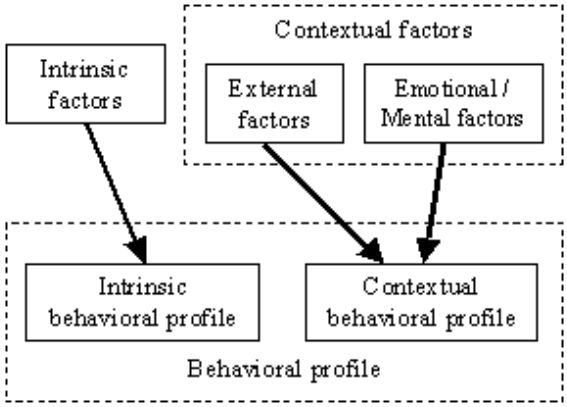


Figure 5: Agent's behavioral profile

```
 1. < agentDefinition
 2.     nameAgent        = "Agent1"
 3.     redundancyStrategy = "mono"
 4. >
 5.     <modHierarchy
 6.         face     = "1.0"
 7.         gesture  = "0.5"
 8.         position = "0.3"
 9.         gaze     = "0.4" />
10.     <intrinsicfactors
11.         face     = "0.4"
12.         gesture  = "0.4"
13.         posture  = "-0.2"
14.         <!--gaze     = "0"-->
15.     />
16. </agentDefinition>
```

Figure 6: Agent's definition

(see figure 4). These libraries contain for each communicative act a list of (meaning, signal) pairs. They also contain information about *expressivity domain* of these signals and about the way to modulate their expressivity. These *expressions libraries* have also the advantage to be independent of the format of the *signals libraries*.

To indicate the expressivity associated to signals, we use fuzzy set that define domains where the signals are appropriately usable. The attribute *ref* represents the expressivity associated to the related signal. The attributes *min* and *max* represent respectively the minimal and the maximal expressivities that can be associated to the signal. We assume that the attributes *min* and *max* take into account the signal distortion possibilities; that is the modulation of the signal does not change the meaning associated to it. The *expression libraries* indicate also the extreme distortion coefficients *coefMin* and *coefMax* to compute the distortion coefficient related to a given expressivity (see section 6.3).

Behavior expressivity may be expressed not only through the signals, and their expressivity, but also by combination of signals dispatched over modalities. We differentiate the *modal signal expressivity*, which concerns the qualitative parameters that determine the choice between several signals of a same modality with a same meaning, and the *inter-modal signal expressivity*, which is modelled by defining the functions that relates the behaviors across the modality, such as redundancy (i.e. expression of the same meaning with several signals of different modalities), complementarity (e.g. saying "*he goes to the stadium*", and complementing it with an iconic gesture that means "*he drives to the stadium*"), substitution (e.g. straight index over the mouth to mean silence), and masking (e.g. masking sadness by a smile).

5 Agent's Model

In previous section, we have shown that even two agents may have in mind a same communicative act, they may behave differently, using various expressivity or using different modalities. In this section, we feather refine our model by specifying the agent's preferences to use a given modality (e.g. an agent may have a very expressive face). We also model how the agent dispatches her behavior over different modalities.

In our model, the information that allows the system to obtain these results is described within the tag $< agentDefinition >$ (see figure 2). In this section, we describe the different elements of this tag.

5.1 Intrinsic behavioral profile

In figure 2, we associate to the agent a behavioral profile, which specifies, on the one hand, the agent's expressivity, i.e. the intrinsic factors, and, on the other hand, the effects of the contextual factors. This profile specifies the agent's expressivity depending on the modalities. It allows one to define that an agent has a very expressive face or that she rarely uses wide arm movements.

These intrinsic factors is described in the element $< intrinsicFactors >$. It associates a numeric value to the attributes *face*, *posture*, *gaze* and *gesture*. These values lessen or accentuate the expressivity of the tag meaning for the related modality. According to the description of *Agent1* in figure 6, this agent is more expressive for the face and for the gestures than the *default agent* (i.e. facial moves and her gestures are more accentuated than the *default agent's* ones but less for the posture. At line [14] the default value of the gaze attribute is 0 meaning the agent does not use this modality to communicate this specific meaning. This intrinsic profile, given aa input, is constant during a dialog session.

Figure 7: Face variation of Greta

5.2 Modality Hierarchy

In order to choose the modality that the agent will use for a given tag, we define a priority scheme on the behavior. We associate to each modality (face, gaze, gesture and posture) a numeric value that represents their preferential level in the hierarchy.

In case several modalities have the same hierarchical level the system considers the expressivity of all the signals of the concerned modalities to choose a signal for this level. In the agent's definition (see figure 6), the lines [6] to [10] describe this hierarchy. According to this description, *Agent1* uses mainly facial expressions.

5.3 Inter-modal functions

From the *communicative expressivity*, the system obtains a *behavioral expressivity*. The system computes how to express this expressivity (see algorithm described in section 6) according to the signals expressivity. This latter expressivity is defined in the expression libraries (see. figure 2). For a same meaning, several signals of different modalities may be associated (e.g. anger can be express with a frown thin lips, looking straight and tense movement). The *behavioral expressivity* is then related to how the signals are dispatched over all the modalities.

Substitution and complementarity modify the text content and are represented by tags (e.g. saying "*he goes to the stadium*", and complementing it with an iconic gesture that means "*he drives to the stadium*" for complementarity, or raising straight index over the mouth to mean silence, for substitution). Therefore, these tags must be defined in the input text.

Strategies for redundancy and masking, from a cognitive point of view as well as from a computational point of view, may vary according to the context or to the considered agent. For the moment, we consider that the masking is mainly contextual and we define its strategies in the section 6.1. Conversely, an agent may mainly employ a specific redundancy strategy. This information is given in the agent's definition (see line [3] in figure 6). We define several strategies :

- "*mono*" : only signals of one modality is used.

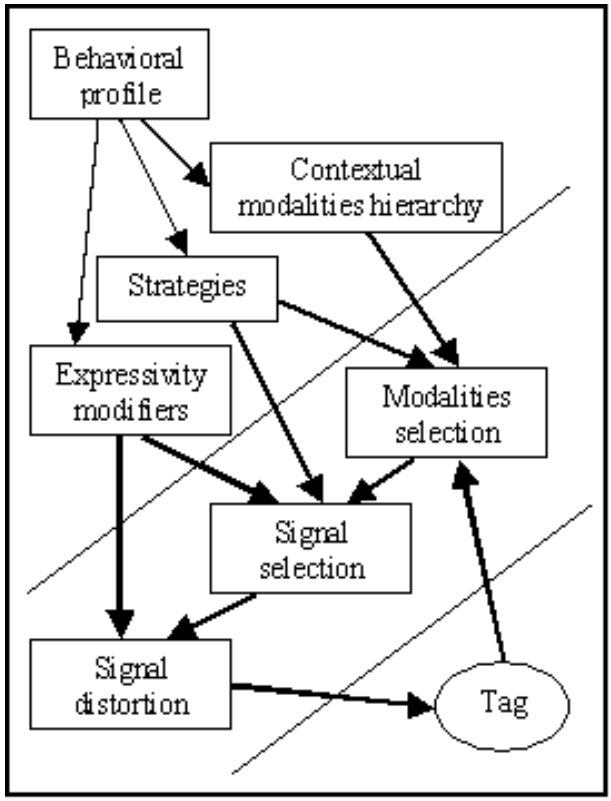


Figure 8: Tags transformation Steps

- "*maximal*" : signals of all possible modalities are selected.

- "*additional*" : redundancy is used only for the expressivity superior to a given threshold.

These strategies may be restricted a specific communicative act, such as "*certainty*" or "*performative*".

The redundancy, more than any other signal association, raises the problem of the coherence of the signals choice. Let us imagine that the system choose to express an emotion with facial and gestural signals. By the process described in the section 6, we obtain a facial signal inconsistent with the gestural signal. For example, in order to express redundantly the agent's anger, the system decides to use, on the one hand, signals related to the face crispation, and on the other hand, wide arms movements.By considering, for the sake of the example, that these signals are inconsistent, our system has to be able to determine this inconsistence (thanks to a *coherence library*) and to propose a substitution solution.

6 System overview

Our system takes as input a *intrinsic behavioral profile* that represents the agent's communication characteristics

```
 1. <agent
 2.    nameAgent       = "Agent1"
 3.    contextCoeff    = "0.5"
 4.   <!--MaskingStrategy =
 5.                  "multimodal"-->
 6. >
 7.   <performative type = "inform">
 8.      <rheme>
 9.        <affective type = "joy"
10.           expressivity = "0.8">
11.            I'm happy.
12.        </affective>
13.      </rheme>
14.   </performative>
15. </agent>
```

Figure 9: Example of input

(see figure 2 and figure 5) and a text with tags specifying communicative functions (Poggi (2003)).

The representation language used for the tags is the Affective Presentation Markup Language (APML) (DeCarolis et al. (2004)). A tag represents the meaning associated to a given communicative function. Most tags contain an expressivity attribute.In the *expression libraries*, each meaning is associated to a list of possible signals that may describe it.

From a given intrinsic behavioral profile, the system instantiates the tags into a set of signals that are then translated into animation parameters. During this instantiation phase, the process of the agent individualization is done in three steps: the modality selection, the signals choice and the signal dynamic instantiation (see figure 8).

6.1 Modalities selection

For each tag, the system has to decide the modality (*face*, *gesture*, *gaze* or *posture* one) to use to express the given meaning.

In most cases, the decision is based on the modality hierarchy: among the modalities that have at least one expression which allows the system to represent the meaning, it choose the one with the highest priority and that is not used yet, in order to prevent conflicts. Conflicts may occur for embedded tags acting on the same text span. These tags may use the same modalities for their corresponding signals. Conflicts may arise if the tags require to use the same modality. In case conflicts may not be solved using the behavior hierarchy scheme. We select one tag to prevail over the others. For gaze and gesture we choose the most embedded tag while for face we choose the outer tag (Poggi (2003)).

Some contextual factors may however modify this hierarchy. For example, for an agent that expresses her communication acts mainly by facial expression, the anger or the nervousness may incite her to use gestures more intensively.

In the tag $< agent... >$ in the input text (see figure 9), the value *mono* of the attribute ***redundancyStrategy*** indicates that *Agent1* does not use redundancy. Therefore the system selects the face modality to express *"joy"*, as this modality is not yet employed to express the performative *"inform"* (expressed by the gaze signal *"look_at"*.

The attribute *maskingStrategy* specifies the strategy to mask an expression by another one (see section 5.3). Agents may not mask their emotions in the same way and with the same efficiency than others. Masking strategy depends on the context and, in particular, on the relationships between the agent and its interlocutor. We define three strategies:

- "*attenuation*": the system attenuates the signals of the expressions, as to aim to show a neutral expression
- "*addition*": the system dissimulates the signal with a signal of an another modality, such as a hand in front of the mouth.
- "*replace*": the system replaces a signal by another one, such as a sad expression replaced by a polite smile.

6.2 Signals selection

6.2.1 Computation of the *behavioral expressivity*

First, in order to select the appropriate signals according to the influences, for the modality selected at the previous step, the system computes the *behavioral expressivity* for the desired modality. It sums up the *communicative expressivity* (given in the input text) to the value of the *intrinsic behavioral profile* related to the selected modality. This operation allows it to take the behavior of the agent for the modality into account. For example, the intrinsic behavioral may express that the agent is inclined to use wide gestures. The result of this operation is modified according to *contextCoeff*. This coefficient only aims to lessen or to accentuate the *behavioral expressivity*. This value is defined for all the text included between $< agent... >$ and $< /agent >$. The attribute *contextCoeff* expresses the effects of the contextual factors only at the level of the expressivity.

In the input text presented in figure 9, at the line [3], the attribute *contextCoef* indicates that the contextual coefficient *contextCoeff* is equal to *0.5*. This value may model for example a signal attenuation resulted on masking an expression by another one (e.g. in some situations, anger may not be shown and a polite smile may have to be displayed). Consequently, since its value is inferior to *1*, the expressivity of the tag "*inform*" and of the tag "*joy*" is lessened. In our example, the tag "*joy*" is associated to an expressivity of *0.8*. Considering that during the previous steps, the system has chosen to express the tag "*inform*" by the signal "*look_at*" of the gaze modality, it choose,

according to the modality hierarchy and to the presence of related signals in the different modalities, to use the facial modality to express the tag "*joy*". It adds to the *communicative expressivity* of the tag "*joy*", the *Agent1*'s facial expressivity, expressed by the attribute *face* of the element *intrinsicFactors* (see line [12] in figure 6). The value of this attribute is *0.4*. It obtains thus an intermediary expressivity of *1.2* (i.e. *0.8 + 0.4*). It applies then the contextual coefficient *contextCoeff* whose a value is *0.5*. The system outputs a *behavioral expressivity* e_{joy} of *0.6* (i.e. (*0.8*+*0.4*) x *0.5*).

6.2.2 Selection in the libraries

In the expression libraries, each signals description contains an *expressivity domain* (defined by a minimal and a maximal values) and a reference value. In the *expression library* described in figure 3, three signals set can represent the emotion related to the tag *"joy"*: the *basic* signals *"smile"* and *"joy_eyelids"*, defined independently of each other (Bui et al. (2003)), and the *hight-level* signal *joy_1*. The name of these signals allow the system to retrieve them in the related signals library (see figure 3).

The *behavioral expressivity* is compared to the expressivity domain of these signals, described in the *expression library* (see figure 4). We compare the value of the *behavioral expressivity* e_{joy} related to the tag *"joy"* with the boundary values *min* and *max* for each of these signals. We select the signal whose *domain* of expressivity contains the value e_{joy}.

If several signals can be selected, the system chooses according to the distance between e_{joy} and its reference expressivity or between e_{joy} and the nearest bound. There are several possible strategies. In our system, they are configurable in order to test their efficiency.

If e_{joy} does not belong to any domain (i.e. no signal with such an expressivity exists for this given meaning and this particular agent), the system chooses the expression with the nearest domain, and redefines the value of e_{joy} according to value of the nearest bound of this domain. In our example, the system selects the signal denoted "*smile*".

6.3 Signal dynamic instantiation

6.3.1 Computation of the dynamic instantiation coefficient

As seen in the previous section, the system obtains the name of the selected signal from the expression library, for a given modality and for a given expressivity. Now, it has to compute the dynamic instantiation to apply to this signal. This dynamic instantiation allows us to obtain the widest range of expressions and to modulate the expressivity. In our example, for an influence coefficient *contextCoeff* with a value of *0.4* instead of *0.5*, the system obtains a *behavioral expressivity* with a value of *0.48*. In this case, the system also uses the signal "*smile*", but *the dynamic instantiation* applied to the signals would have been able to make perceivable the difference.

To compute the dynamic instantiation *l'*, for a given *behavior expressivity e*, we consider:

- *l*: the distance between *ref* and *e*;
- *L*: the distance between *ref* and the appropriate boundary: *min* if *e* is inferior or equal to *ref*, *max* otherwise;
- *L'* the distance between *1* (i.e. the default coefficient, which indicates that the system has not to modify the signals stored in the library) and the coefficient related to the considered boundary (i.e. *coeffMin* or *coeffMax*).

The distance *l'* between the dynamic instantiation coefficient *dynCoeff* and *1* is such as that the ratio *l/L* is equal to the ratio *l'/L'*. Thus, $l' = L' * l/L$. The coefficient *distortCoeff* is superior to 1 iff *e* is superior to *ref*. In our example, for the signal "*smile*", as e_{joy} = *max*, we have *dynCoeff = coeffMax=1.5*.

For *contextCoeff* equal to *0.4*, we have said that the behavioral expressivity have a value of *0.48*. As the expressivity reference ref is *0.4* the *dynCoeff* is equal to *1.2* (i.e. *1+(1.5-1)*x(*0.48-0.4*)*/(0.6-0.4)*).

We consider that the evolution of the dynamic instantiation coefficient is linear between 1 (i.e. the default coefficient) and the extreme values, but not necessarily between the extreme values.

Given the behavioral profile and a specific meaning, the system computes the appropriate value of the signal dynamic instantiation using a fuzzy logic approach. We point out that we are dependent on the signals libraries content: for two different agents for which the system use the same modality and the same signals to express a given meaning or for a same agent in two different contexts but that use in the both cases the same modality and the same signals to express a given meaning, the difference between the *behavioral expressivities* may induce at the level of dynamic instantiation coefficient, and consequently at the level of the animation a difference imperceivable for a human being.

6.3.2 Dynamic instantiation types

Expressivity ought to be modelled differently depending on the modalities (face, gesture, gaze and posture) it applies to. We consider several types of dynamic instantiation: *temporal* (e.g. mutual gaze duration, duration of a raised eyebrow), *spatial* (e.g. facial muscular contractions, width of the arms aperture) or *repetition*. For facial expression, variation of expressivity can be expressed through variation of muscular contractions as well as variation of its temporal course; while when talking about gaze, expressivity variations may be related to factors such as length of mutual gaze or length of looking at the conversation partners; while when talking about gesture,

```
 1. <agent
 2.    nameAgent       = "Agent1"
 3. >
 4.   <signal name = "look_at"
 5.     coeffDistort = "1"
 6.     distortion = "temporal">
 7.       <rheme>
 8.         <signal name = "smile"
 9.           coeffDistort = "1.2"
10.           distortion = "spatial"
11.         >
12.            I'm happy.
13.         </signal>
14.       </rheme>
15.   </signal>
16. </agent>
```

Figure 10: Example of output

it may be related to parameters such as the strength of a movement, its tempo, its dynamism or its spatial amplitude. Variation of expressivity may also be expressed by the rapid repetition of the same gesture (rapid head nods, fast beat gestures). In our example, the signal "*smile*" is associated to an "*amplitude*" dynamic instantiation, that is the system accentuates by *20%* the facial movements. Conversely, the system modulates the expressivity of the signal "*look_at*" by varying its duration (see figure 10).

6.4 Output

From the input text, the system applies several modifications until to obtain a text where tags $< signal >$ replace the communication act tags. Each of these modifications corresponds to a level of influences integration. The tags $< signal >$ are not associated to *expressivity* any more, but to the attribute *dynCoeff*. The figure 10 presents the output that the system processes.

At the computation level, these modifications of XML texts are applied according to XSLT stylesheets (see figure 11). XSLT (e**X**tensible **S**tylesheet **L**anguage **T**ransformation) is a language for transforming XML documents into other XML documents (see XSLT). The first transformation computes the *behavior expressivities* from the *communicative expressivities* in order to individualize the behavior according to the agent. The various libraries (*expression libraries* and *signal libraries*) are specified for a given input text, as well as the agents' definition. The second transformation creates, from the communication act tags, signals tags that are directly exploitable by the animation engine. The algorithms described in the previous sections are implemented in the stylesheets.

The output text of the figure 10 is simplified. A non simplified output contains temporal information related the signals. For example, for the facial signals, several other attributes are defined. They may specify the time

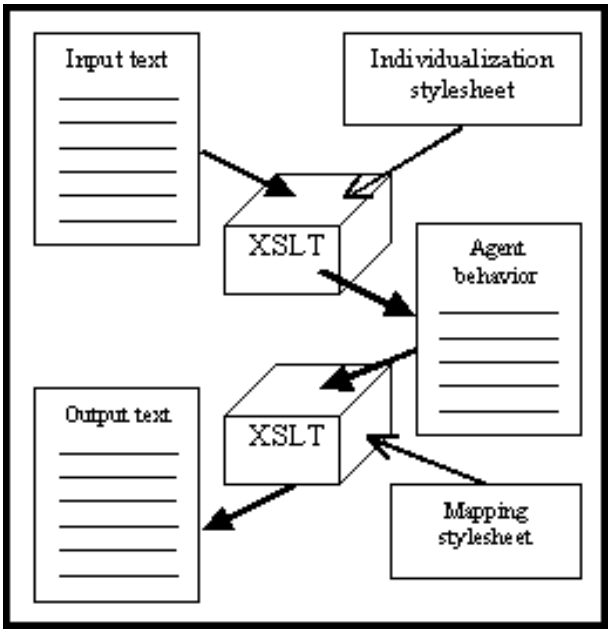


Figure 11: XSL Transformations

that the expression takes to reach its maximal intensity, the time during which the expression maintains its maximal intensity, the time that, starting from the maximal intensity, the expression changes into another expression, the time an expression waits until it raises, the time considered from the beginning of the tag, or the time an expression finishes to be shown before the end of the tag.

At the end, the system outputs a list of facial and body parameters that are used to drive the animation engine. To this end, we are using an MPEG-4 compliant animation engine, described in Pelachaud (2002).

7 Conclusion

We aim at creating an individual agent. Individuality is forged by several factors such as personality, social role, culture. Modelling such factors is extremely complex. To overcome this difficulty we propose to model influences by their impact they have on the behaviors expressivity. At first, we have described a taxonomy of influences as well as a set of parameters that characterize expressivity. The system is still being developed. We then foresee to do evaluation tests to validate the strategies for the intermodal expressivity. We also aim at testing the validity of the dynamic instantiation coefficient for each signal: we have to verify if adding expressivity does not create other meaning perceivable in the agent's behavior.

Acknowledgements

References

G. Ball and J. Breese. Emotion and personality in a conversational agent. In S. Prevost J. Cassell, J. Sullivan and E. Churchill, editors, *Embodied Conversational Characters*. MITpress, Cambridge, MA, 2000.

R Brislin. *Understanding culture's influence on behavior*. Hartcout Brace College Publishers, New-York, 1993.

C.M. Brown and G. Nichols-English. Dealing with patient diversity in pharmacy practice. *Drug Topics*, pages 61–69, September 1999.

The Duy Bui, D. Heylen, M. Poel, and A. Nijholt. Generation of facial expressions from emotion using a fuzzy rule based system. In D. Corbett & M. Brooks M. Stumptner, editor, *Proceedings of 14th Australian Joint Conference on Artificial Intelligence (AI 2001)*, pages 83 – 94, Adelaide, Australia, 2003. Springer.

V. Carofiglio, F. de Rosis, and R. Grassano. Dynamic models of mixed emotion activation. In L Canamero and R.Aylett, editors, *Animating Expressive Characters for Social Interactions*. John Benjamins, Amsterdam, in press.

J. Cassell. Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. In S. Prevost J. Cassell, J. Sullivan and E. Churchill, editors, *Embodied Conversational Characters*. MITpress, Cambridge, MA, 2000.

B. DeCarolis, V. Carofiglio, M. Bilvi, and C. Pelachaud. APML, a mark-up language for believable behavior generation. In *Embodied conversational agents - let's specify and evaluate them!, Proceedings of the AAMAS'02 workshop*, Bologna, Italy, July 2002.

B. DeCarolis, C. Pelachaud, I. Poggi, and F. de Rosis. Behavior planning for a reflexive agent. In *IJCAI'01*, Seattle, USA, August 2001.

B. DeCarolis, C. Pelachaud, I. Poggi, and M. Steedman. APML, a mark-up language for believable behavior generation. In H. Prendinger and M. Ishizuka, editors, *Life-like Characters. Tools, Affective Functions and Applications*, pages 65–85. Springer, 2004.

M.S. El-Nasr, J. Yen, and T. Loerger. FLAME - fuzzy logic adaptive model of emotions. *International Journal of Autonomous Agents and Multi-Agent Systems*, 3 (3):1–39, 2003.

H. McBreen H. Morton and M. Jack. Experimental evaluation of the use of ECAs in eCommerce applications. In Z. Ruttkay and C. Pelachaud, editors, *From Brows till Trust: Evaluating Embodied Conversational Agents*. Kluwer, To appear.

K. Höök. User-centred design and evaluation of affective interfaces. In Z. Ruttkay and C. Pelachaud, editors, *From Brows till Trust: Evaluating Embodied Conversational Agents*. Kluwer, To appear.

K. Isbister and C. Nass. Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies*, Forthcoming.

Scott A. King, Alistair Knott, and Brendan McCane. Language-driven nonverbal communication in a bilingual conversational agent. In *Proceedings of CASA 2003*, pages 17 – 22, 2003.

S. Kshirsagar, C. Joslin, W. Lee, and N. Magnant-Thalmann. Personalized face and speech communication over the internet. In *Proc. of IEEE Virtual Reality*, pages 37–44, Tokyo, Japan, 2001. IEEE Computer Society.

E.-J. Lee and C. Nass. Does the ethnicity of a computer agent matter? An experimental comparison of human-computer interaction and computer-mediated communication. In *WECC'98, The First Workshop on Embodied Conversational Characters*, October 1998.

S. Marsella, W.L. Johnson, and K. LaBore. Interactive pedagogical drama. In *Proceedings of the 4th International Conference on Autonomous Agents*, Barcelona, Spain, June 2000.

Stacy Marsella and Jonathan Gratch. Modeling coping behavior in virtual humans: Don't worry, be happy. In *proceedings of the /2nd International Conference on Autonomous Agents and Multiagent Systems*, Melbourne, Australia, 2003.

A. Paiva, G. Andersson, K. Höök, D. Mourao, M. Costa, and C. Martinho. SenToy in FantasyA: Designing an affective sympathetic interface to a computer game. *Journal of Personal and Ubiquitous Computing*, 6(5-6):378–389, 2003.

C. Pelachaud. Visual text-to-speech. In Igor S. Pandzic and Robert Forchheimer, editors, *MPEG4 Facial Animation - The standard, implementations and applications*. John Wiley & Sons, 2002.

C. Pelachaud and I. Poggi. Subtleties of facial expressions in embodied agents. *Journal of Visualization and Computer Animation*, 13:301–312, 2002.

P. Piwek, B. Krenn, M. Schröder, M. Grice, S. Baumann, and H. Pirker. RRL: a rich representation language for the description of agents behaviour in NECA. In *Embodied conversational agents - let's specify and evaluate them!, Proceedings of the AAMAS'02 workshop*, Bologna, Italy, July 2002.

I. Poggi. Mind markers. In N. Trigo M. Rector, I. Poggi, editor, *Gestures. Meaning and use.* University Fernando Pessoa Press, Oporto, Portugal, 2003.

I. Poggi, C. Pelachaud, and F. de Rosis. Eye communication in a conversational 3D synthetic agent. *AI Communications*, 13(3):169–181, 2000.

H. Prendinger, S. Descamps, and M. Ishizuka. Scripting affective communication with life-like characters in web-based interaction systems. *Applied Artificial Intelligence*, 16(7-8):519–553, 2002.

T. Rist and M. Schmitt. Applying socio-psychological concepts of cognitive consistency to negotiation dialog scenarios with embodied conversational characters. In *Proc. of AISB'02 Symposium on Animated Expressive Characters for Social Interactions*, pages 79–84, 2003.

Zs. Ruttkay, V. van Moppes, and H. Noot. The jovial, the reserved and the robot. In *proceedings of the AAMAS03 Ws on Embodied Conversational Characters as Individuals*, Melbourne, Australia, 2003.

VHML. **Virtual Human Markup Language**. http://www.vhml.org.

XSLT. **eXtensible Stylesheet Language Transformation**. http://www.w3.org/TR/xslt.

Developing a Virtual Ballet Dancer to Visualise Choreography

R. J. Neagle, K. Ng and R. A. Ruddle
*ICSRiM and School of Computing
The University of Leeds, England
{royce,kia,royr}@comp.leeds.ac.uk

Abstract

This paper describes the development of a virtual ballet dancer that is designed to help visualise choreography. Recreating dance not only involves how to achieve the steps but also the quality of the movement to express emotions. The aim is to create a visualisation system that could be used to understand the choreography with expressive movement when resurrecting ballet scores. Real-time computer graphics are ideally suited to bridge the gap between written choreographic notation and performance, via the creation of a virtual dancer. In theory, a realistic virtual performance can be driven from a machine-readable version of dance notation with a virtual dancer. This paper presents the setting and evaluation of key poses which form the foundation for ballet *steps* (combinatin of key poses) and how the application of Laban's Effort factors can be used for expressive interpolation between these poses.

1 Introduction

Like music, the choreographed movements that make up a dance performance can be written down, and the best known systems for doing so are Labanotation (Brown and Parker, 1984; Laban, 1966), Eshkol-Wachman (Eshkol and Wachmann, 1958) and Benesh notation (Brown and Parker, 1984; Benesh and Benesh, 1983). Choreography is primarily written down for archival purposes and to promote its dissemination to a wider audience. Unlike music, dance notation is not widely understood by dancers. There are few professional performers who can read written choreography let alone visualise the movements involved, and this represents a considerable barrier to the utility of choreography in its written form (see Figure 1).

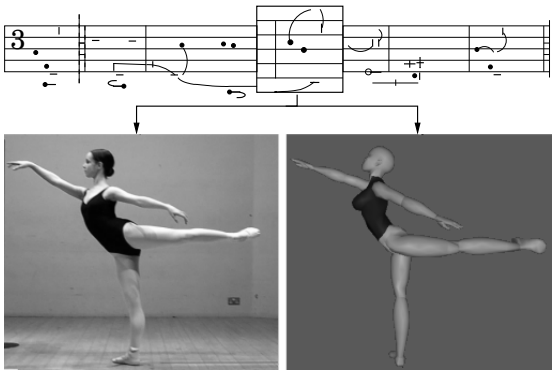


Figure 1: Example of a pose annotating using Benesh notation of an *arabesque* posed by a real dancer and virtual ballet dancer

Both the Labanotation and the Benesh notation give a rich vocabulary for describing human movement. They model structural and positional aspects of the performer at specific times and provide information on the timing and speed of movement, as well as qualitative aspects (i.e. how one moves). Benesh Movement Notation was designed by Rudolf and Joan Benesh in 1947 to represent classical ballet with its inferred rules such as turnout, rounded arms and straight legs. From the notation it is possible to reconstruct and visualise each pose and the movement required between the key poses.

Reconstructing dance using notation provides the foundation or blue print to the dance movement, the quantitive aspects. However, human activity is dynamically and rhythmically charged and structured, and people are recognised by the dynamic and rhythmic make-up which personalises their style of moving (Davies, 2003), the qualitative aspects. Our research is combining the two, and developing a virtual ballet dancer (VBD) where both aspects are required. The remainder of this paper starts by providing a brief background to the Benesh notation, the rules of classical ballet and Laban's dimensions of expressive movement. The machine-readable format used to input the Benesh notation is described followed by a detailed description and evaluation of the rules of ballet that were encoded into the VBD to allow it to adapt to the notated poses. The paper concludes by describing the methods that will be used to bring the VBD to life, animating its movement according to different motive themes in ballet.

2 Background

2.1 Benesh Movement Notation

In computer graphics, a skeleton-based approach is generally considered to be the most flexible way to animate

characters (Badler et al., 1999; Herda et al., 2000). Benesh notation provides information, directly or inferred, that can be mapped to an articulated skeletal structure to specify the positions of the bones or the angle of each joint of a dancer. The symbolic representation of dance notation is similar to music and with an underlying mathematical content that also lends itself well to machine representation.

Written from left to right on a five-line stave, the stave lines map to the height of a person's feet, knees, waist, shoulders and top of the head. To record a pose, the Benesh notation notes the exact locations occupied by the four extremities (the hands and feet). In addition, the position of a bend, such as the knee or elbow, may also be defined. Given these points and the body and head positions, it is possible to reconstruct and visualise the whole pose (see Figure 2).

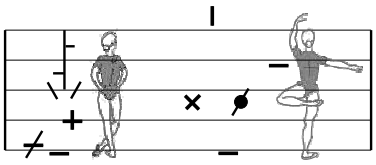


Figure 2: Examples of the Benesh notation showing signs to pose a dancer including orientation of the head and body, contact to the body (hand on hips), crossed positions of the extremities and different feet positions when in contact with the floor. The human figures show the actual poses that were defined.

Benesh also included floor patterns: direction, location and travel for the dancer or group of dancers written below the stave. Further details on travelling, direction, location can be found at page 87 in *Dance Notation for Beginners* (Brown and Parker, 1984) and page 70 in *Reading Dance* (Benesh and Benesh, 1983). Rhythm and phrasing are shown above the stave and include pulse beats, tempo and common dance rhythms. The pulse can be split into half, quarter and third beats and there are three methods for specifying the tempo: (1) set instructions, such as: Fast, Moderate, etc. (2) specifying the number of beats per minute by giving the pulse beat a metronomic speed; and (3) instructions as shown in music scores using Latin terminology, for example *Presto*, *Allegretto*, and *Adagio*.

For this paper, we are only concerned with the *in*-stave signs which provide the direction and orientation of the limbs, head, body and head either directly or by inferred ballet rules.

2.2 Ballet Rules

The basic rules of ballet were defined by Jean-Georges Noverre (1727-1810), the author of *"Lettres sur la Danse et sur les Ballets"* (Letters on Dance and Ballet) (Noverre, 1760), and Carlo Blasis (1797-1878), the author of *"The Code of Terpsichore"* (Blasis, 1830). The material in these books is virtually indistinguishable from ballet as it is taught to-day and specify many ballet rules. As ballet grows and choreographers define new positions and in the process create more exception to the rules. However, the basic rules remain unchanged and these rules will be used for the development of the current VBD system.

The ballet rules for the arms include rounded arms (with the exception of the *arabesque*) and arm orientation. The stance of the dancer is kept erect unless specified with the neck straight over the spine. The position of the legs always assumes a degree of turnout from the hip joint unless the notation specifies otherwise by defining the knee and feet positions.

This paper will be continually making reference to the following arm and leg ballet positions and the ballet rules required to pose them as described by the Cecchetti method. There are five principal positions of the arms:

1st Position. The hands are held at the side with the finger-tips near to the outside of the thigh.

2nd Position. The arms are extended to the side sloping downward and the position of the arms must not pass beyond the line of the shoulder.

3rd Position. One arm is placed in the fifth position, and the other placed at a slight distance from the side.

4th Position. One arm is placed in the second position and the other is in the fifth position, *en bas* (low), *en avant* (forwards) or *en haut* (high).

5th Position. Has both arms placed at shoulder width in front of the body. There are three derivatives of the fifth position: *en bas* (low), *en avant* (forwards), and *en haut* (high).

These five principal positions are posed with rounded arms, so that the point of the elbow is imperceptible. When the arms are to the side, the finger-tips of the hand or hands should be just within the range of vision.

For this research we will also consider the three principal positions for the arabesque line of the arms:

First Arabesque. The front arm is raised in line and above the shoulder line and the second arm is below and level or slightly behind the shoulder.

Second Arabesque. The front arm is in front and just above the shoulder line and the back arm is below and behind the shoulder line continuing the line of the front arm.

Third Arabesque. Both arms are in line with the shoulder and in front of the shoulder with one arm higher than the other.

This paper will also refer to five principal positions of the feet:

First position. Standing with heels together and toes turned out to the side.

Second position. Keeping the turnout established in First position, the heels are aligned under the shoulders.

Third position. Cross one foot to the middle of the other with hips centred equally over the feet and not twisted.

Fourth position. The feet are separated forward and back from either fifth or first position approximately one foot length with the weight evenly distributed between the feet.

Fifth position. The front heel crosses to the big toe joint of the back foot with hips centered over the feet and the weight equally distributed.

2.3 Expressive Movement

Analysis of expressivity in movement, especially dance movement, is not a simple task and there have been many approaches. Most of these use motion capture and then analyse the motion for patterns (Brand and Hertzmann, 2000; Price et al., 2000) or expressive cues (Camurri and Trocca, 2000). The accuracy of professional ballet dancer to position-match (Ramsay and Riddoch, 2001) and the constrained movement of ballet allowed Campbell and Bobick (1995) to use phased space constraints to recognise the different movements used in ballet.However, our research is interested in understanding the expressive layer placed over movement defined by ballet rules.

Laban's Effort and Shape theory (Dell, 1977) characterises the way in which people move. We are concerned with three particular factors: space, weight, and time. In EMOTE, Chi et al. (2000) gives excellent descriptions:

Weight: is the sense of impact of the movement and exertion required ranging from *light* (buoyant and delicate) to *heavy* (powerful with impact e.g. pushing or punching).

Time: describes the qualities of sustainment and quickness of movement as opposed to speed measured by the clock or tempo marked by a metronome. Time ranges from *sustained* (lingering and indulging in time) to *sudden* (agitated, jerky movements).

Space: is spatial focus and attention to surroundings, overlapping shifts in the body among a number of foci ranging from *indirect* (multi-focused) to *direct* (pinpointed and single focused).

Laban provides a fourth parameter called Flow. However, for our research this is superseded by the bounding imposed by the strict rules of classical ballet.

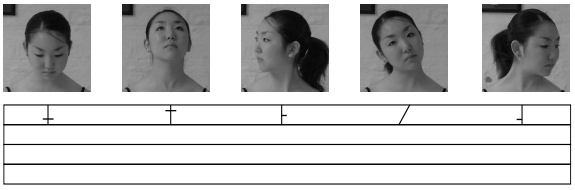
Experimental investigation of the fidelity required for showed that participants could correctly distinguish between different pairs of emotions in almost 80% of trials, even when the resolution, size, and display aspects of fidelity were reduced (Neagle et al., 2003). Participants performance remained largely unchanged when the video framerate of the experiments stimuli was reduced to as low as 5Hz. The results highlighted the time factor with other visual clues such as Laban's space and weight dimensions that participants used to make their judgements.

3 Inputting Key Poses

Virtual dance has developed over the last few years with three main approaches: (1) motion capture (Moeslund and Granum, 2001; Camurri and Coglio, 1998), (2) scripting (Badler et al., 1999), and (3) animations driven from machine-readable versions of dance notations (Badler, 1989; Neagle and Ng, 2003). Motion capture has mainly been used with contemporary dance with some development toward classical ballet including Stevens et al. (2002) and Camurri et al. (2000). Several programmes exist to create, store, and modify dance notations and include Benesh Notation Editor; MacBenesh; Calaban; LabanPad PDA (currently: Apple Newton); LabanWriter; and LED. There are limited visualisation tools and the best known research was LINTER (Herbison-Evans and Hall, 1989). However, while motion capture can provide many nuances in the movement, control of the animation is very limited and it is also labour intensive and costly. Movement of the VBD described in this paper are driven by a machine-readable version of the Benesh notation. The format ASCII text (Neagle et al., 2002) is parsed using a bespoke tokeniser. Examples of the format are shown below, and it is described more fully in Neagle et al. (2002).

The orientation states for the head, body and pelvis are any combination of *tilt*, *turn*, *bend* (see Figure 3 and 4), and orientations represent rotations around the three cardinal axes. Benesh notation specifies these states using the top three spaces in the five-line stave. We see Benesh notation, using single signs to represent combinations of orientation. Examples of head orientation together with the associated machine-readable ASCII text are shown in Figure 3. Example of body and pelvis orientation are shown in Figure 4

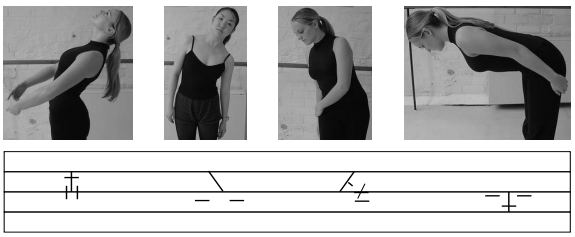
To record a position or pose, the Benesh notation also notes the exact locations occupied by the four extremities, the hand and feet, in relation to the body of the dancer on the coronal plane. The position of extremity signs within the stave for each key pose denotes the position of the dancers hands and feet in relations to their anthropometrics. Benesh notation provides the height and width of the extremities' position in relation to a dancer standing upright. There are five defined height positions represented by the five lines of the Benesh notation stave: the floor (height is zero); knee height; waist height; shoulder height; and the head height. Width is specified proportionally to the horizontal reach of the limb along the



(a) (b) (c) (d) (e)

```
frame 1{head(bend,forward)}                    #(a)
frame 1{head(bend,back)}                       #(b)
frame 1{head(turn,right)}                      #(c)
frame 1{head(tilt,right)}                      #(d)
frame 1{head(bend,forward)(turn,right)}        #(e)
```

Figure 3: Examples of the Benesh notation and machine-readable text format for head orientation. (a) and (b) are bends, (c) is a turn, (d) is a tilt, and (e) is a combination of a turn and a bend. NB. the notation is represent from behind the body and therefore the signs are notated in the opposite direction of the photos for the rotation and tilts.



(a) (b) (c) (d)

```
frame 1{                                #(a)
  body(bend,back)
  lefthand(0.1,3.0)front
  righthand(0.1,3.0)front
}
frame 1{                                #(b)
  body(tilt,left)
  lefthand(0.1,2.75)level
  righthand(0.1,2.75)level
}
frame 1{                                #(c)
   body(tilt,right)(turn,right)
  lefthand(-0.1,3.0)contact
  righthand(0.1,2.75)level
}
frame 1{                                #(d)
  pelvis(bend,forward)
  lefthand(0.1,2.75)level
  righthand(0.1,2.75)level
}
```

Figure 4: Orientation of the Benesh notation body and pelvis signs. (a) is a bend of the body above the waist, (b) a tilt, and (c) a turn and tilt. Images (d) shows a bend of the body below the waist (pelvis).

cardinal plane. For example, a width specification for the hand of half way from the centre of the body to the maximum reach would position the hand approximately where the elbow is.

The depth of an extremity is inferred from its position mapped onto the coronal plane plane and the orientation of the relevant limb(s). Different sets of signs are used to add depth information to the two-dimensional information provided by the position of a sign on the stave. These sets of signs are: front (| , + , ⁄), level (— , ⊥ , ⊿) or behind (• , × , ⊿) the body. The distance in front or behind the body is not specified as only one place is possible due to the fixed limb length. It is also not required to specify how bent the limb is, as this is governed by the position of the extremity and the position of the bent joint. The legs are positioned in a similar manner to the arms with the exception that the floor contact is specified. Unless a bent knee position is notated to define the orientation of the legs, the default turnout rule is applied (a rotation in the hip socket for the patella bone to face the side).

4 Defining Key Poses of a VBD

To teach dancers the movements prescribed by a particular piece of ballet notation, choreographers combine the information provided by the notation with their knowledge of the rules of ballet. The notation defines the positions of the key elements of a dancer's body, who then adopts a pose consistent with those positions and the standard rules of ballet. To define key poses for a VBD, the rules of ballet need to be expressed as a set of mathematical equations which were merged with the machine-readable Benesh notation data.

The VBD used in this research was the Cal3D software (Heidelberger, 2001) and has the degrees of freedoms (DOF) shown in Figure 5. The following sections explain how the information contained in Benesh notation can be combined with equations that capture the rules of ballet, to define the DOFs of a VBD's body. The centre of

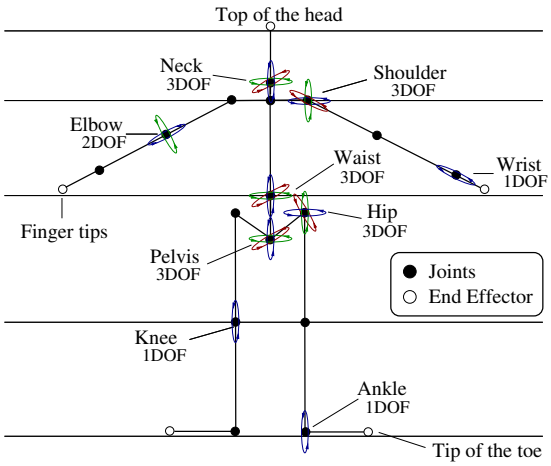


Figure 5: The skeletal structure of the VBD, showing the end effectors and the DOFs at each joint.

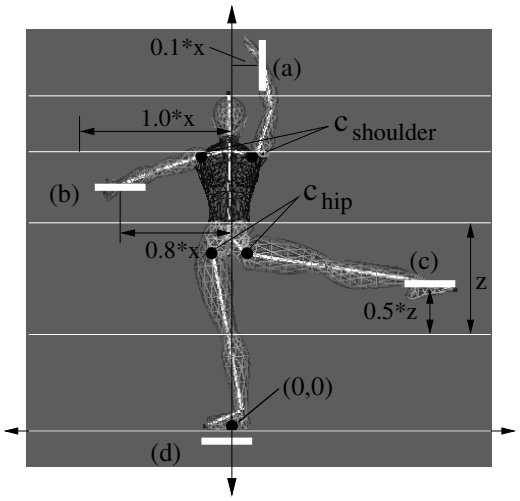
body nodes (pelvis, waist and neck), the shoulder nodes and the hips have three DOFs. The elbow nodes have two to: (1) bend the elbow; and (2) rotate the forearm to orientate the palm of the hand. The wrist nodes have one DOF to raise and lower the hands when setting rounded arms.

The knee nodes have one DOF to flex and straighten the leg and the ankle nodes have one DOF to flex and point the foot.

4.1 Setting Joint Positions

Before calculating positions the VBD is loaded in a default standing pose. The height (z) values of the knee, waist, shoulder and head are calculated from the VBD's bone structure and set as the stave heights. For example, the waist height z_{waist} is stored as the third stave line height. The maximum width for positioning the hands and legs are taken form the sum of the bone lengths that make up the limbs for those extremities.

Positions are explicitly notated in terms of height (z), width (x), and depth (y). Figure 6 demonstrates the position of the extremities. The file format representation




```
frame 1{
   righthand(0.1,5.5)front #(a)
   lefthand(0.8,3.5)level  #(b)
   rightfoot(0.9,2.5)level #(c)
   leftfoot(0.1,0.75)level #(d)
}
```

Figure 6: An example of Benesh notation, the file format representation and the VBD posed inputed from the file.

in this example specifies the position of each extremity as a tuple and a word, e.g. level, that defines the depth. The width parameter, though redundant and unused for calculating the position for extremities which are level, it is however required in the calculation of the ballet rules. The right foot, in Figure 6, is notated half way between the second and third stave line. The application therefore obtains the height of the second stave (knee height of the VBD) and the height of the third stave line (waist height) and calculates the height of the foot as:

$$z_{position} = z_{stave2} + p(z_{stave3} - z_{stave2}) \qquad (1)$$

where p is the proportion value (0.5).

Given the value of the height, the end effector can be calculated. When the appendage is in the body plane, this is a simple 2D problem using Pythagoras. Given the length of the limb (l), the height ($z_{position}$), and the rotation point of the extremity i.e. the hip or shoulder joint in absolute space ($c_{(x,y,z)}$, $c_{shoulder}$ or c_{hip} in Figure 6) the width position is calculated as:

$$x_{position} = \sqrt{l^2 - (z_{position} - \underline{c}_z)^2} + \underline{c}_x \qquad (2)$$

The position of the extremity (in this case the right ankle joint) is set as $(x_{position}, c_{hip\ y}, z_{position})$.

The above approach is easily extended into three-dimensional space for positioning the extremities in front or behind the body plane. The width (x) is calculated proportional to the length of the straight limb. For example if the arm span of the VBD is 1 metre, a notation width parameter 0.1 defines the hand to be 0.1m from the shoulder (see Figure 6, the raised arm above the head is notated with width 0.1 and height half way between the top of the head and a raised stretched arm (5.5)). Once the $x_{position}$ and the $z_{position}$ has been calculated (see above), the depth ($y_{position}$) is calculated using Pythagoras:

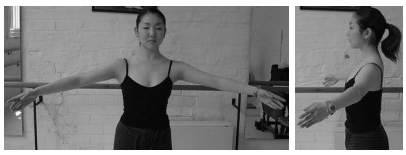
$$y = \pm\sqrt{l^2 - (x^2_{position} - \underline{c}_x) - (z^2 - \underline{c}_z)} + \underline{c}_y \qquad (3)$$

where $y > c_y$ is in front of the coronal plane and $y < c_y$ is behind the coronal plane.

4.2 Rounded Arms and Elbow Positions

Like the Benesh notation, the VBD will assume the arms are rounded unless elbow joints are specified or the position is identified as one of the three *arabesque* positions where the elbow and wrist rotation is set to $0°$. The rounded arm is produced by a slight bend of the elbow and wrist, examples are shown in Figure 7. The amount of curvature varies depending on the ballet method and teachers preferences. The prototype therefore is designed for the user to specify the amount curvature. Screenshots in this paper with rounded arms were set at $154°$ based on the first authors professional opinion. The amount of rotation affects only the depth position of the wrist and the height and width values are obtained as described in §4.1 with the length l in Equations 1–3 calculated using trigonometry given the fixed length of the upper and lower limbs and the specified rounded angle. Once the wrist position is obtained, the vector u is the vector from the position of the shoulder joint to the position of the wrist joint, see Figure 8.

The position of the elbow in classical ballet is on a plane defined by the vector from the shoulder to the wrist u and a vector parallel to the floor w. The ballet rule states the elbow should not be dropped down or raised too high but continue the line of the slope to the hands. The position of the elbow joint is calculated using sphere intersection to obtain a point (p_{inner}) on the vector (u) and Pythagoras to translate the p_{inner} along a vector (v)



(a) *secondé* position

(b) *pirouette* position of the arms

Figure 7: (a) *a la secondé* demonstrates rounded arms with the hand brought forward within the dancers line of sight. (b) *pirouette* position of the arms demonstrating the rounded arms the elbow position on a plane continuing the line from the shoulder to the wrist.

which is on the plane and parallel to the floor from p_{inner} (see Figure 8)

$$p_{inner} = p_{shoulder} + \left(\frac{|u|^2 - l^2_{upper} + l^2_{lower}}{2|u|^2}\right) u \quad (4)$$

The vectors are normalised to unit vectors and the cross product of w and u provides n and taking the cross product of n and u provides the vector v. The elbow position is calculated as:

$$\begin{aligned} d &= \sqrt{l^2_{upper} - (p_{shoulder} - p_{inner})^2} \\ p_{elbow} &= p_{inner} + dv \end{aligned} \quad (5)$$

where d is a scalar and the magnitude required to position the elbow from p_{inner} in the direction of vector v.

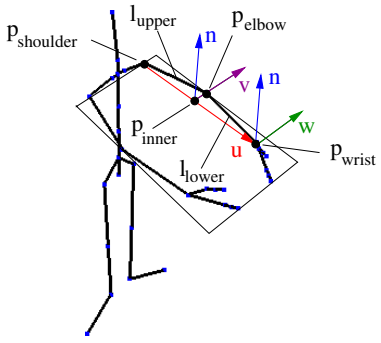


Figure 8: Vectors, points and limb lengths used to calculate the position of the elbow on a plane defined by the vector u (shoulder to wrist) and w (parallel to the floor).

4.3 Specified Elbow Position and Rotation

Benesh notation can override the ballet rules by specifying the elbow and wrist positions individually. The elbow position is determined as described earlier (Equations 1–3), given the width and the height of the elbow, the depth is calculated in relation to the shoulder position. The notation also provides the width and height of the wrist position and the depth direction. Using the elbow as the rotation point the wrist positions coordinates are calculated and the amount of elbow rotation is calculated from the three coordinates, (u, v, and w) as:

$$\theta = \cos^{-1}\left(\frac{(u-v)\cdot(w-v)}{|u-v||w-v|}\right) \quad (6)$$

where u is the shoulder coordinates; v, the elbow coordinates; and w, the wrist coordinates.

4.4 Arm Orientation

For poses with bent arms, the z-axis of the shoulder and elbow is orientated perpendicular to the plane that contains the lower and upper arm. Vectors running along the length of the upper and lower arm are used to calculate the orientation of the shoulder. Arabesque poses have straight arms, and so require a slightly modified procedure (see Figure 9). Once calculated, the vectors are set in a rotation matrix which for the model used with the Cal3D libraries in the VBD is:

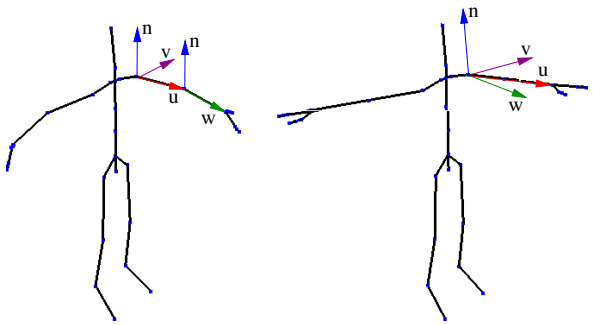


Figure 9: The left diagram demonstrates the orientation of the shoulder determined using the upper and lower arm vector. The cross product of w and u provides n where u is the x-axis and n is the z-axis. The y-axis v is obtained from the cross product of n and u. The right diagram demonstrates *arabesque* with straight arms where a vector (w) parallel to the floor and not parallel with u is substituted for the lower arm vector. All of the vectors are unit vector.

$$r_{matrix} = \begin{pmatrix} u_x & u_y & u_z \\ v_x & v_y & v_z \\ n_x & n_y & n_z \end{pmatrix}$$

The matrix is converted to a quaternion to set the rotation.

The rotation of the wrist has been simplified and for rounded arms uses the same rotation angle calculated for

the rounded elbow and rotated around its one DOF. The orientation for rounded arms to orientate the palms to face the correct direction is calculated in the elbow rotation rotated at 90° around th axis running along the lower arm bone. For the arabesque pose, the wrist rotation and elbow rotation are set as the identity quaternion.

4.5 Head, Body and Pelvis Rotations

The Benesh notation defines three rotations around each axis: *bend*, a *x-axis* rotation, *tilt* a *y-axis* rotation, and *turn*, a *z-axis* rotation where for the head bone of the VBD the x-axis is left to right; y-axis is front to back; and the z-axis is bottom to top. For this report we will use this coordinate system and the Euler angles are $(r_{bend},\ r_{tilt},\ r_{turn})$. The VBD currently calculates the rotations to the limit for that particular DOF. Each joint DOF limit have been selected based on the authors judgement within the range specified in Grosso et al. (1987) based upon the NASA Man–System Integration Standard Manual in *Occupational Biomechanics* by D. B. Chaffin.

The rotation limits are used specifically for setting key poses and the amount a rotation is variable depending on the expressive movement parameters to be devised for the next stage of research. The rotations of the centre joints as discussed in §3 is simple achieved by using quaternion multiplication where the quaternion is calculated for each Euler angle. The order of multiplication is important and rotations of 0° are set as identity quaternions.

$$q_{rotation} = q_{turn} \times q_{tilt} \times q_{bend} \tag{7}$$

4.6 Rules for the Lower Extremities

Positioning the legs follows the same rules as the arms replacing the rules for calculating rounded and straight arms with turnout and floor position. Turnout of the legs in the hip socket is one of the fundamental rules for classical ballet. The perfect dancer is aiming to have 90° rotation. However because few dancers have perfect turnout. The applications allows for the user to specify the turnout for the VBD up to a maximum value of 90°. The rotation is around the axis of the bone

Benesh notation specifies the position of the feet on the floor. Using the width parameter in the depth direction it is possible to determine if the feet are posed in one of the five basic feet positions, is a supporting foot or a position not defined by the ballet rules. Ballet rules in the application set the position of the feet for the basic positions and if a supporting leg is specified (see Figure 6, left leg). The width position of the feet and the depth provided by the Benesh notation and using Equation 3, all other feet positions can be calculated as a vertical distance from the hip joint. Unless the foot position is in a line from the hip to the floor, and the calculation is taken from the hip joint, the height of the foot position will not be connected to the floor. A vertical translation of the parent node is calculated to reset the foot to the floor.

5 Evaluation of Key Poses

5.1 General Methodology

The evaluation compared key poses of the VBD, as determined by combining machine-readable Benesh notation with the rules of ballet programmed into the VBD application, with corresponding poses described in *The Manual* (Beaumont and Idzikowski, 1977). The images from The Manual were used by permission of the Imperial Society of Teachers of Dance (ISTD). Images of each pose taken from the text book were set adjacent to the image screen grabbed from the VBD and only differences highlighted for comparison. The poses were selected to demonstrate the use of the different rules used by the VBD.

5.2 General Pose Descrepancies

The VBD uses the human model provided with the Cal3D libraries. Although this is an excellent starting point there are some discrepancies that need to be addressed. The most obvious is the shape and proportions of the Cal3D model in respect to the real-world ballet dancers used in this research. These includes, most noticeably, the breast size and the shape of the legs. Figure 11 shows the VBD has noticeably curved thigh muscles and a ‘s’ shape to the leg, whereas ballet defines this shape as a straight leg. The three major classical ballet factors which are less obvious to non-professionals are the orientation of the head; the shape of the hands; and there being no toe joint.

Head Orientation

Blasis states *“Take especial care to acquire perpendicularity and an exact equilibrium”* (Blasis, 1830). Figure 10(a) shows the spine in the head is sloped slightly backward and is not perpendicular to the ground and the mesh is both forward and down . For a classical ballet stance as described by Blasis, the head should be perpendicular to the ground and the centre of the mesh aligned so there is a sense of equilibrium. Currently the VBD appears top heavy and forward.

The Hands, Legs and Feet

The Manual describes the shape of the hands for classical ballet and provides variations in the positions of the fingers for different positions. These variations are minor and therefore for the VBD, a single classical shape would would suffice. However the Cal3D model currently used has an open hand which is incorrect for a classical dancer, as shown in Figure 10(b). This incorrect hand pose has been highlighted in both evaluated poses. See Figure 12, difference (c) and Figure 13, difference (c).

The pointed foot currently is an issue when the foot is lifted off the ground or defined touching the ground on full point (*sur la pointe*). This like the hands is a theme

(a) (b) (c)

Figure 10: (a) The VBD with skeleton and wire mesh surface display, with the head positioned in the default anatomical position. (b) A classical hand position as defined in *The Manual*, Plate IV Fig. 17 (left) and the open palmed VBD hand (right). (c) The pointed foot extended as much as possible with the instep forced well outwards and the *pointe* forced downwards taken from Plate III, Fig 11 (left) and the VBD without a toe joint showning the rigid foot with the *pointe* created with only the ankle rotation.

that will run through every pose when a pointed foot is required. To have every possible foot pose requires the toe joint to be added to the skeleton which is currently not part of the Cal3D model (see Figure 10(c)). Because the shape of the foot remains unchanged from flat to pointed, when a pointed foot is required the shape is noticeably rotated only at the ankle joint. See Figure 13, difference (e) on all images. The basic feet positions have minor errors as shown in Figure 11 for the *first* to *fourth* position and Figure 12 for the *fifth* position.

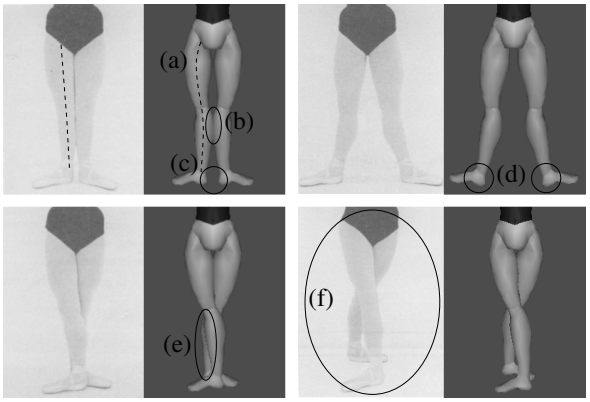


Figure 11: Positions of the feet from Plate I of The Manual (Beaumont and Idzikowski, 1977) and VBD in the same pose . The poses are from left to right: *first, second, third,* and *fourth croisée* position. Highlighted differences: (a) straight legs on the VBD appears 's' shaped (b) no mesh deformation resulting in intersecting mesh; (c) distance between heels due to leg shape of the VBD model; (d) ankle rotation is currently incorrect; (e) back leg can be viewed due to leg shape of the model; and (f) the real-world dancer has been photographed with the hip orientation off centre creating a slight twist in the shape.

Both the reshaping of the hands and skeletal change in the feet are required to add a greater level of fidelity to the classical pose. However, for this research, the fidelity of the dancer is high enough to demonstrate classical ballet poses that professionals and non-professional can recognise and/or compare to real-world examples.

5.3 Standard Poses in Fifth Position

The poses that were evaluated combined of poses position the feet in *fifth* position (the corner stone of ballet) and a combination of different arm and body positions based on the five basic ballet positions and their derivatives that were outlined in §2.2. The example in Figure 12 is slightly lower than normal as the Manual is also describing the pose from the *First* Exercise on the *Port de Bras*, No. 2.

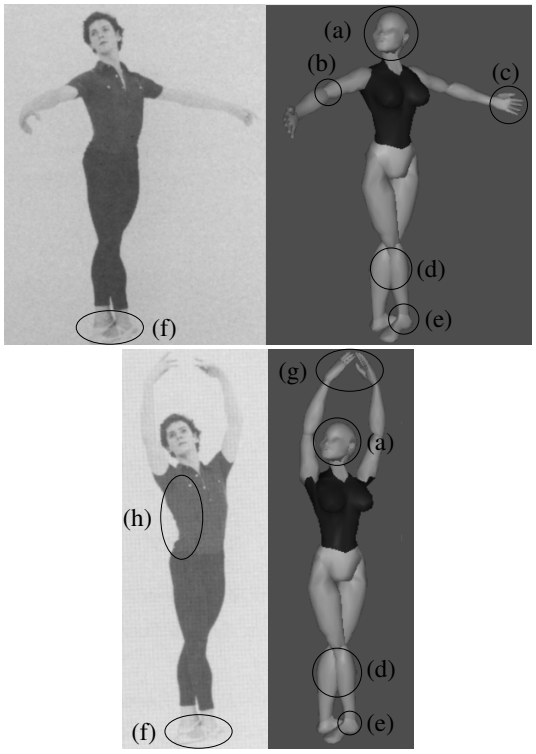


Figure 12: Pose from The Manual (Beaumont and Idzikowski, 1977) and VBD in the same pose . The poses are: *fifth position* with arms *a la seconde* (left) and *fifth position* with arms *fifth en haut* (right). Highlighted differences: (a) head turn and incline combination; (b) amount of elbow bend for rounded arms and orientation; (c) classical hand position; (d) collision detection and mesh deformation; (e) ankle rotation (toe position is slightly below floor level); (f) twisting of the joints to squeeze a better fifth position (common in posed photographs); and (g) width of hand placement.

Comparing the real-world pose with the VBD in Figure 12, the following differences were identified:

(a) Current coding of the combination head rotations are currently incorrect. The assigned values for the rota-

tions are set as a value not taking into consideration the arm positions and the authors at the time of coding misunderstood the different variation in the Benesh notation and the affect the combinations have on the orientation of the head. The orientation therefore is currently only a close approximation.

(b) The current elbow bend and orientation specified by the authors professional opinion varies from the manual. The value for the bend is currently set from measurements taken from the authors own pose of the rounded arm as discussed earlier. The orientation is set to 90° around the bone axis. From Figure 12 we observe the VBD's lower arm orientation is greater than the real-world dancer. However the amount of rotation to orientate the lower arm varies between dance methods and the dancers themselves and other poses of the same arm position taken from The Manual demonstrate a greater rotation.

(c) The pose of the hands has been discussed earlier (see §5.2)

(d) The libraries used to create the VBD currently have no collision detection and therefore deformations of the mesh are only related to rotations around the joint. The real-world dancer's calf muscles have been compressed as we see that the legs are still straight (i.e. no bend at the knee joint). When the VBD is posed in the same position, the meshes intersect. Correcting this is beyond the scope of this research.

(e) The VBD is currently not being placed in a virtual environment and therefore no compensation of the ankle rotation has been coded to check if parts of the anatomy intersect the floor plane.

(f) The real-world dancer is using the pressure of the floor to have what is termed a tighter fifth position (toes are pressed against the heel of the other foot). This is noticeable on the real-world dancer as the direction of the hips (to the corner) do no match the direction of the feet (direction closer to facing the front). Current teaching tries to avoid this, however, when asked to pose for a still photo, most professional dancers will use floor pressure to rotate the ankle and knee joints incorrectly to create a tighter position. The Manual demonstrates the *fifth* position of the feet which the VBD maps accurately with respect to the defined turnout.

(g) The hands of the VBD are too close together. Increasing the separation will allow the elbow bend rotation to increase to visually make a more rounded shape.

(h) The hint of the back bend shown by the real world dancer is a level of fidelity that the VBD does not achieve. The slight bend is a part of the Cecchetti style and the image is a pose from the *Third* Exercise of *Port de Bras*. This variation in shape is hoped to be achieved from the movement algorithm discussed in the next chapter.

5.4 The *Arabesque* Poses

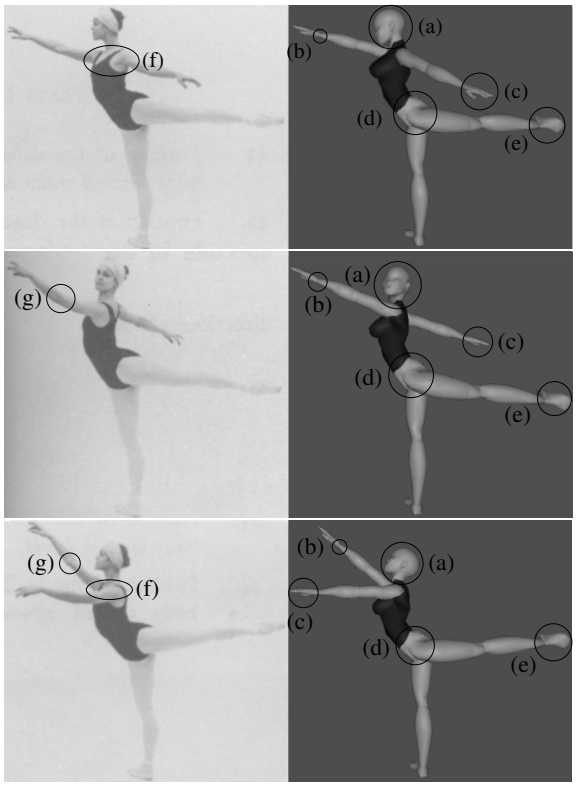


Figure 13: Poses from The Manual (Beaumont and Idzikowski, 1977) (left) and VBD in the same pose (right). The poses are: *first arabesque* from Plate VIII, Fig. 36 (top); *second arabesque*, Plate VIII, Fig. 37 (middle); and *third arabesque*, Plate VIII, Fig. 38 (bottom). Highlighted differences are: (a) head descrepencies; (b) break in the wrist joint; (c) classical hand position; (d) pelvis compensation for turnout; (e) classical foot shape; (f) rotation of the shoulders; and (g) slight break in the elbow to create a softer appearance (common with the female dancer).

The *arabeque* pose is one of the most used in classical ballet choreography and therefore has been selected as our second evaluated pose. As in the earlier section the visually noticeable differences have been highlighted:

(a) This is a recurring theme and was discussed in the previous evaluated pose. An extension to the problem can be seen in the bottom pose of Figure 13 where the head line (direction of the head) is slightly raised to look at the top hand. Whether a choreologist would define a head back position as seen in the VBD of the same pose or leave it undefined is

at this stage unsure. If undefined, a new level of fidelity to the ballet rules would need to be added to compensate 'looking at the raised front hand'.

(b) Most classical dancers place a slight break in the wrist to create the illusion of a softer position. The amount of break varies from dancer to dancer and therefore the VBD was coded without. When compared with the real-world pose however, the VBD's arms appear very rigid and stiff would be corrected by a dance teacher.

(c) The pose of the hands has been discussed earlier, see §5.2.

(d) The rotation of the hips by the real-world dancer is used to provide the required turnout of the raised leg. Depending on the dancers body this can vary a great deal. Dancers spend years of training to minimise the amount of twist required. The VBD is therefore in a technically correct but unrealistic pose in the sense that most professional dancers will use some amount of hip rotation to create a better leg line.

(e) We see that the bottom half of the foot does not match the real-world pose. This is due to the toes of the VBD not having the functionality to be pointed. See §5.2 for the discussion on the pose of the feet.

(f) The real-world dancer has rotated the shoulder. Though, technically, students are trained to be in what is classified as a square position (both shoulders facing the direction and not a corner), many professionals rotate the shoulders to create a better arm line and to make it easier to raise the hips to create better turnout of the legs. Because of the different teaching methods and the technical description of an *arabesque* pose the VBD will keep its shoulders square.

(g) The real-world dancer in this pose, arguably, has too much bend in the elbow joint. It is of the authors professional opinion that the arabesque pose has straight or nearly straight (if creating a softer appearance) arms. The amount of bend is difficult to assign and varies between dancers and therefore the VBD currently sets the arabesque pose with straight arms.

6 Emotive Animation

Ballet dancing not only requires technique to perform the poses and movement, but expressiveness to create a performance and not a series of steps. Benesh notation provides movement arcs to define the path of the extremities between poses and general movement descriptors to orientate (direction of the body) and position (height of the *saute*/jump or *plié*/bend) the dancer. Using Laban's Effort parameters it is possible to define variations in the interpolation that are distinguishable to the audience. The Effort factor are ranged between the two extreme values associated with each effort to capture the many different expressive movement.

Time is not the speed of the movement between key poses but how the speed of movement varies. Given the two extreme parameters are sustained and sudden, an equivalent concept is in music where sustained is playing a long consistent note and sudden is an accented note with a quieter sound afterwards. A movement equivalent is shown in Figure 14. The attack segment of the movement will require and acceleration and deceleration period at the beginning and end of the attack period as motor motion does not have jumps in velocity.

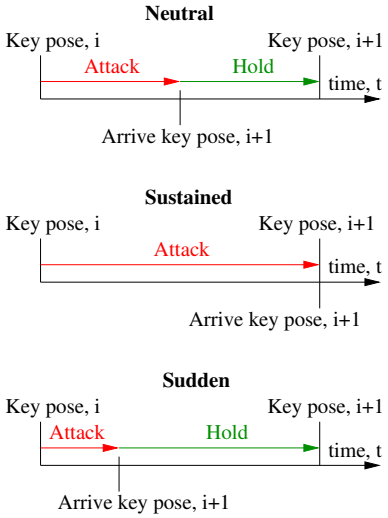


Figure 14: Variations on Laban's time factor for interpolation

Laban's space factor ranges between direct and indirect. For classical ballet a different application toward movement in space is required when looking at arms, legs and the spinal orientation. The main cue for expressive movement comes from the use of the arms in classical ballet. A direct path is not a straight line but the shortest path between key poses. Ballet movement is based on a circular geometry and Benesh notation infers that the shortest most direct path will be taken by the extremity unless specified otherwise (see Figure 15). By defining movement rules for space that deviates away from the inferred movement path within the bounds of the ballet rules, the VBD arm movement will utilise more space around the body and therefore more indirect movement. Currently, the VBD poses the centre joints (head, waist and pelvis) rotated at their respective maximum joint limits for each DOF. A direct specification rotates only around the axis specified by the notation and an indirect specification specifies rotations with a minor value to the other two axes. In classical ballet the legs not only achieve ballet poses but is fundamentally to supporting the dancer and must be considered when applying the expressive layer. The movement of the legs is therefore more bounded to

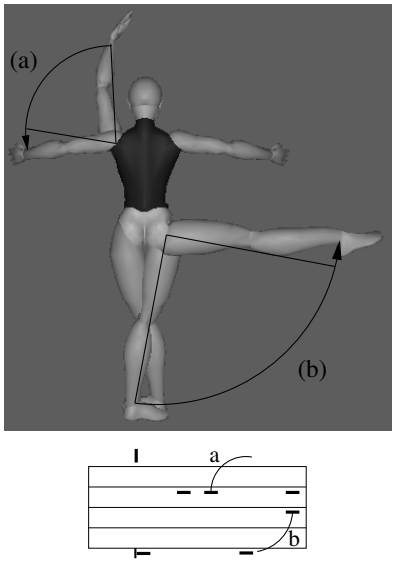


Figure 15: Example of the shortest path from a pose with feet in *fifth* position and arms in *fourth en haut* to standing *a la seconde*. Variation from the specified path leads towards indirect movement. Unspecified movement paths will always assume the shortest ballet movement path.

keep the dancer balanced and supported. Altering the position of the supporting legs as specified in the notation could create unbalanced poses and movement a real dancer is physically unable to achieve. For this research the space factor will only affect the raised leg movement and the path taken within the the strict bounds of the classical ballet rules. For example variations from the movement path can not bend the knee unless specified in the notation as this breaks the rules of classical ballet and limits the amount of spacial variation compared to the arms.

The final factor is the weight of the movement, described by Laban as ranging from heavy to light. The major cues for distinguishing differences in the weight are from the arms and legs. In classical ballet, the torso is strongly held and a strong abdomen is required for a dancer to have adequate technique to perform ballet steps. Variations of the weight effort would therefore be minimal for the spinal joints and we will focus on the extremities. Weighty dancing can be distinguished by the amount of *plié* used by the dancer. The greater the knee bend the more weighty (heavier) the feel. For the arms the approach is different. How the arm moves between key poses provides cues to the audience on the weight of the movement being performed. The most observable joint is the wrist rotation. For the wrists Blasis proposed that: *"There are two methods of moving the wrists, upwards and downwards. When the movement is to be made downwards, the wrist must be bent inwards, moving the hand demi-circularly, by which movement the hand returns to its first position; but care must be taken not to bend the wrist too violently, for it would then appear as if broken. With respect to the second movement, which is upwards, the wrist must be bent in a rounded position, allowing the hand to turn upwards, making a demi-tour or half-turn*" (Blasis, 1830). A light weight would have more flex in the wrist joint creating an illusion of a feather movement similar to the wrist action when painting with a brush. A heavy movement will have little rotation of the wrist giving a strong more solid and therefore more powerful feel to the movement.

The motivation of this project is to provide a system that will aide the realisation of a notated piece of dance into a live performance during rehearsals. A system that can simulate performed dance sequences as required by the choreographer or choreologist using the key poses as defined by the notation. The full development of a software animation system to represent and simulate dance would be beneficial to historians, choreographers and choreologists to: (a) evaluate ballet choreography with expressive styles, and (b) aide professionals to visualise the movement required for resurrecting ballet scores. Historically, dance provides a medium for open expression and conveying feelings revealing inner thought. The final system hopes, by synthesising expressive dance movement using VE technologies, to lead to better understanding of how animation provides visual cues to the virtual characters feelings while performing a very specific constrained rule based task.

Acknowledgements

A special thanks to Jon Singleton and Linda Pilkington from the I.S.T.D. for granting me permission to used images from *The Manual* for evaluation. Also, to Kate Simmons at Kate Simmons Dance Ltd. for allowing me to photograph their full-time professional dance students in ballet poses required for the VBD.

References

N. Badler. *A Computational Alternative to Effort Notation*, chapter 3, pages 23–44. National Dance Association, 1989.

N. Badler, C. Phillips, and B. Webber. *Simulating Humans: Computer Graphics, Animation and Control*. Oxford University Press, 1999.

C. Beaumont and S. Idzikowski. *A Manual of The Theory and Practice of Classical Theatrical Dancing*. Beaumont Publication, revised edition, 1977. Manual for the Cecchetti Method of Classical Ballet, Preface by Maestro Cav. Enrico Cecchetti.

R. Benesh and J. Benesh. *Reading Dance: The Birth of Choreology*. McGraw-Hill Book Company Ltd, first edition, 1983.

C. Blasis. *The Code of Terpsichore*. London, E. Bull, 1830. The Code of Terpsichore. The art of dancing,

comprising its theory and practice, and a history of its rise and progress, from the earliest times by C. Blasis. Translated under the author's immediate inspection by R. Barton.

M. Brand and A. Hertzmann. Style machines. In Kurt Akeley, editor, *Siggraph 2000, Computer Graphics Proceedings*, pages 183–192. ACM Press / ACM SIGGRAPH / Addison Wesley Longman, 2000.

A. Kipling Brown and M. Parker. *Dance Notation for Beginners*. Dance Books Ltd, 1984.

L. W. Campbell and A. F. Bobick. Recognition of human body motion using phase space constraints. In *ICCV*, pages 624–630, 1995.

A. Camurri and A. Coglio. An architecture for emotional agents. *IEEE Multimedia Journal*, 5(4):24–33, December 1998.

A. Camurri, S. Hashimoto, M. Ricchetti, A. Ricci, K. Suzuki, R. Trocca, and G. Volpe. Eyesweb - toward gesture and affect recognition in dance/music interactive systems. *Computer Music Journal*, 24(1):57–69, 2000.

A. Camurri and R. Trocca. Analysis of expressivity in movement and dance. In *Colloquium on Musical Informatics Proceedings*, July 2000.

D. Chi, M. Costa, L. Zhao, and N. Badler. The EMOTE model for effort and shape. In Kurt Akeley, editor, *Siggraph 2000, Computer Graphics Proceedings*, pages 173–182. ACM Press / ACM SIGGRAPH / Addison Wesley Longman, 2000.

M. Davies. *Movement and Dance in Early Childhood*. Paul Chapman Publishing, second edition, 2003.

C. Dell. *A Primer for Movement Description*. Dance Notation Bureau, Inc, forth edition, 1977.

N. Eshkol and A. Wachmann. *Movement Notation*. Weidenfeld and Nicolson, first edition, 1958.

M. Grosso, R. Quach, E. Otani, J. Zhao, S. Wei, P. Ho, J. Lu, and N. Badler. Anthropometry for computer graphics human figures. Technical Report MS-CIS-87-71, University of Pennsylvania, 1987.

B. Heidelberger. B/e/o/s/i/l: Cal3d, a free 3d animation library. `http://cal3d.sourceforge.net/`, 2001. Last accessed February 2004.

D. Herbison-Evans and N. Hall. The computer interpretation of classical ballet terminology. *Technical Report*, (TR364), 1989.

L. Herda, P. Fua, R. Plänkers, D., R. Boulic, and D. Thalmann. Skeleton-based motion capture for robust reconstruction of human motion. In *Computer Animation*, Philadelphia, PA, May 2000.

R. Laban. *Choreutics*. MacDonald and Evans Ltd, second edition, 1966.

T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding: CVIU*, 81(3):231–268, 2001.

R. J. Neagle, Ng K., and R. A. Ruddle. Notation and 3D Animation of Dance Movement. In *Proceedings of the International Computer Music Conference (ICMC2002)*, pages 459–462, September 2002.

R. J. Neagle and K. Ng. Machine-representation and Visualisation of a Dance Notation. In *Proceedings of the Electronics Imaging and the Visual Arts (EVA2003)*, pages 22.1–22.8, July 2003.

R. J. Neagle, K. Ng, and R. A. Ruddle. Studying the fidelity requirements for a virtual ballet dancer. In *Proceeding of Vision, Video and Graphics Conference (VVG2003)*, pages 181–188, July 2003.

J. G. Noverre. *Lettres sur la danse et sur les ballets*. Editions Ramsay, 1760. Jean-Georges Noverre, Letters of Dancing and Ballet, Translated from revised and enlarged edition (St Petersburg 1803) by Cyril W. Beaumont, London 1930.

R. Price, C. Douther, and M. A. Jack. An investigation of the effectiveness of choreography for the portrayal of mood in virtual environments. In *Proceedings of the fourth international conference on Autonomous agents*, pages 54–55. ACM Press, 2000. ISBN 1-58113-230-1.

J.R.E. Ramsay and M.J. Riddoch. Position-matching in the upper limb: professional ballet dancers perform with outstanding accuracy. *Clinical Rehabilitation*, 15 (3):324–331, June 2001.

C. Stevens, S. Mallock, R. Haszard-Morris, and S. McKechnie. Shaped time: A dynamical systems analysis of contemporary dance. In *Proceedings of the 7th International Conference on Music Perception and Cognition, Sydney*, pages 161–164, 2002.

Virtual Human Signing as Expressive Animation

John Glauert* Richard Kennaway* Ralph Elliott*
Barry-John Theobald*
*School of Computing Sciences
UEA Norwich, UK
{jrwg,re,jrk,bjt}@cmp.uea.ac.uk

Abstract

We present an overview of research at UEA into the animation of sign language using a gesture notation, outlining applications that have been developed and key aspects of the implementation. We argue that the requirements for virtual human signing involve the development of expressive characters. Although the principal focus of work has been on sign language, we believe that the work can be generalised easily and has a strong contribution to make to future research on expressive characters.

1 Signing Research at UEA

1 in 1000 people become deaf before they have acquired speech and may always have a low reading age for written English. Sign is their natural language. British Sign Language (BSL) has its own grammar and linguistic structure that is not based on English.

Sign language is expressive in its own right, and is multimodal, combining manual gestures, other bodily movements, and facial expressions. Facial information is especially important, conveying key semantic information. Just as intonation can affect the meaning of a sentence, for instance, turning a statement into a question, or indicating irony, so, facial gestures modify manual gestures in crucial ways. In addition, certain signs use the same manual content combined with different mouthings, often related to speech, to distinguish closely related concepts.

Research at UEA addresses the linguistics of sign language, where little is documented about grammar and semantics, and explores generation of signing, using gesture notation. We have developed SiGML (Elliott et al., 2001) (Signing Gesture Markup Language) for representing sign language utterances.

SiGML is used to generate realistic animation of signing using Virtual Human Avatars. The Animgen system (Kennaway, 2001) employs advanced techniques for skeletal animation to realise precise hand shapes and movements, leading to accurate bodily contacts. In addition, a range of facial gestures is animated by weighting morph targets giving appropriate displacements for facial mesh points.

In collaboration with Televirtual Ltd, a local multimedia company we have developed description formats for specifying avatars and their streams of animation parameters.

Complete systems have been produced allowing control of content animation using a range of avatars embedded in a range of applications including support on web pages. The framework integrates SiGML processing through Animgen, and supports a number of avatars developed separately at UEA and by Televirtual.

Early work was based on signing captured via motion sensors, using blending techniques to concatenate motion sequences. Further work is based on capture via video, especially for facial expressions, providing a basis for recognising signs from motion data.

2 Virtual Signing Applications

Since deaf people do not necessarily find information easy to absorb in text, their access to services is restricted, despite the requirements of recent legislation. There is little support for digital services in sign.

Recent projects by colleagues at UEA include Simon the Signer (Pezeshkpour et al., 1999), winner of two Royal Television Society Awards, and TESSA (Cox et al., 2002), winner of the top BCS IT Award, undertaken within the EU ViSiCAST project (ViSiCAST, 2000). Both Simon the Signer and TESSA (see Figure 1) used motion captured signs that are blended into sequences on demand.

2.1 Simon the Signer

Simon the Signer took words from a television subtitle stream and rendered a sequence of signs in Sign Supported English (SSE) to appear as an optional commentary on screen. SSE is widely used in education of deaf people, using a subset of BSL signs presented in English word order. Although technically successful, the use of SSE rather than true BSL was not fully accepted by the deaf community since it does not provide the required cultural richness.

There are obvious benefits for broadcasters if signing can be generated from an existing low-bandwidth data

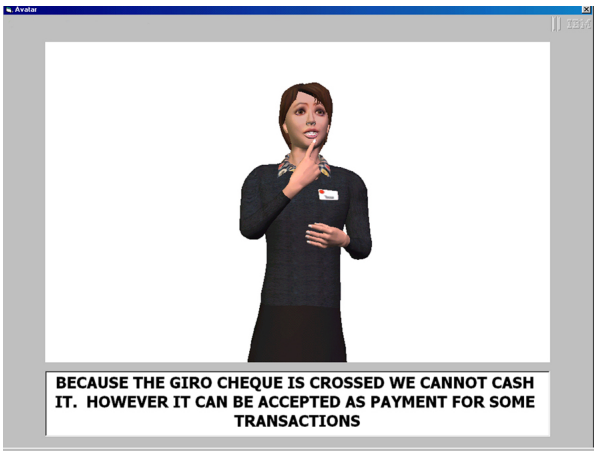


Figure 1: TESSA

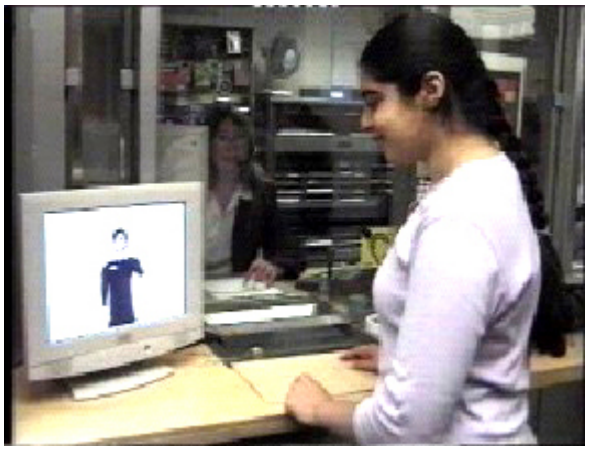


Figure 2: TESSA in use in a Post Office

stream such as subtitles. However, this seems a distant prospect. The use of video of sign interpreters is established. However, open captioned signing is not acceptable to hearing audiences and is only broadcast for a limited range of programmes at unsocial times. The bandwidth requirements are too high to broadcast a separate video stream for every channel that can be composited with the standard data streams in a set top box.

An alternative approach being explored in a current project is to capture the performance of a sign language interpreter, and transmit motion data parameters to drive an avatar in the set top box. Experiments show that the bandwidth requirement would be of the same order as for a speech channel.

2.2 TESSA

TESSA enables a Post Office clerk to communicate with a deaf customer by the use of speech recognition and avatar animation. Phrases used in standard transactions at Post Offices are recognised automatically and trigger animation of the corresponding sign language phrase in BSL. In addition to recognising fixed phrases, TESSA will handle phrases containing variable values such as days of the week or amounts of money and will substitute the corresponding BSL signs.

TESSA is an example of the use of an expressive character for communication. As the system is interactive and covers an extensive, though finite, domain, it provides genuine mediation between a hearing clerk and a deaf customer.

In addition to exploring applications in broadcast and to support face to face transactions, the ViSiCAST project also developed tools for providing low-bandwidth signing on the Web through a plugin for Internet Explorer.

2.3 Signed Weather Forecasts

Although the weather varies hour by hour, summary weather forecasts conform to a fixed pattern. The domain can be fully described for a number of natural spoken languages and natural sign languages. A system has been developed that enables forecasts to be presented by seamless blending of captured sign phrases using the web plugin.

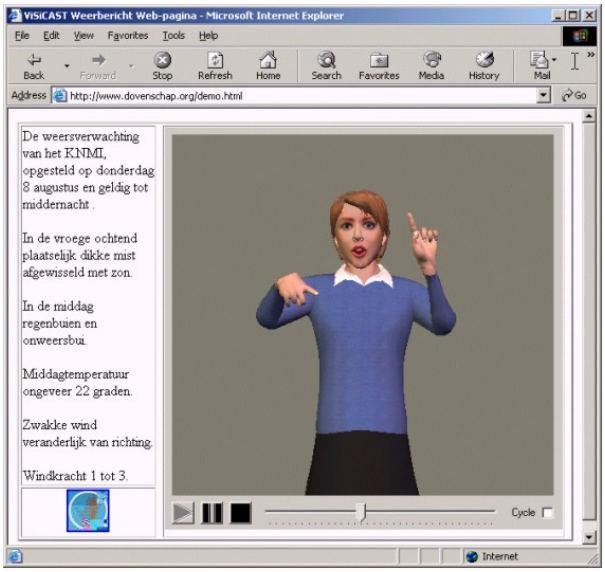


Figure 3: Weather Forecast on the Web

A tool has been developed which allows a non-signer to build forecasts, using standard weather phrases, for conversion into text and sign for a number of languages. Our implementation covers English, BSL, Dutch, SLN (Sign Language of the Netherlands) (see Figure 3), and DGS (German Sign Language). The Weather Forecast Creator, illustrated in Figure 4, may be used with a user interface in English, German, or Dutch and may be used to generate signing and text for all three countries. Hence it is not necessary for the content creator to know signing, or even the national language to be provided as text.

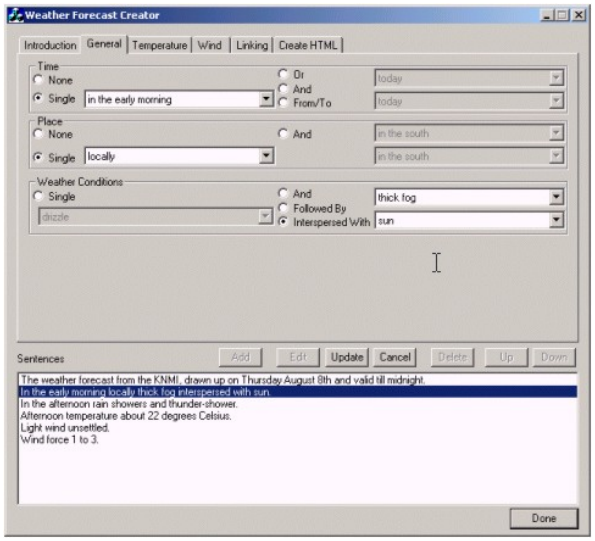


Figure 4: Weather Forecast Creator Application

Figure 6: German Website

2.4 Signing on eGovernment Websites

While earlier projects have been based on seamless concatenation of motion captured signs, the eSIGN project focuses on content created by synthesis from notation. As a result, information can easily be updated without the need for an expensive capture session. Information of an ephemeral nature can be generated automatically and interactively.

To enhance the usefulness of the internet for sign language users, the eSIGN project is developing signed commentary to accompany eGovernment forms. Figures 5 and 6 show web content under development.

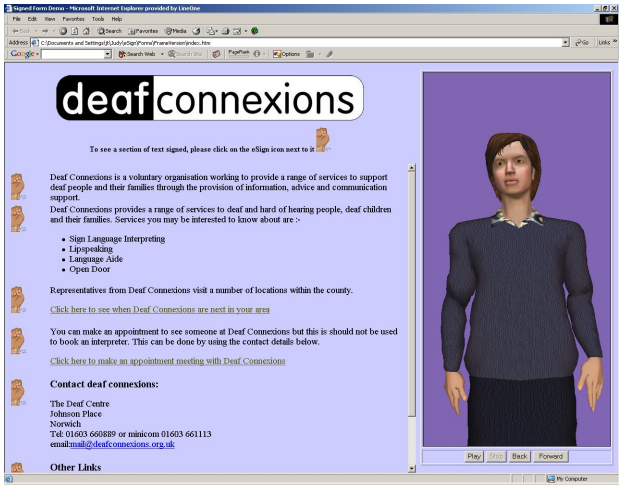


Figure 5: British Website

3 SiGML Notation

SiGML – Signing Gesture Markup Language (Elliott et al., 2001) – was initially developed in the ViSiCAST project as a key component of a prototype "natural-language-to-signed-animation" system developed in that project. Thus the primary purpose of SiGML is to support the definition of signing gestures in a manner allowing them to be animated in real-time using a computer-generated virtual human character, or avatar.

SiGML is an XML application language. We focus attention here on the major component of SiGML, referred to as "gestural" SiGML, which is used to drive the synthetic signing system. However, it should be noted that SiGML also allows the incorporation into the definition of a signed performance of data obtained by other means, including "motion capture" data, that is, motion parameters obtained by recording the actions of a human signer. Gestural SiGML is based on HamNoSys, the long-established Hamburg Notation System (Prillwitz et al., 1989) developed at the Institute for Deutsche Gebärdensprache (IDGS) at the University of Hamburg. Although it is based on HamNoSys, SiGML allows some physical features of the signer's posture to be specified with a greater degree of precision than HamNoSys. However, the semantic relation between the two notations is close: HamNoSys can be (and is) translated into SiGML; no significant information is lost in this process, and so any SiGML sign thus generated can generally be translated back into HamNoSys.

The purpose of HamNoSys is to support the transcription and analysis of human signing in a manner that is independent of the particular sign language used by the signer. Hence HamNoSys supports the transcription of signs at the phonetic level, providing special symbols and

structuring devices for representing phonetically significant features of signing such as hand shape and positions in “signing space”. HamNoSys effectively embodies a model of sign language phonetics, a model which is retained, largely unmodified, in SiGML.

A distinctive feature of sign language, in contrast with speech, is that is allows several distinct articulators to be in play concurrently. The most important articulators are the signer’s hands, but other bodily movements and various forms of facial gesture such as eye gaze and mouthing also have significance at the phonetic level in signing. Thus Gestural SiGML, like HamNoSys, has both a “manual” and a “non-manual” component. The manual component has a richer structure and is more fully specified than the non-manual, reflecting the fact that some non-manual aspects of signing, and their phonetic status, are less well-defined than are manual aspects.

3.1 Manual Signing

The manual component of SiGML allows a sign to be defined in terms of transitions between static postures, each of which may involve either or both of the signer’s hands. A hand posture is determined by the location of the hand in signing space, its shape, and its spatial orientation.

There is a core set of commonly occurring hand-shapes, such as “flat hand”,“fist” and“cee” (the shape of the hand when it is wrapped round a cylindrical object like a cup). A much larger repertoire of hand-shapes can be defined by applying modifications to these basic hand-shapes, for example, bending of individual fingers or the thumb, splaying of fingers, and various forms of contact between fingers. For two-handed signs, the notation allows precise specification of the relative configuration of the hands with respect to each other: this is achieved through the concept (taken from HamNoSys) of a “hand constellation”.

Various forms of hand motion may be specified: straight line, circular, or zig-zag. Each of these motions can be modified or refined in a wide variety of ways, of which the following is a small sample: a straight-line motion may be arced; the number of quarter-turns may be specified for a circular motion, whose radius may be varied dynamically to give a spiral effect; the plane in which a zig-zag movement is performed can be explicitly specified. Several more specialised forms of motion such as finger fluttering and wrist rotation are also supported. Another form of motion consists of a change of hand-shape or of hand-orientation. Motions may be combined in sequence and in parallel. There are modifiers which control the manner in which a motion is performed

3.2 Non-manual Signing

The definition of non-manual signing features in SiGML is based on the corresponding definitions for HamNoSys 4 (Hanke et al., 2000). A hierarchy of independent tiers, corresponds to distinct articulators. These may specify shoulder, body and head movements and eye gaze. Facial expressions control eye-brows, eye-lids, and nose. A repertoire of mouthings covers visemes for speech, along with other mouth gestures.

Here, “facial expression” refers to expressive uses of the face which form part of the linguistic performance, rather than those which communicate the signer’s attitude or emotional response.

3.3 SiGML Examples

```
<sigml>
<hamgestural_sign gloss="film">
  <sign_manual>
    <split_handconfig>
      <handconfig handshape="flat" extfidir="u"
            palmor="d"/>
      <handconfig handshape="finger2" thumbpos="across"
            extfidir="r" palmor="r"/>
    </split_handconfig>
    <split_location>
      <location_hand digits="2" contact="touch"/>
      <location_hand location="wristback"
            side="palmar" contact="touch"/>
    </split_location>
    <wristmotion motion="swinging"/>
  </sign_manual>
</hamgestural_sign>
</sigml>
```

Figure 7: SiGML for BSL sign “film”

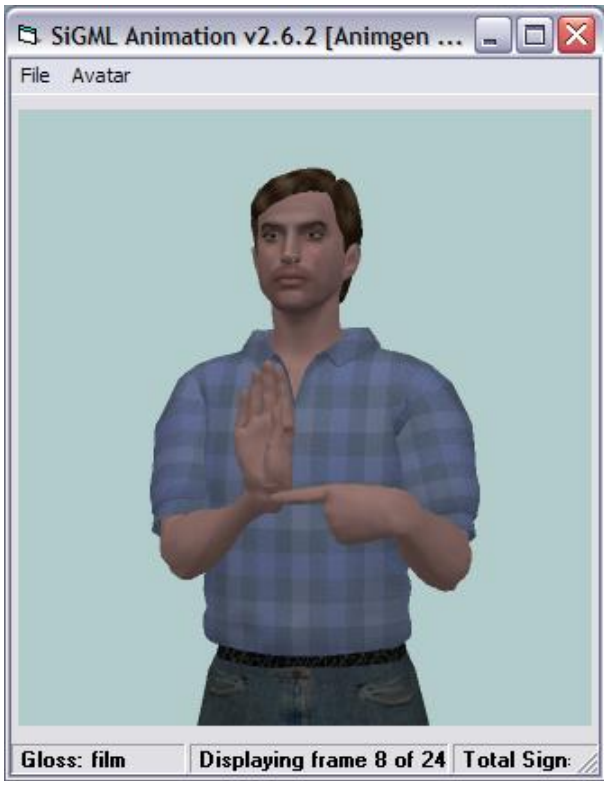


Figure 8: Initial configuration for BSL “film” sign

Figures 7 and 9 show two examples of individual SiGML signs. The first of these is the SiGML definition for the sign "film" in British Sign Language (BSL). This sign has a manual component but no non-manual component. Most of the former is devoted to the definition of the sign's initial configuration, shown in Figure 8. Both hands are involved in this configuration, and so for both hands there are specifications of their shape and orientation, followed by a specification of their locations with respect to each other: here the back of the wrist of the dominant hand is in contact with the extended index finger of the non-dominant hand. The motion for this sign is expressed comparatively succinctly in the `<wristmotion ...>` element, which specifies a swinging motion of the dominant (raised) hand.

```
<sigml>
<hamgestural_sign gloss="tell_story">
  <sign_manual both_hands="true"
               lr_symm="true" outofphase="true">
    <handconfig handshape="flat" thumbpos="out"/>
    <handconfig extfidir="ul"/>
    <handconfig palmor="ul"/>
    <handconstellation contact="close">
      <location_hand location="tip" digits="3"/>
      <location_hand location="palm"/>
      <location_bodyarm location="shoulders"/>
    </handconstellation>
    <rpt_motion repetition="fromstart">
      <tgt_motion>
        <circularmotion axis="l"/>
        <handconstellation contact="close">
          <location_hand location="tip" digits="3"/>
          <location_hand location="palm"/>
        </handconstellation>
      </tgt_motion>
    </rpt_motion>
  </sign_manual>
</hamgestural_sign>
</sigml>
```

Figure 9: SiGML for BSL Sign "Tell-story"

Figure 9 shows the SiGML definition for the BSL sign "tell-(the-)story". Here, in contrast to the previous example, both hands are involved not only in the sign's initial configuration (shown in Figure 10), but also in the subsequent motion. The attributes in the main `<sign_manual ...>` element specify that both hands participate in the motion, that there is left-right symmetry in this motion, and that the motions of the two hands are to be performed out-of-phase with each other. In this sign the "location" part of the intitial configuration consists of a hand-constellation which specifies the precise configuration of the hands with respect to each other, as well as the position in signing space of the two hands thus combined. Given the symmetry characteristics already specified for the two-handed motion as described above, an explicit movement specification is required for the dominant hand only. In this case, the required motion has significant internal structure explicitly defined in the notation: the motion has an explicit target, and is repeated once from its starting configuration.

Recently, as described in (Elliott et al., 2004), we have been developing support in our synthetic animation framework for the non-manual features of SiGML. The SiGML example shown in Figure 11, illustrates the fact that the system described there can be adapted to synthesise expressions of emotion, as well as the linguistically significant facial expressions required for signing. Figure 12 shows a pair of frames from the resulting animation: in the first of these frames the avatar's face is still in its neutral posture, in the second it has reached a much more expressive one.

Figure 10: Frame from BSL "tell-story" sign

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE sigml SYSTEM
  "http://www.visicast.cmp.uea.ac.uk/sigml/sigml.dtd">
<sigml>
<hamgestural_sign gloss="vguido-sad">
  <sign_nonmanual>
    <extra_tier>
      <extra_par>
        <extra_movement movement="X14"/>
        <extra_movement movement="X15"/>
        <extra_movement movement="X24"/>
        <extra_movement movement="X38"/>
        <extra_movement movement="X41"/>
      </extra_par>
    </extra_tier>
  </sign_nonmanual>
  <sign_manual/>
</hamgestural_sign>
</sigml>
```

Figure 11: SiGML for a tearful facial expression

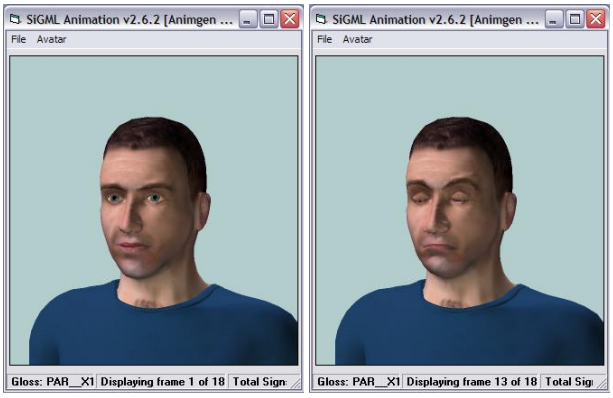


Figure 12: Two frames from the tearful facial animation

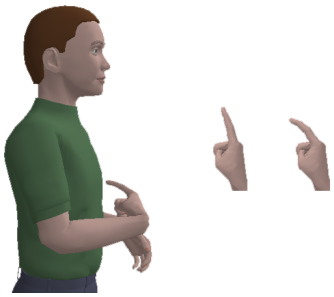


Figure 13: Pointing inwards

4 Animation of Signing

The ViSiCAST and eSign projects have achieved significant results in creating signing animations from HamNoSys (Kennaway, 2001, 2003). Although this notation was originally developed for researchers into signing to communicate with each other about signs, it has proved a suitable basis for synthetic animation.

4.1 Manual Animation

To translate the human-meaningful notations of HamNoSys into numerical animation data, several problems must be solved. The fuzzy categories of HamNoSys—the "chest", a "large" movement, a "close" proximity, a "fist" handshape etc.—must be replaced by numerical locations, distances and joint rotations. The locations named by HamNoSys, of which there are a few hundred, must be provided as part of the definition of the avatar. In general there is no way to automatically determine these locations by calculation from the surface mesh or the animation bones (which do not always closely correspond to the physiological bones). Given these, the various sizes of movements and proximities can be defined as multiples of measurements of the avatar. For example, we define near and far distances from the torso as a certain proportion of the length of the arms, since in signing these are primarily used as locations at which to place the hands.

HamNoSys transcriptions often leave out information which is obvious to the human reader and writer of the notation. Sometimes it simply takes a standard default value: absence of any explicit location for a sign means that it happens in the middle of the signing space. Sometimes it is dependent on the context: when the hands touch each other, the location at which they touch is often not specified.

HamNoSys specifies various types of repetition: repeating a movement from its starting position, repeating a movement from where its previous occurrence finished, repeating it several times getting larger and larger, etc. Various modes of repetition can be combined, and it can be quite complicated to determine exactly what an arbitrary repetition specification really means.

There are other features of the posture and movement which HamNoSys does not record at all. For example, it mostly describes what the hands do; what the rest of the arm must do to place the hands in the positions specified is not described. The animation software must be programmed with rules to decide how high the elbows are raised, and whether the collarbone joint moves. Physical objects are prevented by their nature from penetrating each other. The avatar's body parts are under no such constraint, except for whatever has been explicitly programmed.

To synthesise a lifelike movement from one posture to another, we use a semi-abstract biocontrol model to determine the accelerations and decelerations, parameterised in a way that lets us animate the various manners of movement which HamNoSys can specify: fast, slow, tense, with a sudden stop, etc.

Sometimes, the simplest way to resolve the problem of what a given piece of HamNoSys means is to make it more detailed, explicitly specifying information that is impractical to calculate: for example, specifying which points on the hands are in contact instead of merely saying that the hands contact each other. We are currently moving towards a version of SiGML that will allow the specification of more detail of this sort, and thus separate the problem of filling in the missing information from that of animating the gesture. This allows the trade-off between the effort of the transcriber and the effort of the animator to be made in different ways.

In some instances, HamNoSys transcriptions have been found to be incorrect, even when made by experienced users of the notation. There is a tendency for people to write down not the actual motion, but an idea of the motion that sometimes does not closely match it. An example occurring frequently in the HamNoSys corpus of over 3,000 signs of German Sign Language is that of an inward pointing finger (see Figure 13).

Often, the hand shape has been transcribed as if it were the first of the two hands in that figure, with the wrist bent sharply so as to point the whole hand at the signer. In reality, the hand shape will be more like the second shape

shown, with the fingertip pointing towards the signer, and the back of the hand pointing in a direction above, left, and behind the signer. This is perhaps an indication that a signing notation should transcribe these signs by recording the direction in which the finger points, rather than the direction in which the back of the hand points. In general, we can say that in any gesture, some geometric properties of the posture and movement are significant and some are not. A possible definition of the significant aspects is: those which would remain the same even when the sign was performed by a different avatar, with different body proportions. We are currently considering a revised version of the notation which would attempt to record signs in terms of such significant properties, and would be intended from the beginning for computer animation. Extending HamNoSys or SiGML to other classes of movement, such as those required by interactive characters in virtual environments, will be the subject of future research.

4.2 Facial Expressions

As mentioned above, non-manual signing includes a range of bodily movements, of head and shoulders, that can be animated by controlling the articulation of appropriate joints. In addition, there are facial expressions that are animated by controlling the vertices of the facial mesh.

Some expressions, denoted *mouth gestures*, come from a set of gestures used in signing, such as puffing out a cheek or raising the eyebrows. Other expressions, denoted *mouth pictures* consist of the visemes corresponding to an arbitrary phonetic (IPA) string. For convenience, this viseme string is expressed using the SAMPA (Wells, 2003) conventions for transcription of the IPA.

As reported in (Elliott et al., 2004), Animgen assumes that each avatar comes with a set of facial deformations, which are named morphs, which can be applied in combination to animation frames. Animgen has no detailed model of morphs but specifies a weighting for each morph for each animation frame.

Facial non-manuals used in SiGML are encoded as *morph trajectories*. A trajectory consists of a morph name, the maximum weighting of that morph to be applied, and an envelope describing the attack, sustain, and release for the morph.

Morph trajectories can be combined in series and in parallel to build up an arbitrarily complex definitions that are specified in a configuration file specific to each avatar. The creator of the avatar creates the avatar's morph set and the mapping of SiGML facial elements.

Figure 14 gives such a specification for a mouth gesture, which is defined as a mouthing of "bEm". It is realised by a sequence of three morphs corresponding to the three phonemes, where the first and third have been given identical visual representations.

The timings (slow, medium, fast, zero, or sustain to end

```
<mouth_gesture sigmlName="L09">
    <morph name="mbp" amount="1" timing="mt-ft"/>
    <morph name="aaa" timing="ft-mt"/>
    <morph name="mbp" timing="mteml"/>
</mouth_gesture>
```

Figure 14: Definition for Mouth Gesture L09

of sign) can be given symbolically (s, f, m, −, or e). The symbolic tokens are mapped to times in another configuration file.

Additionally, "manner" components determine how the morph approaches its full value during the attack, and how it tails off during the release. The possible values for this are "t" (tense) and "l" (lax). They are mapped to sets of parameters for a general model of accelerations and decelerations.

An extension of this format, illustrated in Figure 15, is used to define morph trajectories for viseme sequences derived from SAMPA strings.

```
<sampa phonemes="EIszi">
    <morph name="cgng" amount="0.7" timing="m t - m t"/>
</sampa>

<sampa phonemes="a_I">
    <morph name="aaa" timing="m t - m t"/>
    <morph name="cgng" amount="0.7" timing="m t - m t"/>
</sampa>
```

Figure 15: Defining SAMPA codes E, I, s, z, i, and aI

Several phonemes may correspond to the same viseme, for example *E, I, s, z,* and *i*. Hence a single specification is used to animate any of the phonemes in a list. Diphthongs are often required, but they vary from language to language. In order that a single set of definitions can be used for all languages, we require that diphthongs are tied together with an underscore. Hence the diphthong *aI* is encoded as *a_I*.

In order to handle coarticulation in a viseme sequence, the release of one trajectory is overlapped with the attack of the next. However, this is a largely untested approach and we are not confident that it will provide mouth movements suitable for lip-reading, for example. Instead, we are working to incorporate leading work on audio-visual speech synthesis based on appearance models that has been undertaken at UEA. This work is discussed below.

4.3 Animation System Architecture

As stated earlier the synthetic animation system we describe here was developed as part of a complete prototype system in the ViSiCAST project, in which the input is a natural language (English) text for which a signed animation is generated. This system divides into a front-end, which applies natural language processing techniques based on DRT and HPSG (Safar and Marshall, 2002), and a back-end — the system described here. The interface between these two subsystems is a

phonetic-level definition of the required signing sequence expressed in SiGML. The main data flow in this back-end SiGML-to-Animation system can be viewed as a pipeline as shown in Figure 16. Of the three processing stages shown in this figure, the most significant (because the most novel) is the central one which converts SiGML sign definitions into the corresponding animation frame definitions, as described earlier in this section. One important feature of the architecture not explicitly represented in the diagram is the fact that the synthetic animation module has an additional input, namely the description of the avatar's geometry, which is supplied with each avatar as an essential part of its definition.

The first stage in the pipeline decomposes the input stream into individual "signing units". A signing unit is typically an individual sign expressed in gestural SiGML, but it may instead consist of motion capture data for a sign (in which case it by-passes the synthetic animation stage). The first stage can also perform translation of a sign definition from HamNoSys to SiGML. The final stage in the pipeline consists of rendering software, which applies conventional 3-D animation techniques to each packet of frame data, first to determine the configuration of the avatar's surface mesh corresponding to the given configuration for its virtual skeleton (and morph weights), and then to render this mesh with the appropriate colouring and texture on-screen. A separate controlling module manages the scheduling of the necessary data transfers between the individual stages shown in Figure 16.

Figure 16: Processing Sequence for Synthetic Animation of SiGML

4.4 Appearance Models for Faces

Appearance models (Cootes et al., 1998) are statistical models of the shape and appearance variation of the face, which are learnt from hand-labelled facial images. Traditionally these have been used in the computer vision community to track and recognise faces (and other objects) in video sequences. Analysis is done by synthesis, i.e. the model is able to synthesise realistic example images by applying the appropriate parameters and an optimiser is used to update an estimate of the parameters such that the original and model generated images coincide. The face in an image is then encoded in terms of the parameters of the model, or is mapped to a point in a face-space spanned by the model.

Work at UEA on modelling talking faces has focussed on the use of shape and appearance models. A talker first recites a series of training sentences and the video analysed using the shape and appearance model. Since the face in a single frame forms a point in the model-space, a sentence forms a trajectory in this space. These trajectories are segmented according to their phoneme boundaries, derived from the corresponding acoustic signal.

To synthesise a novel utterance, a sequence of phoneme symbols is required. The synthesiser then selects a sub-trajectory from the original data that corresponds to the desired phoneme in the closest context to that in which it appears in the new utterance. These sub-trajectories are concatenated to form a new trajectory of model parameters, which are then applied to the model to create a realistic synthetic talking face.

This approach provides the flexibility and efficiency of traditional graphics-based talking faces with the realism of traditional image-based talking faces. A further advantage is that a complete avatar can be animated using this technique, so the talking head can be coupled with signing and other manual gestures (Theobald et al., 2003). Here the geometry of the face of the avatar is updated using the shape component of the appearance model, while the appearance component provides a texture update that significantly improves the realism when, for example, only a single texture is used.

5 Future for Signing and Expressive Characters

To develop virtual human signing it has been essential to address issues of both human animation and content creation. Animation only becomes acceptable once it achieves good visual realism with relatively natural motion. To support useful quantities of signed content it was necessary to develop scripting techniques soundly based in signing linguistics.

A benefit of using notation is that semantically unimportant information can be left implicit. An example is the position of elbows during signing. During animation, such implicit information is reconstructed using inverse kinematics. For representing more general gestures it is likely to be necessary to provide the option of being more explicit about aspects of gesture that do not matter for signing, but the principle of minimising the amount of explicit information is crucial.

The choice of a high-level representation is important if animation is to be scripted without knowledge of the physical dimensions of the avatar. A crucial part of our work has been an extended avatar definition format that enables the Animgen software to generate acceptable animation for any compliant avatar. A number of different avatars have been used in illustrations in this paper, but the software does is generic.

The notation concentrates on gestures for signing and only addresses upper body movement. There are few features relating to interaction with the environment, al-

though contacts between parts of the body are addressed in detail. We intend to develop SiGML to encompass a wider range of movement and gesture including conversational gesturing, whole-body actions such as walking and running, and interaction with physical objects.

An immediate application will be through the EPOCH Network of Excellence (Arnold, 2003), using avatars to help the user visit virtual cultural heritage sites constructed using the CHARISMATIC UEA/TU Braunschweig modeller (Day et al., 2003). A scenario would be that the user follows a walking, talking, multi-lingual virtual guide to places of interest in the scene. Ideally the visitor should be able to interact (via speech) with the virtual guide as well as the rest of the model.

We have introduced the leading work on audio-visual speech synthesis that is undertaken at UEA. To date, this work has focussed on the synthesis of the visible articulators associated with speech production only, i.e. the lips, teeth and tongue. It is well known that realistic conversational characters require expressive speech, which is lacking in the current system. To determine whether the model is able to re-synthesise the range of expressions required by a conversional character, it is currently being used to analyse the face of a signer and re-synthesise the facial movements on a virtual signer.

Much of our experience with signing appears to have wider application to work on expressive characters. The repertoire of techniques is clearly applicable to animation of more general gestures, although it remains to be seen how much extension is necessary to notations for signing and to animation techniques to achieve this purpose.

Acknowledgements

We acknowledge with thanks financial support from the European Union, and assistance from our partners in the ViSiCAST and eSIGN projects, in particular Televirtual Ltd. who supplied one of the avatars and the supporting rendering software.

References

D.B. Arnold. Plans for the EPOCH Network. In *VAST2003 Symposium*, Brighton, 2003.

T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active Appearance Models. In H. Burkhardt and B. Neumann, editors, *Proc. European Conference on Computer Vision 1998, Vol 2*, pages 484–498. Springer-Verlag, 1998.

S.J. Cox, M. Lincoln, J Tryggvason, M Nakisa, M Wells, M. Tutt, and S Abbott. TESSA, a system to aid communication with deaf people. In *ASSETS 2002, Fifth International ACM SIGCAPH Conference on Assistive Technologies*, pages 205–212, Edinburgh, Scotland, 2002.

A.M. Day, D.B. Arnold, D. Fellner, and S. Havemann. Combining polygonal and subdivision surface approaches to modelling of urban environments. In *Proceedings of Cyberworld 2003*, Singapore, December 2003 2003.

R. Elliott, J.R.W. Glauert, and J.R. Kennaway. A Framework for Non-Manual Gestures in a Synthetic Signing System. In *Proc. Cambridge Workshop Series on Universal Access and Assistive Technology (CWUAAT)*, 2004.

R Elliott, JRW Glauert, JR Kennaway, and KJ Parsons. D5-2: SiGML Definition. working document, ViSiCAST Project, 2001.

T Hanke, G Langer, C Metzger, and C Schmaling. D5-1: Interface Definitions. working document, ViSiCAST Project, 2000.

J.R. Kennaway. Experience with and requirements for a gesture description language for synthetic animation. In *5th International Workshop on Gesture and Sign Language Based Human-Computer Interaction*, LNAI, to appear. Springer-Verlag, 2003.

R. Kennaway. Synthetic animation of deaf signing gestures. In *4th International Workshop on Gesture and Sign Language Based Human-Computer Interaction*, LNAI, pages 146–157. Springer-Verlag, 2001.

F. Pezeshkpour, I. Marshall, R. Elliott, and J.A. Bangham. Development of a legible deaf signing virtual human. In *IEEE Multimedia Systems '99 (IEEE ICMCS '99)*, 1999.

S. Prillwitz, R. Leven, H. Zienert, T. Hanke, J. Henning, et al. *Hamburg Notation System for Sign Languages — An Introductory Guide*. International Studies on Sign Language and the Communication of the Deaf, Volume 5. Institute of German Sign Language and Communication of the Deaf, University of Hamburg, 1989.

E. Safar and I. Marshall. Sign Language Translation via DRT and HPSG. In A. Gelbukh, editor, *Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, Lecture Notes in Computer Science (LNCS), pages pp58–68, Mexico City, Mexico, 2002. Springer-Verlag.

B.-J. Theobald, A. Bangham, I. Matthews, J.R.W. Glauert, and C.C. Cawley. 2.5D Visual Speech Synthesis Using Apprearance Models. In *British Machine Vision Conference (BMVC 2003)*, Norwich, UK, 2003. BMVA.

ViSiCAST. Virtual Signing: Capture, Animation,and Storage. http://www.visicast.cmp.uea.ac.uk, 2000.

J Wells. SAMPA computer readable alphabet. http://www.phon.ucl.ac.uk/, 2003.

Defining the Gesticon: Language and Gesture Coordination for Interacting Embodied Agents

Brigitte Krenn, Hannes Pirker
*Austrian Research Institute for Artificial Intelligence (ÖFAI)
Freyung 6, A-1010 Vienna, Austria
{brigitte,hannes}@oefai.at

Abstract

In this paper we address problems of the automatic assignment of speech accompanying gestures and present solutions we have developed and still develop in the IST-project NECA. Special emphasis is put on the presentation of the central repository of information necessary for this assignment: the so called *gesticon*.

1 Introduction

The task of automatic gesture assignment discussed in this paper, can be described as follows: Given a dialogue between two or more embodied agents, specify their non-verbal behavior by automatically selecting ”appropriate” gestures and facial expressions from a given set. Take care of the temporal alignment of gestures with the spoken utterance and provide the information in a way that it subsequently can be used as input for an animation engine. The crucial task here is to design a gesture repository with representations general enough to be reusable in different multimodal generation systems and to be applicable in combination with different animation engines. What is required from the generation system is the availability of information on the dialogue structure, the dialogue related emotion, and the prosody and timing of speech.

Our discussion will be centered around

a) the design and representation structure of a central gesture repository, which we call *gesticon* in analogy to lexicon,[1]

b) the methods and strategies employed in gesture generation and the alignment of gestures with speech.

As regards a), we discuss what information shall be represented in the gesticon, and how this information shall be structured and represented. As regards b), we make proposals for gesticon-based gesture generation and gesture timing in collaboration with multimodal natural language and speech generation. In order to do so, we introduce a general purpose multimodal representation language, the RRL (Rich Representation Language, see `http://www.oefai.at/NECA/RRL`), which is the back-bone of the whole gesture-assignment process and functions as an interface to the individual system components. Both gesticon and RRL are represented in XML format, thus ensuring compatibility with a variety of existing representation and standardisation efforts of multimodal information. For an overview see (Pirker and Krenn, 2002). Coupling gesticon and RRL also allows us to design a component which makes the gesture representations and gesture generation strategies and methods independent from implementation details of individual system modules.

Thus, even though we describe gesture representation and gesture assignment in the context of the NECA system[2], our proposals are general in nature and not restricted to NECA.

The paper is organized as follows: To set the context, we briefly introduce the NECA project (section 2.1) and the architecture of the NECA system (section 2.2). In section 2.3 we give an outline of gesture encoding in the RRL (Rich Representation Language), a general purpose multimodal representation and scripting language which has been developed in the NECA project. In sections 3.1 to 3.6 the overall gesticon structure is defined and the organization of gesture relevant information is discussed. Gesticon entries are exemplified in section 3.7.

2 The NECA System

2.1 Outline

NECA (”Net Environment for Embodied Emotional Conversational Agents”) aims at the development of a toolkit that allows for time- and cost efficient implementation and adaption of Web-applications for the following scenario: Animated scenes are generated where two or more

[1] Other terms in use for a repository of gesture definitions are *gestuary*, a term coined by (deRuiter, 1998) and subsequently employed by (Kopp and Wachsmuth, 2000), and *gestionary*, a term used by Isabella Poggi (Poggi, 2002a) to refer to a dictionary of symbolic gestures, or *dictionary of gestures* such as ’The Berlin dictionary of Everyday Gestures’ (Posner et al., 2002) or ’The Nonverbal Dictionary of Gestures, Signs & Body Language Cues’ (Givens, 2002).

[2] `http://www.oefai.at/NECA/`

virtual human-like characters communicate with each other using expressive (emotionally rich) speech, gesture and facial expression.

Due to bandwidth restrictions, the use of lean player technologies is necessary.

For various reasons it is important that the system can easily be adapted to different player technologies. For instance, in different applications varied animation styles are preferred, the state-of-the-art in player technology is rapidly changing, improvements in bandwidth capacities increase the choice of web compatible player technology. Thus special emphasis needs to be put on keeping the influence of player-specific aspects as small as possible. In the two NECA demonstrators we currently work with two fairly different animation/player technologies, namely Charamel (`http://www.charamel.de`) and Macromedia Flash (`http://www.macromedia.com`).

2.2 Architecture

Because in NECA whole dialogues are planned in advance in the way a playwright designs a scene, a strict pipeline-architecture as depicted in Figure 1 can be employed. The information between modules is passed on using NECA's XML-compliant Rich Representation Language (cf. `http://www.oefai.at/NECA/RRL`, (Piwek et al., 2002)). First, the scene generation and an affective reasoning component (Gebhard et al., 2003) specify the dialogue acts to be produced and feed into the multi-modal natural language generator (M-NLG) (Piwek, 2003). M-NLG is responsible for the generation of the textual representation of the agent's utterances as well as the selection of semantically motivated gestures and emotion-driven facial expressions. Relevant information is encoded in the <function>-element of a gesticon entry.

The following concept-to-speech synthesis module (Schröder and Trouvain, 2003) does not only produce speech files containing emotional speech, but also provides full timing information, i.e., the exact position and duration of all phonemes, syllables, words, phrase boundaries and tonal accents.[3] This information is crucial for the Gesture Assignment module (GA). Here the final selection of gestures takes place and the animation is timed. Phonemes are mapped to visemes, tone accents are aligned with eyebrow raises, and selected parts of an intonation phrase are aligned with specific components of a gesture. The GA module makes use of information encoded in the <form>-element of a gesticon entry. Both M-NLG and GA make use of constraints encoded in the <restrictions>-element. After GA, a further component, the Animation Generator, produces the player-specific animation instructions. Player-specific information can be accessed via the <playercode>-element of a gesticon entry. While the input to this component is an RRL document, the output is code which can be directly rendered by the player employed.

[3]The speech synthesis system MARY can be tested online at `http://mary.dfki.de`

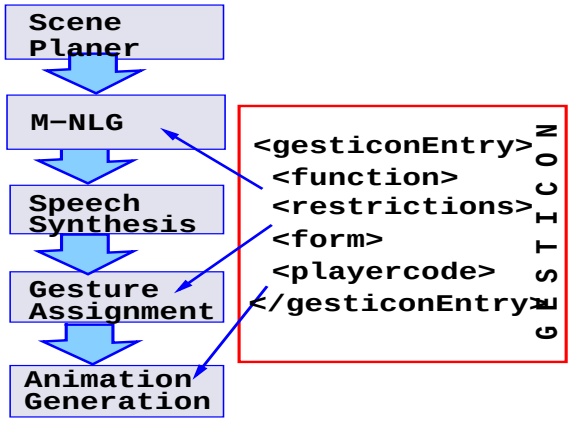


Figure 1: *Schematic diagram of NECA-architecture and how different modules make use of specific information provided in the gesticon.*

2.3 Gesture Encoding in the RRL

Generally speaking, dialogue accompanying gesture generation is a two-step process.

1. During multimodal language generation, gestures are selected on the basis of the semantic and pragmatic content of the utterances, and are symbolically linked to whatever entity is appropriate, e.g. a word or a sentence.

2. Based on the prosodic and temporal information produced by a speech synthesis component a fine-grained alignment between the verbal and nonverbal communication systems is performed.

The relevant information is encoded by means of the RRL. The interplay of the different aspects of multimodal information is exemplified in the following. The RRL-snippet below illustrates the result of step 1) multimodal generation.

```
<gesture identifier="hipshift"
         id="g001"
         aligntype="seq_before"
         alignto="s001"/>
<gesture identifier="wave"
         id="g002"
         aligntype="par_end"
         alignto="s001"/>
<sentence id="s001">
  Hello, how are you?
</sentence>
```

Two classes of gestures – identifier="hipshift" and identifier="wave" – have been selected to accompany the sentence "Hello, how are you?". This is specified via the value of the *alignto* attribute, i.e., the unique "id" of the

sentence. The *aligntype* attribute designates the temporal relationship between gesture and anchor element. In this case the gesture ”hipshift” would be realised before sentence ”s001” starts and the gesture ”wave” should end when the sentence stops.

The speech synthesis system produces the according soundfile for the sentence, and also provides information on its internal structure (syllables and phonemes) as well as information on the location and type of tonal accents and prosodic phrase boundaries, represented in ToBI format (Baumann et al., 2001). See the RRL representation below.

```
<sentence id="s001" src="s001.mp3">
  <word id="w_1" accent="H*" pos="UH"
        sampa="h@l-'@U">
     Hello
     <syllable id="syl_1" sampa="h@l">
        <ph dur="75"  p="h"/>
        <ph dur="48"  p="@"/>
        <ph dur="100" p="l"/>
     </syllable>
     <syllable id="syl_2" sampa="'@U"
               stress="1" accent="H*">
        <ph dur="230"  p="@U"/>
     </syllable>
   </word>
  <prosBoundary breakindex="4"
                dur="200"
                p="_"
                tone="H-L%"/>
  <word id="w_2" ... />
  ...
</sentence>
```

With the availablility of exact phoneme durations the alignment-specifications produced by multimodal generation can now – in step 2) of the gesture assignment process – be transformed into concrete time-measures. More sophisticated alignto-types can be processed such as the alignment of a certain gesture component to the syllable which bears the nuclear accent of a phrase, information not available at step 1) of gesture processing.

The output then is an unambiguous specification of the animation stream, which is expressed by means of a subset of W3C's Synchronized Multimedia Integration Language (SMIL 2.0, `http://www.w3.org/TR/smil20/`), i.e., via a collection of <seq> and <par> elements. At this step all linguistic information is discarded and replaced by an <audio>-element which holds the name and duration of the speech soundfile. The symbolic alignment between gestures and language-related entities (e.g. sentences, words, syllables) is replaced by the specification of the exact temporal alignment between this <audio>-element and the according <gesture>-objects.

The example from above would render to:

```
<animationSpec>
  <seq>
     <gesture key="g023"
              identifier="hipshift"
              id="g001"
              dur="1650"/>
     <par>
        <audio src="s001.mp3"
               dur="1459"/>
        <seq>
          <!-- visemes -->
          <viseme identifier="v_h"
                  dur="75"/>
          <viseme identifier="v_@"
                  dur="48"/>
          <viseme identifier="v_l"
                  dur="100"/>
          <viseme identifier="v_@U"
                  dur="230"/>
          ...
        </seq>
        <gesture key="g012"
                 identifier="wave"
                 id="g002"
                 begin="259"
                 dur="1200"/>
     </par>
 <seq>
</animationSpec>
```

It can be seen, that the <sentence>-element of the input is now replaced by an <audio>-element, which refers to the soundfile to be played. The sequence of visemes is of course parallel to the audio-element, and the aligntype “par-end” for the “wave”-gesture is reflected by the temporal offset specified in its *begin*-attribute. The *id* attributes used as unique identifiers throughout the processing are redundant at this stage, and are kept for debugging purposes only.

3 The Gesticon

As already indicated, the gesticon is designed as a general repository of meaningful bits and pieces of animation descriptions which are relevant for the generation of dialogue accompanying nonverbal behaviour. In other words, the gesticon is the direct equivalent to the lexicon in language-processing systems. As the latter is a mapping from phonetic form to the meaning of words, the gesticon represents the mapping between the form and the semantics of a gesture. In analogy to words in a dictionary, gesticon entries store information about the form (phonology), the meaning (semantics), the combinatory properties (syntax) and the pragmatics of gestures. Thus our conception of gesticon corresponds to Poggi's notion of (gesture) 'lexicon'. In (Poggi, 2002b) it reads

> In a “codified” communication system, the

> signal-meaning link is shared and coded in the memory of both a Sender and an Addressee (as it is the case, for example, with words or symbolic gestures) and a whole set of these links makes a "lexicon".

Note though, that Poggi's work focuses mainly on a verbal description of symbolic (emblematic) gestures, i.e., gestures with a conventionalized meaning within a certain community such as 'thumbs up' meaning 'o.k.' in many western countries. In contrast we aim at a machine readable gesture repository, which functions as the basic resource for the automatic generation of all different types of gestures. With the gesticon we propose the foundations for a framework for the uniform symbolic representation of different nonverbal communication systems such as gesture and facial expression. Without doubt, descriptive work such as the one by Poggi or the descriptions available in the Berlin dictionary of Everyday Gestures (Posner et al., 2002) will be valuable resources to instantiate the gesticon structure. As a precondition, however, these works need to be made machine-readable. Another open question is how effectively the textual descriptions can be transformed into appropriate entries for automatic gesture generation.

In the following we present the general structure of a gesticon entry and discuss the representational details of entries for facial expression and gesture. An illustrative example is provided in section 3.7. The gesticon is represented in XML format. Each entry comprises a form, a function and a restriction element, and pointers to player-specific representations. The fact that currently only information on facial expressions and hand-arm gestures is represented in the gesticon results from the NECA context where animated characters do not move within the scene.

3.1 Overall Structure of a Gesticon Entry

We propose the following overall structure for a gesticon entry.

```
<gesticonEntry>
      <verbatim/>
      <function/>
      <form/>
      <restrictions/>
      <playercode/>
</gesticonEntry>
```

The attributes *key* and *identifier* in the gesticonEntry are both used for naming the entry. The first is the entry's unique key, while the identifier is used as common name for gestures that share the same meaning, i.e., there can be numerous gestures with the identifier "greeting".

Gesticon entries are classified according to the main modality expressed. This information is specified via the *modality* attribute. In our examples the value is either "arms" which means the entry is a representation of a gesture or "face" which indicates that the entry is a representation of a facial expression. In the context of NECA, a further modality is "body", which stands for posture such as relaxed versus upright, etc. In the long run, however, the modality "body" needs to be further subclassified, for example, into posture, movement, and spatial location.

3.2 The <verbatim>-element

In the verbatim element, a verbal description of the gesticon entry is stored. This is information for the human reader.

3.3 The <function>-element

The function element contains information about the meaning and type of an entry, where the entry is attached to, and which type of temporal alignment is to be used (before, after, parallel, etc.)

The *type* attribute is not defined for facial expressions. As regards gestures, we distinguish between the following types:

- deictic (indicative or pointing gesture)
- beat (repetitive or rhythmic movement mainly coordinated with speech prosody)
- iconic (a gesture which "bears a close formal relationship to the semantic content of speech" (McNeill, 1992) quoted after (Serenari, 2002), p. 57, e.g. the hands forming a box in order to depict a container)
- emblematic ("gestures that have a specific social code of their own" (McNeill, 1985) quoted after (Serenari, 2002), p. 57, e.g. a nod meaning 'yes')
- illustrator (e.g. a wave accompanying or substituting a greeting act; in our use illustrators are similar to emblems, but are less strict as regards their social or cultural norms than emblems)
- metaphoric ("similar to iconics in that they present imagery, but present an image of an abstract concept" (McNeill, 1992) quoted after (Serenari, 2002), p. 57)
- adaptor ("part of adaptive efforts to satisfy self or bodily needs, or to perform bodily actions, or to manage emotions, or to develop or maintain prototypic interpersonal contacts, or to learn instrumental activities" (Ekman and Friesen, 1968) quoted after (Serenari, 2002), p. 59)
- idle (we have introduced a number of idle gestures which are selected when the animated characters "do nothing", i.e., they are not engaged in a dialogue, they are waiting till data transmission is completed)

Summing up, we have drawn the values for our *type* attribute mainly from work by Ekmann/Friesen and McNeill, cf. (Ekman and Friesen, 1968), (McNeill, 1985), (McNeill, 1992). The selection was guided by practical decisions, i.e., which classification is useful in the context of the NECA demonstrators. In general the classification of gestures is somewhat controversial in the literature, see for instance (Krauss et al., 2000) or (Serenari, 2002) for an overview of gesture classifications.

At the current stage of development, the values for the *meaning* attribute are simple atomic labels. Of course this is a shortcoming and reflects a rudimentary semantic classification of gestures and facial expression. This approach, however, is sufficient for the current stage of development of the NECA system. Especially for the generation of metaphoric gestures, however, the encoding of meaning via a symbolic label is inappropriate. Instead a more complex representation structure of the meaning and the pragmatics of gestures needs to be developed. Currently this is approached from different angles such as descriptive work as represented in (Poggi, 2002a) or work on coupling gesture recognition and gesture generation such as (Kopp et al., 2004).

Meaning in gesticon entries for facial expression refers to the six basic emotions (happy, sad, anger, fear, disgust, surprise) known from (Ekman, 1993) and a few other labels which are appropriate in the context of the demonstrators such as 'neutral', 'false laugh', 'melancholy', 'reproach' etc. which are inspired by (Faigin, 1990).

The *alignto* attribute is mainly used for gestures and specifies the type of entity the particular gesture shall be aligned to. This can be a sentence, a word, an accented syllable, etc. At the current stage of development of the NECA system, facial expressions are per default aligned at sentence level.

In the *aligntype* attribute it is specified how a gesture G and an entity X from the verbal communication system are coupled together.

par	G starts exactly when X starts
par_end	G stops exactly when X stops
par_adjust_to_fit	G's duration is forced to be the same as X's, i.e., they start and stop at the same time
atstress	G is aligned to the STRESSED position of X
seq_before	G is performed before X, i.e., G preceedes X
seq_after	G is performed after X, i.e., G succeeds X

3.4 The <form>-element

In the form element, information on the basic physical properties of a gesture or facial expression is specified. The form element comprises two sub-elements: the <position>-element providing information on static (spatial) aspects of a gesture or facial expression, and the <components>-element encoding information about the dynamics (the sub-parts and temporal properties) of a gesture or facial expression.

As we treat facial entries as snapshots of facial expressions, the components element is reduced to the specification of a duration range and a default duration. The position element in facial entries specifies eyebrows (up, relaxed, center down, ...), eyes (relaxed, open wide, open narrow, ...) and mouth shapes (open smile, closed relaxed, pursed, ...). These values are inspired by (Faigin, 1990). An alternative, more fine grained representation of form information of the face are the Face Animation Parameters (FAPs) used in MPEG4, see for instance (Tekalp and Ostermann, 2000). This information can be used as extra filter for selecting appropriate facial expressions during multimodal generation. In the NECA system, however, facial expressions are currently selected according to the emotion specified for the individual dialogue acts by the affective reasoning component.

Regarding gestures, the availability of information on the basic physical properties as encoded in the position element is a prerequisite for performing basic reasoning on the well-formedness of combinations of gesture. Minimal positional information is required to decide whether two gestures can be directly concatenated or whether the combination of two gestures requires an intermediate gesture for the sequence to look natural.

Information on gesture dynamics as encoded in the components element is required for the calculation of the temporal alignment of gestures to speech as well as for modulation of the expressivity of a gesture.

In the position element spatial information of gestures is encoded very coarsely specifying the position of the left and right wrists at the very beginning and end of a gesture. This is encoded by a two-dimensional grid (top, mid, down) $\times$ (center, outwards) distinguishing 6 possible positions per wrist. This information is required for reasoning on the necessary time of moving from the end-position of one gesture to the start-position of its successor. Depending on the available time and on the interpolation capabilities of the animation technology used, the information in the position element is employed to decide on either ruling out a particular gesture, directly interpolating between two gestures or inserting movements to neutral (idle) positions in between the gestures to be concatenated.

The mechanism can also be extended in order to cope with gestures that rely on the existence of specific predecessors, e.g. return-movements from special gestures. For these an attribute *special* is added to the <start> or <end>-element, and it is enforced, that only gestures which share the same *special*-value can be combined.

As already mentioned, our proposal to positional encoding of gesture information is a minimal approach. An example for a much more detailed encoding is MURML (Kransted et al., 2002). As both our gesticon structure and MURML are XML compliant, an enhancement of the

proposed gesticon entries by MURML representations is straight forward.

For the components element of gestures the following sub-elements are defined: prepare, stroke, hold, retract (cf. (McNeill, 1992)). Each of these elements has its duration element <dur> where an appropriate range and a default for the duration of the respective phase is specified in milliseconds.

Note, that a majority of our gesticon entries are gesture fragments which only comprise stroke and hold phases, whereas the prepare and retract phases result from player-specific interpolation between adjacent gestures. In general, stroke and hold are the most important phases for aligning gesture and speech. The stroke phase for instance is employed to fine-tune the timing of gesture and speech. The stroke phase is typically aligned with a particular (accented) syllable. In cases where a gesture needs to be elongated, the hold phase is of importance, as it will be unproportionally more affected than any other phase of a gesture.

3.5 The <restrictions>-element

While in the function and form elements semantic and structural aspects of a gesture or facial expression are described, the restrictions element serves as a repository for all kinds of additional constraints that specify the applicability of a particular gesticon entry in the context of a specific system. For instance, in the NECA system for each dialogue act an emotion category is calculated by an affective reasoning component (Gebhard et al., 2003) implementing the OCC model (Ortony et al., 1988). These emotion categories need to be related to emotion expressing entries in the gesticon such as facial expressions and adaptor gestures, so that appropriate nonverbal behaviours can be selected from the gesticon. This is reflected in the constraint element <constraint name="occ_emotion" val="..."/>.[4] Another example is the activation constraint <constraint name="activation" val="..."/> by means of which we specify for which affective activation level or range a particular gesticon entry is applicable.

The structure of the restrictions element is defined as follows: It holds a set of <constraint>-elements, which can be logically combined by bracketing <and>, <or> and <not>-elements (i.e. conjunction, disjunction and negation).

In the current form each constraint element just contains an attribute *name* which holds the name of a constraint and an attribute *value* or *range* that is to be used as argument of that test.

In order to facilitate the processing of the different constraints used under <restrictions> and to ensure consistency, maintainability and readability of the gesticon, a macro-mechanism is offered in the gesticon:

For the most common type of constraints, namely the lookup of a certain value already stored in the RRL, the semantic of that constraint can be specified within the gesticon itself, using a separate <constraintCode>-section.

The example in section 3.7 shows such a <constraintCode>-entry for the constraint "occ_emotion". It defines, what a program really has to do in order to test whether <constraint name="occ_emotion" val="anger"> is fulfilled: Under the current dialogueAct (this is the scope) look for the element <emotionExpressed> and test whether the value of its *type*-attribute equals "anger".

For the constraint with the name "gender" it states, that the information on the speaker has to be dereferenced and that the gender value is to be found under the element <gender>, more precisely in the attribute *type*.

This should facilitate the authoring of individual gesticon-entries and helps to keep constraint-entries consistent. The inclusion of novel constraints or changes in the structure of the RRL thus do not necessarily require changes in the code of the interpreting programs.

3.6 The <playercode>-element

Finally, the necessary mapping to player-specific gesture-code is defined in the playercode element. For the players currently used in NECA, this element is very simple. For Charamel the playercode directly points at a animation-file, for Flash it contains the key to entries in an external gesture-repository. This playercode information is embedded in the SMIL-based timing specification and forms the output to the player-specific Animation Generator.

3.7 Example Gesticon Entries

Gesture Entry

```
<gesticonEntry key ="g001"
               identifier="Thinking"
               modality="arms">
  <verbatim>
    Thinking: adaptor:  Tina: adaptor:
    moves right hand to chin but in
    addition left hand moves to
    shoulder-hight +
    palm up
  </verbatim>
  <function type="adaptor"
            alignto="sentence"
            aligntype="par" start="-200"
            meaning="think"/>
  </function>
  <form>
    <position>
      <!-- starts with D(own) O(ut) -->
      <start left="DO" right="DO"/>
      <!-- ends with T(op) C(enter) -->
```

[4] Note, that mapping between emotion categories resulting from an OCC-based approach and the basic emotion categories for facial expressions a la Ekman is in general problematic. A principled way still needs to be developed.

```
      <end   left="TO" right="TC"/>
    </position>
    <components>
      <stroke>
        <dur min="1000" default="1300"
             max="2000"/>
      </stroke>
      <hold>
        <dur min="500" default="1000"
             max="50000"/>
    </components>
  </form>
  <restrictions>
    <and>
      <constraint name="gender"
                  val="female"/>
      <constraint name="speaker"
                  val="tina"/>
      <constraint name="occ_emotion"
                  val="anger"/>
    </and>
  </restrictions>
  <playercode type="charactor"
     id="tina/char/motions/gs_thinking"/>
 </gesticonEntry>
```

Facial Expression Entry

```
<gesticonEntry identifier="happy"
               key="18"
               modality="face"  >
 <verbatim>
    flash
    eager_smile
    applicable to John and Vanessa
 </verbatim>
 <function>
    attach_to="sentence"
    aligntype="unknown"
    meaning="happy"
 </function>
 <form>
    <position>
       <eyebrows>
       <eyes>
       <mouth type="smile_open"/>
    </position>
    <components>
       <hold>
          <dur min="50"
           default="400"
           max="5000"/>
       </hold>
    </components>
 </form>
 <restrictions>
   <and>
     <or>
       <constraint typ="occ_emotion"
                   val="joy"/>
       <constraint typ="occ_emotion"
                   val="liking"/>
     </or>
     <constraint typ="activation"
                 range="1.0:0.2"/>
   </and>
 </restrictions>
 <playercode>
    type="flash"
    length="400"
    id="f_eagersmile"/>
 </gesticonEntry>
```

Constraint Code

```
<constraintCodes mapgoal="neca_rrl.0.4">
    <constraintCode name="gender"
                    typ="attributeEquals"
                    scope="speakerInfo"
                    element="gender"
                    attribute="type">
      <verbatim>
        this specifies that the gender of
        the SPEAKER has to have a certain
        value
      </verbatim>
    </constraintCode>

    <constraintCode name="occ_emotion"
                 typ="attributeEquals"
                 scope="dialogueAct"
                 element="emotionExpressed"
                 attribute="type">
      <verbatim>
        for constraint "occ_emotion":
        look under emotionExpressed
      </verbatim>
    </constraintCode>
  ...
</constraintCodes>
```

4 Conclusion

Summing up, we have outlined an overall structure for a *gesticon*, a reusable, system independent repository of gesture snippets and facial expressions relevant for the generation of dialogue accompanying nonverbal behavior. To achieve a seamless integration of gesture and language we rely on XML-based gesture representations (the gesticon) that closely interact with the RRL, a multi-modal representation structure/language used as interface to the individual system components of a multimodal generation system for spoken dialogue. Both RRL and gesticon have been developed in the context of the NECA project, but are designed to be system independent.

As regards the representation of the physical properties of gestures, our work draws upon MURML, but, for practical reasons, does not implement a similar level of detail. The general approach taken, however, allows for an extension to MURML. In contrast to MURML which concentrates on the representation of gestures, we aim at

defining a uniform representation for gestures as well as facial expressions. Moreover, due to interlinking the gesticon and the RRL, we have defined a clearcut interface to individual processing components. The linking between gesture descriptions and an XML-compliant multimodal representation language relates our work to the work described in (Ruttkay et al., 2003). Here the scripting language STEP is used to define and process gestures for h-anim[5] agents. While our aim is to separate the representation structure of a gesture repository from the processing and animation components, STEP representations are a genuine part of the STEP animation engine. Nevertheless it would be a beneficial exercise to separate out the STEP representations for gestures and incorporate the knowledge into the gesticon. On the one hand this would enhance the gesticon entries by the joint information available in h-anim and the dynamism of gestures encoded in STEP. On the other hand it would foster the understanding of which information shall be represented in a gesticon and which information belongs to a rule system for gesture generation.

As regards language and gesture coordination, the approach presented in this paper is comparable to the one pursued in the BEAT system (Cassell et al., 2001). However, other than in BEAT where thematic structure is widely used for fine-tuning of gesture assignment, we strongly rely on the prosodic information (intonation phrases, accents) directly available from speech synthesis. Another recent system for dialogue related gesture animation utilizing an XML-based framework is presented in (Hartmann et al., 2002). This work also comes with its own gesture repository.

All in all, a number of gesture repositories exist, typically being closely tied to specific gesture animation systems. Partially these repositories encode similar information, partially the information differs regarding the dimensions and the granularity of the representations. In the current situation, it would be an advantage for the work on ECAs if the community could agree on common representation structures for gesticons to decouple the gesture repositories from the individual gesture generation systems, and thus to enable the exchange of data sets. We hope that with the presented work we have made a small contribution to a common structure for gesticons which comprise definitions of elements of nonverbal communication systems (gestures, facial expressions etc.), rather than encode concrete body-specific or animation system-specific instances of such communication elements.

Acknowledgments

The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry of Education, Science and Culture and the Federal Ministry for Transport, Innovation and Technology. The work reported in this paper is supported by the EC Project NECA IST-2000-28580. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

[5] See `www.hanim.org`. H-anim agents are built in VRML (`www.web3d.org/vrml/vrml.htm`).

References

Stefan Baumann, Martine Grice, and Ralf Benzmüller. GToBI - a phonological system for the transcription of German intonation. In *Proceedings of Prosody 2000. Speech Recognition and Synthesis*, pages 21–28, Poznan: Adam Mickiewicz University, Faculty of Modern Languages and Literature, 2001.

Justine Cassell, Hannes Vilhjálmsson, and Timothy Bickmore. BEAT: The Behaviour Expression Animation Toolkit. In *Proceedings of SIGGRAPH '01*, pages 477–486, 2001.

Jan-Peter deRuiter. Gesture and Speech Production. MPI Series in Psycholinguistics. Technical report, Ph.D. dissertation, University of Nijmegen, 1998.

Paul Ekman. Facial expression of emotion. *American Psychologist*, 48:384–392, 1993.

Paul Ekman and Wallace V. Friesen. Nonverbal behavior in psychotherapy research. In John M . Shlien, editor, *Research in Psychotherapy: Vol. 3*, pages 179–216. American Psychological Association, 1968.

Gary Faigin. *The artist's complete guide to facial expression*. Watson-Guptill Publications, 1990.

Patrik Gebhard, Michael Kipp, Martin Klesen, and Thomas Rist. Adding the emotional dimension to scripting character dialogues. In *Proceedings of IVA'03*, Kloster Irsee, Germany, 2003.

David B. Givens. *The Nonverbal Dictionary of Gestures, Signs & Body Language Cues*. Center for Nonverbal Studies Press, Spokane, Washington, 2002.

Björn Hartmann, Maurizio Mancini, and Catherine Pelachaud. Formational parameters and adaptive prototype instantiation for mpeg-4 compliant gesture synthesis. In *Proceedings of Computer Animation*, pages 111–119, 2002.

Stefan Kopp, Timo Sowa, and Ipke Wachsmuth. Imitation games with an artificial agents: From mimicking to understanding shape-related iconic gestures. In Antonio Camurri and Gualtiero Volpe, editors, *Gesture-Based Communication in Human-Computer Interaction, 5th International Gesture Workshop, Genova, Italy, April 15-17, 2003, Selected Revised Papers*, volume 2915 of *Lecture Notes in Computer Science*, pages 436–447. Springer, 2004.

Stefan Kopp and Ipke Wachsmuth. A knowledge-based approach for lifelike gesture animation. In Werner Horn, editor, *Proceedings of the 14th European Conference on Artificial Intelligence*, pages 663–667, Berlin, Germany, 2000.

Alfred Kransted, Stefan Kopp, and Ipke Wachsmuth. MURML: A Multimodal Utterance Representation Markup Language for Conversational Agents. In Andrew Marriott et al., editor, *Embodied Conversational Agents: Let's Specify and Compare Them!, Workshop Notes AAMAS*, Bologna,Italy, 2002.

Robert M. Krauss, Yihsiu Chen, and Rebecca F. Gottesman. Lexical gestures and lexical access: A process model. In David McNeill, editor, *Language and gesture: Window into thought and action*, pages 261–283. Cambridge University Press, 2000.

David McNeill. So you think gestures are nonverbal? *Psychological Review*, 92:350–371, 1985.

David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago Press, 1992.

Andrew Ortony, Gerald L. Clore, and Allan Collins. *The Structure of Emotions*. Cambridge University Press, 1988.

Hannes Pirker and Brigitte Krenn. Assessment of markup languages for avatars, multimedia and multimodal systems. Technical report, Austrian Research Institute for Artificial Intelligence, Vienna, 2002.

Paul Piwek. A flexible pragmatics-driven language generator for animated agents. In *Proceedings of ACL-2003*, pages 151–154, East Stroudsburg, PA, 2003.

Paul Piwek, Brigitte Krenn, Marc Schröder, Martine Grice, Stefan Baumann, and Hannes Pirker. RRL: A rich representation language for the description of agent behaviour in NECA. In Andrew Marriott et al., editor, *Embodied Conversational Agents: Let's Specify and Compare Them!, Workshop Notes AAMAS*, Bologna,Italy, 2002.

Isabella Poggi. Symbolic gestures: The case of the Italian gestionary. *Gesture*, 2(1):71–98, 2002a.

Isabella Poggi. Towards the alphabet and the lexicon of gestures, gaze and touch. In *Multimodality of Human Communication. Theories, problems and applications. Virtual Symposium edited by P.Bouissac (http://www.semioticon.com/virtuals/index.html)*, University of Toronto,Victoria College, 2002b.

Roland Posner, Reinhard Krüger, Thomas Noll, and Massimo Serenari. The Berlin Dictionary of Everyday Gestures. Version 9 2002. Technical report, Research Center for Semiotics, TU Berlin, 2002.

Szofia Ruttkay, Zhisheng Huang, and Anton Eliens. Reusable gestures for interactive web agents. In Thomas Rist, Ruth Aylett, Daniel Ballin, and Jeff Rickel, editors, *Intelligent Virtual Agents, Proceedings of IVA 2003. LNAI 2792*, pages 80–87, Springer, 2003.

Marc Schröder and Jürgen Trouvain. The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology*, 6:365–377, 2003.

Massimo Serenari. Survey of existing gesture, facial expression, and cross-modality coding schemes. Technical Report of the NITE project IST-2000-26095. Technical report, TU Berlin, 2002.

Murat Tekalp and Joern Ostermann. Face and 2-D Mesh Animation in MPEG-4. *Signal Processing: Image Communication*, 15(4-5):387–421, 2000.

Artificial companions for older people

Gopal Vaswani
Department of design
Indian Institute of Technology
Guwahati, India
gvaswani@iitg.ernet.in

David Benyon, Stewart Cringean Oli Mival and Greg LePlatre
HCI Group
Napier University
Edinburgh
{d.benyon@, S.Cringean@, O.Mival@, G.Leplatre@} napier.ac.uk

Abstract

In this paper we give a review of our initial research on developing anthropomorphic embodied conversational agents as artificial companions for older people. We are looking at the concept of having a mobile companion that can move between devices such as a radio, TV, PDA, mobile phone, ATM, and so on. We discuss the roles and characteristics of artificial companion and issues of trust and emotional interaction between an older person and companion. Based on ethnographic studies, we developed several personas of older people that represent behavior patterns, goals, skills, attitudes, and environment of the older people. We present various cases when an older person would interact with artificial companion and visualize one such scenario. The paper also discusses the agent architecture we have developed that is necessary if the agent is to have the level of personality-rich features and the mobility that we seek.

Keywords:
Artificial companions, emotional interaction, interaction design, conversational agents

1. Introduction

The UTOPIA (Usable Technology for Older People: Inclusive and Appropriate) project is a three-year research project established to investigate technology use and development by and for older people. The proportion of older people in the population is increasing and with it the demands on long-term care and help for their particular needs. Although many older people are independent and provide much to the community, as we grow older, we will, in general, experience a reduction in our abilities. Finally many people will require support in some activities, eventually even the basic needs of life. The loss of human companions is a natural consequence of growing old. There is a diminishing of the supportive ties of family members, of friends and of other relationships from previous, concurrent, and following generations through death or distancing by migration or relocation. Furthermore, social roles and ties are lost through retirement and any parental function is reduced as children grow up and become independent. This substantial erosion of social networks inevitably leads to the loss of companions and is often accompanied by an experience of emotional impoverishment, not infrequently experienced by the elderly as a pervasive depression "without a reason" [Gory and Fitzpatrick, 1992]. With consideration of this natural decline in human companionship, the potential value of developing artificial companionship becomes distinctly apparent.

Previous work in his project has shown a wide disparity in older people's views on technology, but we hypothesise that when the technology has a perceived usefulness, difficulties with technologies can be overcome [Mival, Cringean and Benyon, 2004]. We believe that older people will welcome new technologies and services, *if* they are designed to serve their needs and aspirations. Too often designers develop technologies for the 'early adopters', packing new versions with features in order to keep this small group on board. The result is that the later adopters and 'laggards' finish up with a technology that is far too complicated and not suited to their needs.

We are employing a methodology from the discipline of interaction design [Benyon, Turner and Turner, 2004] that emphasises design for a specific group of users. Even more specifically we seek to develop some embodied conversational agents that are mobile across devices and which can act as artificial companions for older people. Our work on the UTOPIA project to date has led us to believe that such a concept might be welcomed. The current work is to build and evaluate a high fidelity prototype to see how the work should progress. The basic goal of developing these companions is to assist the older people to maintain their quality of living. Quality of life is a measure of an individual's physical, mental, and emotional well-being, as compared with their needs and capabilities [Hirsch, Forlizzi, Hyder, Goetz, Stroback, Kurtz. The ELDer Project: 2000]. Quality of life is highly individualized, and changes from day to day as a person's capabilities change.

2.Background

2.1 Embodied conversational agents and interfaces

The term 'Embodied Conversational Agent' has come to be used by researchers who aim to develop screen-based characters that are convincing in their appearance and conversational behaviour. Although this is a fast developing research area it remains largely 'in the lab' rather than in commercial or public sector use.
There is also a growth in research dealing with the creation of a new type of computer interface, which breaks away from the traditional desktop metaphor of viewing the computer as just a tool and sees it more as a partner for communication *[Janzen, 2002].* Much of this research is based on a facial representation forming the main object of interaction. While simply supplying a face doesn't make the interface social but if the face responds appropriately to a user's input this adds the social dimension *[Janzen, 2002].*

A common perception is that a conversational interface relies on the quality and accuracy of the recognition engine. However, even human speech recognition can break down, yet we are still able to communicate. Therefore, creating an effective conversational interface must also leverage other aspects of the human dialogue. Further speech recognition is not always practical and must be supplemented by other forms of dialogue interaction.

Realistic gestures must accompany the speech if the image is to be believable. These include lip synchronization and appropriate hand gestures and body movements. These, then become 'personality-rich' agents. The aim is to change an interaction into a relationship and to provide real companionship for the older person.

2.2 What makes an Artificial Companion?

Companionship is a concept that is familiar to all, yet defies simple explanation. Psychology considers it a central need, yet balks at a concise definition of what constitutes a companion beyond "a relationship…with mutual caring and trust" [Gleitman, 2000, p467]. As previously mentioned our own work to date suggests that, for older people, an artificial companion should have some utility. But at the same time should not be so utilitarian that the relationship lacks the emotional involvement need to take interaction and trun it into a relationship. We identify a number of roles that the agent should play and a number of characteristics that it should have.

2.2.1 Roles:

Assistance:

1. The companion should take care of the health of the older person. It should remind him of regular health checkups, taking medicines in time, keeps control on his diet, contacts doctor in case of emergency etc
2. Takes care of the finances of the older person (bank account, Insurance, pension etc)
3. Helps in the activities like shopping, payments of bills, security etc.

Sociability:

1. Helps or participates in the hobbies of older person.
2. Engages the older person in interesting games puzzles, quizzes pertaining to his interests and share jokes, stories and events.
4. Reminds him of any social events and helps in his social interaction.

Knowledge up-date

1. Updates the older person about the news in locality, city, country and world.
2. Helps him learn new tasks and skills.
3. Keeps him updated about the topics of older person's interest and liking.

2.2.2 Characteristics

Most of the characteristics of artificial companions are synonymous with the basic characteristics of software agents. These include functional abilities such as autonomy, communication and cooperation capability , a learning ability, reasoning capability, adaptive behaviour and (in our case) mobility across devices. Beside this there are issues of trust and emotional interaction, which are vital to the success of artificial companions. There are also the personality issues that are key to the agent's believability.

2.2.3 Trust

Must an agent have anything other than believable, or recognisable at least, social behaviour? Trust comes up repeatedly as one of the most important attributes of interaction, real or artificial. Without trust an agent working in the home will probably be useless *[Jönsson and Ingmarsson, 2002]* this is trust in the sense that if an agent is given a task it will complete it without detrimental effect to the user but also in the sense that it will be secure and reliable.

> *"Of course, it would have to be secure…a good friend doesn't give away your secrets. Unless of course, you tell it that you were going to kill yourself, and then it should tell your doctor.*
> *[Bickmore, 1998]*

If trust between a user and an agent builds, the relationship between them cements and as a result more gets done *[Marsh and Meech, 2000].* In order for this trust to build, people must believe that an agent is more than just a machine. This allows for interaction that is fully immersive and at a totally engaging level.

2.2.4 Emotional interaction

People need to connect on an emotional level and therefore it is an important consideration in the design of new interfaces. If embodied conversational agents are ever to interact naturally and intelligently with people, then they need to be able to recognise and express affect, which is an

action that relates to, arises from or deliberately influences emotions *[Picard, 1995].* One of the major area we are exploring for emotional interaction is the use of gesture both for ease of use and developing a relationship between an older person and artificial companion. We are developing some scenarios that involve gesture-based input from the user and gesture-based response from the agent.

3.Designing conversational interfaces

We are employing Interaction design methodology for developing conversational interfaces, which employ speech interface, visual design and tradition GUI principles in perfect harmony so that the older person can interact in a seamless and naturalistic manner.

Interaction Design is essentially story creating and telling. It is at once an ancient art and a new technology. Technology has always affected the telling of stories and the creation of experiences, but currently new technologies offer capabilities and opportunities not yet addressed in the history of interaction and performance. Conversational interfaces are one such innovative medium, which can transform the way people interact with technology.

3.1 Personas

The heart of the whole interaction design process is to understand our users, which in this case are older people. A persona is a user archetype you can use to help guide decisions about product features, navigation, interactions, and even visual design. By designing for the archetype whose goals and behaviour patterns are well understood you can satisfy the broader group of people represented by that archetype. In most cases, personas are synthesized from a series of ethnographic interviews with real people, then captured in a descriptions that include behaviour patterns, goals, skills, attitudes, and environment, with a few fictional personal details to bring the persona to life. Based on the ethnographic studies we developed around 7-8 personas of older people in Scotland. We present one such persona below

William
Age: 70 yrs
Retired bank manager

William is 70 yrs old retired bank manager. He is quite conversant with computers and latest technology and often uses them for e-mail and other uses. He has lived alone with his pet dog 'Tommie' for 4 years after his wife died of cancer. He had a good salary and owns many electronic gadgets and appliances. He is a good cook and normally cooks himself, although he orders take-aways occasionally. He is a Manchester United fan and likes to keep himself updated about the news in sports and politics. He is good natured and likes to watch comedy serials and chat with people. Recently he was diagnosed with diabetes and so has to visit his doctor regularly for checkups. Also he has to take precautions in his diet and take medicines on time. He loves his dog and look after it himself. He also keeps his garden green and tidy (which is also a hobby). He wants to enjoy his life and pursue his interests in his old age, but due to mental and physical degradation, he faces problems tackling everything himself and at times feels lonely and rather helpless

3.2 Scenarios

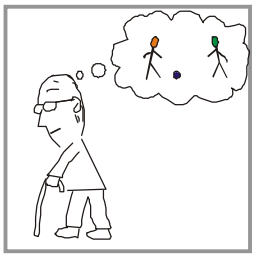
A scenario is a description of a person's interaction with a system. Scenarios help focus design efforts on the user's requirements, which are distinct from technical or business requirements. Scenarios are appropriate whenever you need to describe a system interaction from the user's perspective. They are particularly useful when you need to remove focus from the technology in order to open up design possibilities, or when you need to ensure that technical or budgetary constraints do not override usability constraints without due consideration. Scenarios can help confine complexity to the technology layer (where it belongs), and prevent it from becoming manifest within the user interface.

From an analysis of observed activities, we identified six cases when the older person would need a companion and them came out with various scenarios in each case. The six cases are:

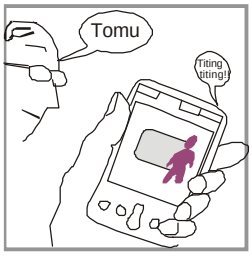
***Case1. I need help*:** This case describes situations when a older person would need help for his day to day tasks like shopping, payment of bills, learning new skills etc
Case2. I am lost: This case accounts for the situations the older person is lost and needs assistance.
Case3. I feel lonely: This describes the interaction between the companion and older person when he feels lonely and depressed.
Case4. Security: Older people are more prone to potentially risky situations that increase when they try to carry on an independent way of life. This case considers scenarios of interaction in these situations with emphasis on security of older person
Case5. I don't feel well: This is similar to the above case, but it considers the scenarios with health related emergencies of older people.
Case6. Agent's proactiveness: This case is a bit general and has caters to situations in which the companion acts proactively without the active input from he older person.

We present below one scenario from case 1, ***I need help***, and visualize the interaction in form of storyboards in fig 1.

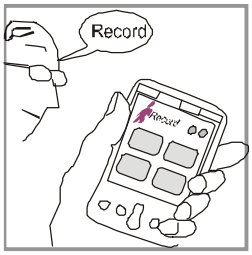
William is at his doctor's clinic for his weekly checkups. At the same time there his a match between Man United and Chelsea, which he does not want to miss. He instructs his agent 'tomu' through his PDA to record the match for him. Tomu connects to the recorder in home and starts recording.



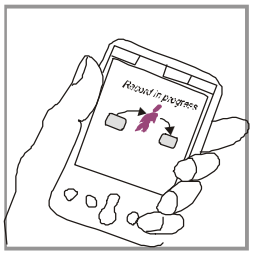
1) William is on his way to Doctor's clinic for his weekly check-up. Suddenly he realizes that at the same time there is a match between Man U and Arsenal. Being a passionate football fan, he doesn't want to miss this thrilling match.



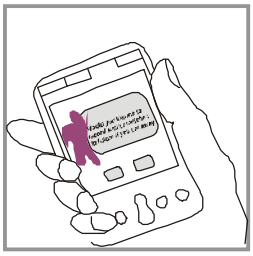
2) William calls tomu, his personal artificial companion on his PDA. Tomu gets activated with its peculiar recognizable sound. Tomu displays an easy to use interface on the pda. Tomu can be operated by speech, pen or even gestures.



3) William is very comfortable with using speech interface, but he knows, it does not work for every tasks. But this is straightforward task and he instructs tomu to 'Record'. Tomu displays some recording options on the interface and William selects the one he wants and says ok.



4) The graphics shows tomu visually moving from the PDA to video recorder. After a few moments Tomu comes back onto the PDA. "All done", he says.



5) After the completion of the task, Tomu asks him "whether he should record similar matches in future, if he is away?" Tomu always asks for feedback after any task completion, In this way it becomes more intelligent and user specific.

Fig1: Scenario Visualization of persona interacting with his companion.

3.3 Character development

Research has been done to investigate the correlation between anthropomorphism and the level of expectations that users have of an interface *[Bonito et al, 1999*]. It was found that in some cases it is better to use cartoon-like characters to suspend the user's disbelief *[Cassell, 2000]*. A human character that does not behave like a human may cause unrealistic expectations from the user. Peoples' expectations of an interface will influence how they important not to misrepresent its abilities to appear more capable than the user thinks it is. It is easier to build up confidence than to rebuild it *[Jönsson and Ingmarsson, 2002].*

It has been shown that an animal rather than a human face may be the best choice for an agent. Users have a preconception of the intelligence of animals, most likely less than their own, thus reducing their expectations of the agent *[Jönsson and Ingmarsson, 2002]* and in a study by Parise *[Parise et al, 1996*] it was found that users rated an agent that looked like a dog as more trustworthy and loyal than one with a human face. Therefore, as with facial emotions, realism is not necessarily a key attribute to an agent's success.

"it is better to go with the less realistic characters which meet the audience's expectations than to go with the more realistic characters which don't."

[Reilly, 1996]

We have developed some initial character ideas with our user personas in mind. Since we are targeting specific community (Scotland), We are also looking into the cultural factors in character development that may affect its acceptability and behaviour.

3.4 Personality

Why are character and/or personality important in an agent? According to Rousseau and Hayes-Roth *[1997]* people are captivated by personality-rich interactive characters and their research shows that

- Mood and attitude are usually recognised by users
- Users are more likely to believe in characters with consistent and non-extreme behaviour
- Users consider and can be influenced by the agent's character when they select actions to perform
- Users prefer to shape the personality of the agent rather than a having a pre-made one
- Users reacted to the agent's character in a similar way to that of a person
- Users have preferences about the type of personalities they wish to interact with

This means that an attempt to portray character in an agent should be a considered and tested process, as the wrong choices will lead to the failure of the agent regardless of its technical capabilities.

Other factors can influence users. Reeves and Nass *[1996]* demonstrated that users like agents more when the character flatters them and are more likely to interact with an agent that has a personality that closely matches their own. Morkes, Kernal and Nass *[1998]* demonstrated that computer agents, which use humour, are rated as more likeable, competent and cooperative than those that do not. Heckman and Wobbrock *[2000]* suggest that agents should have a humble personality and admit when they don't know what is going on. A summary of personality in agents can be found in "The art & science of synthetic character design" *[Kilne and Blumberg, 1999].*

Over and above issues of personality are the issues of gender, age and racial stereotypes. Reeves and Nass *[1996]* showed that users rated male-voiced agents better on technical subjects and the female-voiced agents on love and relationships even though each gave all the subjects exactly the same information. This shows that even in artificial interactions people are susceptible to applying, even if only at a subconscious level, preconceptions they use in the real world and again great care must be taken when designing a character for a specific role.

One of the major question about personality, we hope to answer through our design and evaluation is 'How do we keep the personality of artificial companion consistent along all the devices?'

3.5 Conclusions

Although the design of our companion is not yet complete, many of the key features that it needs to possess have been identified. These have come both from the literature and from our own analysis of older people specific to the purpose of this companion.

4. Architecture

The rapid proliferation of new generation wireless devices such as smart phones and personal digital assistants (PDAs) and their adoption by users across the world has lead developers to consider what new types of applications these technologies will require. Traditional models of client/server applications may no longer be suitable for these devices. Peer-to-peer (P2P) systems readily lend themselves to the more dynamic environments of these new wireless devices. P2P technology allows the creation of, membership to and interaction with peer groups. Members of these groups can offer and solicit resources by advertising them across the network therefore allowing other P2P applications to dynamically find resources they need without either an IP address or knowledge of the network itself.

We are aiming at producing personality rich agents - agents that, through the use of personality and support of multi-modal interactions such as speech, text, visual cues and gestures, engage the user in such a manner as to elicit an emotional investment in the agent. This investment transforms the interaction into a relationship. The creation of systems to enable this style of interaction presents many interesting challenges, not least in the design of the underlying architectures to support such agents.

Fig2 shows an initial architecture model for the type of agent system proposed for this work. An agent compliant network (ACN) consists of all compliant devices within the local area. These may or may not be connected externally to the Internet or other networks. If any device does have an external connection it registers itself, and by inference, all the members of its local network on that network. This allows all devices within ACN's to communicate with each other without the need for each of them to have an external connection. This architecture will be refined, amended or changed as the work detailed above is undertaken. The resulting model will be robust, flexible and be able to support the technological, functional and interaction needs of the proposed systems would need.

Fig 3 shows a proposed architecture for an agent within an individual device. This model shows various layers within each agent. The Peer-2-Peer layer handles the registration and description of a device within an ACN / external network and finds other compliant devices already registered. When an agent requires the services of another device the handshaking layer handles negotiations between the two devices and provides the characteristics,

capabilities and resources of the queried device. The agent space consists of a core agent, which represents the essential functionality needed by the system. Several supporting units supplement this core.

The interface layer, which is tailored for each device, handles all input and output for the device, allowing the core agent to be relatively device independent. This helps to ensure the consistency and credibility of the agent's personality.

This also supports a variety of input mechanisms, where devices do not support traditional interaction techniques. Content layer is responsible for collecting and collating information from different content providers or other agents. The intelligence unit supplies the AI portion of the agent so that it can perform logical decisions and user specific tasks, which are based upon and update user preferences and a collected knowledge base. This knowledge base can be stored locally or remotely depending on the resources of the device

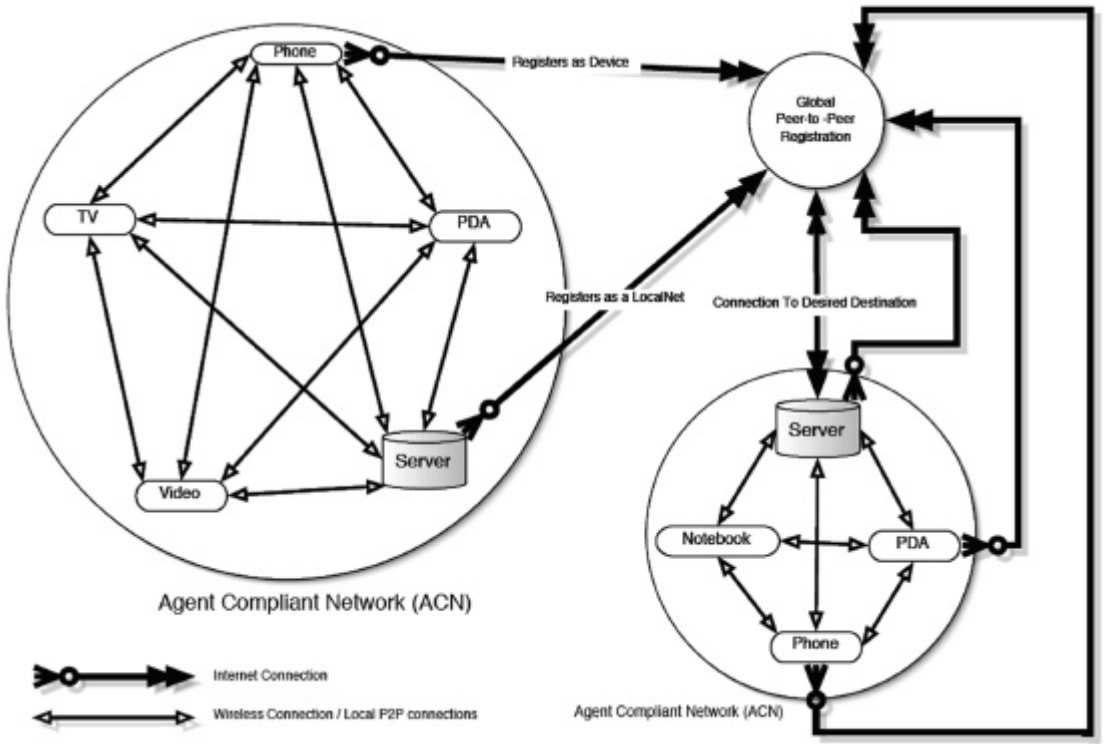


Figure 2: Agent compliant Network, local and global connection

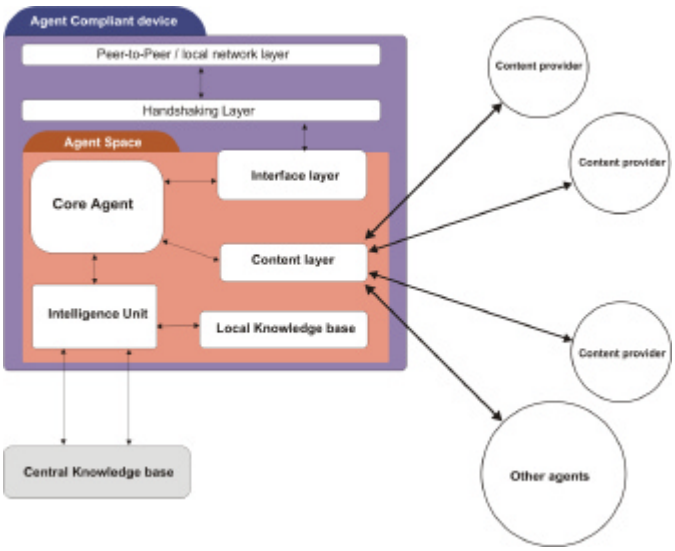


Figure 3: Agent architecture within a device

5. Conclusion

We have adopted a two-way approach to design the conversational agents that act as artificial companions for older people. The interaction design methodology discussed in the paper caters to the social and emotional segment of the artificial companions, while the software architecture side develops the technology needed.

Our **next step** is to develop some prototypes of artificial companions and evaluate them with the older people. Evaluation of systems such as these itself raises interesting methodological issues. Our approach is to make a short movie highlighting the interaction between artificial companion and older person in real life situations. It would be interesting to test how older people react to the expressive characters on the conversational interfaces. The issue of emotional interaction and personality would be the key issue we would like to test.

Simultaneously we would test our software architecture for the scenarios we have developed. Our next step will surely deepen and refine our understanding of how older people react to the concept of artificial companion. At the presentation of this paper at the symposium we expect to have some prototype interactions and some visual examples of the agents.

References

1. AVANTI, 2003. About The Project. AVANTI website http://194.203.41.27/avanti/about+the+project/default.htm]
2. Alan Cooper, The Inmates Are Running the Asylum : Why High Tech Products Drive Us Crazy and How To Restore The Sanity, SAMS; 1st edition (1999)
3. Bartneck, C., 2001, *Affective Expressions of Machines*, CHI2001 Conference Proceedings, Extended Abstracts, Seattle.
4. Benyon, D., Turner, P., Turner, S., 2004. *Designing Interactive Systems*. Addison-Wesley.
5. Bickmore, T., 1998., *Friendship and intimacy in the digital age.*. A final paper presented in MIT course, Systems & Self. [Online: http://www.media.mit.edu/~bickmore/Mas714/finalReport.html]
6. Bonito, J.A., Burgoon, J.K., Bengtsson, B., 1999. *The role of expectations in human-computer interaction*. In Proceedings of the international ACM SIGGROUP conference on Supporting group work, pages 229-238. ACM Press, 1999.
7. Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Hang, K., Vilhjálmsson, H., Yan, H., 1999. *Embodiment in Conversational Interfaces: Rea*. In Proceedings of CHI99, Pittsburgh, PA, 1999.
8. Fabri, M., Moore, D.J., Hobbs, D.J., 2002. *Expressive Agents: Non-verbal Communication in Collaborative Virtual Environments*, in Proceedings of Autonomous Agents and Multi-Agent Systems (Embodied Conversational Agents), July 2002, Bologna, Italy
9. Gratch, J., Marsella, S., 2001. *Tears and Fears: Modelling emotions and emotional behaviours in synthetic agents*. In Proceedings of the 5th International Conference on Autonomous Agents.
10. Hirsch, Forlizzi, Hyder, Goetz, Stroback, Kurtz. The ELDer Project: Social and Emotional Factors in the Design of Eldercare Technologies, in Proceedings on the 2000 conference on Universal Usability
11. King, J. A., 1995, *Intelligent Agents: Bringing Good Things to Life*, AI Expert, February, 17-19.
12. Kline, C., Blumberg, B., 1999. *The Art and Science of Synthetic Character Design*. In Proceedings of the AISB 1999 Symposium on AI and Creativity in Entertainment and Visual Art, Edinburgh, Scotland. 1999. [Online: http://citeseer.nj.nec.com/kline99art.html]
13. Mival, O., Cringean, S. and Benyon, D. R. 2004 Personification Technologies: designing Artificial Companions for Older People. *Proceedings of CHI 2004 Fringe.* ACM
14. Nwana, H. et. Al., 1996. *What is an agent*? Knowledge Engineering Review, Vol. **11**, No 3, pp.1-40, Sept 1996. Cambridge University Press, 1996 [Online:

http://www.labs.bt.com/projects/agents/publish/papers/review2.htm#agent]
15. Picard, R. W., 1995. *Affective computing*. Media Laboratory, Perceptual Computing TR 321, MIT Media Lab.
16. Reilly, W.S.N, 1996. *Believable social and emotional agents*. PhD Thesis. Canegie Mellon University.
17. Gory, K.L. & Fitzpatrick, K. (1992) The effects of environmental contexts on elderly depression. *Journal of Aging and Health*, 4(4):459-479
18. Gleitman, H. (2000) *Psychology.* Oxford Press
19. Mival, O., Cringean, S. and Benyon, D. (2004) Artificial Companions and Older People. *Proceedings of CHI 2004 Fringe*
20. Nathan Shedroff, The unified field theory of design www.nathan.com/thoughts/unified
21. T. Trover, Creating Conversational Interfaces for Interactive Software Agents CHI 97
22. UTOPIA http://www.computing.dundee.ac.uk/projects/utopia/

“To be or seem to be; that is the question” Unnatural but lively, speech synthesis for believable synthetic performers: models from acting

Christopher Newell
Department of Computer Science
University of York
Heslington, York
YO10 5DD, UK
Department of Digital Arts
University of Hull
c.newell@hull.ac.uk

Alistair Edwards
Department of Computer Science
University of York
Heslington, York
YO10 5DD, UK
alistair@cs.york.ac.uk

Abstract

The most natural interface for speech communication appears to be human face-to-face contact and significant effort is going into research towards speaking, anthropomorphic interfaces or synthetic actors. Traditionally acting has been exercised in live performances where special freedoms and constraints contribute to a unique interactive experience we may describe as ‘lively’. Interest in affective computing has led to consideration of high-level issues of believability, naturalness, and emotional truthfulness in synthetic actors. What do these largely interchangeable terms mean to synthetic actors and do they need all of them to appear lively? Can liveliness be disassociated from these other less quantifiable qualities? Human actors routinely tackle these issues but computer science with a few exceptions, has been slow to take up the opportunities offered by performing techniques to train their synthetic thespian counterparts to ‘seem’ to be lively. Unlike modern, post-Freudian actors, Elizabethan actors appear to have reduced the capture, representation and projection of liveliness to a scripting language. A technique based on their methods is proposed as an alternative paradigm for the training of expressive synthetic actors in unnatural but lively voice synthesis.

1 Introduction

It seems indisputable that the task of a synthetic expressive character in a simulated environment is the same as that of a human expressive character on stage. This is evidenced by research in synthetic acting, drawing directly on particular human acting techniques (Wavish and Connah, 1997). For human actors to learn ‘to feign, to simulate, to represent, to impersonate.’ (Zarrilli, 2002) takes many years of training and the application of a range of techniques. To non-actors this process is imagined to be mysterious, profound or even magical. Statements such as ‘You cannot be taught to act.’ are cultivated by the acting community to explain a phenomenon that only in the last one hundred years has been considered worthy of explanation. Modern actors are inclined to a folk psychological approach to acting and embrace such notions as the subtext providing the subconscious motivation for their behavioural manifestations and the super-objective driving the überaction of the play. This dubious pseudo cognitive science has manifested itself in such well-known labels as ‘The Stanislavski Technique’ (Stanislavskii and Hapgood, 1937) ‘Method Acting’ or ‘The Method’ (Strasberg and Morphos, 1987). As directors or developers of human or synthetic actor we have subliminally assimilated and accepted this as the only available formula for good acting.

The evidence suggests that there may be many alternative ways through which the principal virtues of good acting, emotional truthfulness, believability and entertainment may be exploited on synthetic stages. The cost of this new style of acting will be in the representation of naturalness. Naturalness may be an overloaded term, inappropriate either as a design objective or as criterion for the evaluation of synthetic acting.

This paper is an attempt to unravel some of the interpretive issues that are problematic in the domain of human and synthetic acting and expressiveness. Much centers on the interpretation of the criteria by which we seek to evaluate good acting. Is such acting believable? Can it be natural but not believable? Does realistic expression have to equate with believable expression? A provisional conclusion is that these desirable qualities may be best expressed and evaluated within the container term liveliness. Liveliness appears to be actable, at least by humans, and some components of liveliness – such as comic timing – appear to be quantifiable. If liveliness is promoted to the principal design objective, then other less commonplace acting styles – as well as artistic traditions – may serve as alternative models for synthetic acting. At least one his-

torical example exists that points towards the possibility of a formal description for a synthetic but lively voice.

1.1 Bad acting for synthetic characters

Within the field of research into expressive characters, actors have been used as substitute real people, able to simulate 'real, natural' behaviour and emotion for experimental purposes. This has quite rightly raised issues of the authenticity of the behavioural or emotional manifestation and raises the possibility of developing coarse [1] (Green, 1994) 'synthetic actors.' Because an acted emotion may be a caricature of the true one, the predominant view in the community appears to be that data derived from actors may be unreliable. It seems likely that the use of actors in this way will continue given the difficulty in finding an alternative, but is this the best use of actors for research into expressive characters?

In scrutinising the output of these substitute real people for modelling purposes, have we missed an opportunity to enquire more deeply into how they may be generating the output? In so doing we may find that some of the complex mapping of cognitive processes to behavioural output is already done for us. The problem with such an approach is that the modern actor, for reasons already stated, may not be willing or able to describe the process in useful terms. Most actors have little understanding of what they do, when disassociated from the drama they are seeking to perform. They may be able to articulate what they think the character needs to know to be believable, but not how these factors map to the eventual process of embodying those choices in the performance. Historically this would have been less of a problem. An 18th century actor working with a mirror and a manual of instructions would have been able to precisely map an emotion to a gesture, change in posture, grimace or vocal intonation thus anticipating the skills expected of today's animators.

It is both the creed and fear of the modern actor that these processes, reliant as they are upon (as they perceive it) complex psychological intuition are beyond meaningful description. Furthermore, the fear is that were they to be described they might be broken, through a process of destructive self-awareness. Unlike actors, dramatists and theatrical theorists have sensible methodologies for enquiry into these processes. Laurel (1991) has convincingly demonstrated the usefulness of this association between drama theory and computing. However, she is fearful that the power of the actor and their interaction with the audience will mar the 'carefully crafted uncertainty' she is anxious to preserve. Her description of the resultant clutter that occurs when the role of actors and the role of audiences become confused is probably the same clutter that prevents modern actors and audiences attempting to unpick the craft of acting for fear of unravelling it.

[1] The title to Green's book on amateurish acting is now synonymous with bad acting.

It is worth considering what questions we would wish to ask a human actor that would be helpful in developing a synthetic one. An incomplete list could be as follows:

- How do you pretend to be real?
- How do you express emotions you do not have?
- How do you make a character's behaviour natural?

This actor is required to evaluate a formidable collection of loaded concepts. Reality, expression, emotion, character and naturalness cannot be objectively or uniformly described. To explore the underlying philosophical basis upon which these notions (the principal and frequently interchangeable tasks we assign to synthetic expressive characters) are built, is outside the scope of this paper. Yet a healthy cynicism directed at any notion of a simple quantification of a universal, absolute expressiveness, naturalness, realness or believability is a prerequisite for a good human actor and should be of a synthetic one. Actors adapt their expressive techniques to the context in which they are required to act. An expressive synthetic character on a website designed to assist the user in navigating the site, would apply a different technique, with a different set of seductive functions to a non-player character (NPC) in a game. Human or synthetic actors need provide no more than their audience needs to achieve the 'willing suspension of disbelief' and this may frequently be expressed, (judged against existing criteria), as unnatural, unrealistic and unbelievable.

For example: Opera is popularly regarded as a highly expressive and affective but unnatural[2] art form. Acting in opera is principally concerned with the task of delivering the vocal sound and the instructions implicitly or explicitly provided by the composer to communicate affectively. Opera singers train for years to give precise representation to the expression of musical emotion through physical and vocal means. This usually results in a style of stylized 'heart on sleeve' expressiveness that is entirely appropriate to that context but frowned upon by those of the 'Method' school. Conversely most film or TV acting models itself on the accurate representation of an amplification of real life and is intolerant of overtly expressive performances or stylisation. The assumption in this case is that expressiveness is an externalization or embodiment of an internal, usually deemed realistic, psychological truth. Good 'acting' provides a good fit in either case. This is an illusion and has had an impact upon the popular misconception of acting methods.

It is of interest to note that the two 19th century literary and dramatic traditions of naturalism and realism that gave rise to the Stanislavski method, required two distinct conceptual bases from which to understand them. Naturalism was concerned with the representation of low-life struggle, heroism and despair while realism was principally an attempt to capture psychological complexity.

[2] Nobody in real-life sings a song as they die of consumption. In Verdi's La Traviata and Puccini's La Boheme they do.

Neither was concerned with capturing real-life. Just as the visual arts have moved through levels of representation from photographic hyper-realism to abstraction, it is only for about the last one hundred years that expressive acting has been targeted towards objective realism. Actors, rather than dramatists or theorists, know not to expect to be real. Only very rarely do they attempt it, and even less often do they achieve it.[3]

If aspirations toward realistic, natural and believable embodiments are deemed un-actable in humans and therefore likely to be un-computable in synthetics, what useful seductive qualities could be successfully and meaningfully synthesised in an expressive character?

2 Synthetic liveliness

Synthetic liveliness is an expression designed as much to exclude the other more complex synthetic acting attributes already discussed, as it is to include new ones.

These exclusions are made to avoid comparison with other work that may be broadly characterised as 'attempts to pass the Turing Test'. The Turing Test is arguably an acting test anyway, but its principles run much deeper than this research. No attempt will be made to describe the simulation of intelligence, sociability or empathic qualities (Hayes-Roth, 2003) in synthetic actors. The objective is to describe techniques to give the illusion of lively synthetic actors behind a digital, insurmountable 'fourth wall'.[4] Roth's other qualities may be implied by the successful synthesis of liveliness but that will be serendipitous.

How does this differ from a determinist solution such as a fully scripted and animated performance or the acting methods of the 18th century described above? It would always be different only in so far as the envisaged technique would enable the expressive characters to generate a convincingly lively performance of any text, in any environment and in keeping with general assumptions of liveliness. Like real acting the technique would facilitate a progressive process of rehearsal during which anomalies would be eliminated but in performance serendipitous outcomes to environmental variables would also be simulated.

2.1 Modeling liveliness

To attempt to scrutinise and model the cognitive processes of a human actor, label them as lively and reproduce those processes in their synthetic equivalent is going to prove difficult when there is no scientific research or formal evaluation of stage acting methods to base a model upon.

A simple approach may be to scrutinise and model the output of human stage acting after the essential cognitive processes have taken place and at the point when the audience interact with the actor make an evaluation. Perhaps the only realistic evaluation at this point is to quantify laughter as the least conscious and most clearly uncontrollable reflex action likely in these circumstances. We know applause to be meaningless, objective appreciation unreliable and the critics unassailable. It is in this respect that the notion of liveliness as a computable characteristic of good acting leading to laughter becomes a realistic possibility.

The 18th century, upside down approach (modelling the effect rather than the cause) accurately represents the limit to which human-like acting should aspire. Of surprise to some will be the fact that many modern actors work in this way. Examples are to be found in the carefully crafted and planned mistakes that audiences take to be chance occurrences, plants and hecklers that have been rehearsed and the laborious length to which 'out-takes' are planned.

2.2 Synthetic liveliness in music

Other potential models are to be found in the performance of music. Lively (allegro) in a musical performance principally implies speed but also mood. A lively song is rather fast; a natural song may be any speed. A lively song is probably not that quiet, a believable one could be. A lively song is not sad, an expressive one may be. To tell a musician to be natural would tell them nothing; to tell them to be lively would be an instruction they may(possibly with some difficulty) be able to follow. They would probably speed up the delivery of the notes, move about a bit more on stage and try to be more light-hearted. Liveliness in music has very clear and performable associations.

In the field of the communication of emotion in musical performance, Juslin proposes a formal distribution for code usage (the mechanisms of music) in conveying emotion. He explores the five most studied emotions (happiness, sadness, anger, fear, love/tenderness). Using data from other studies in which participants rated the valence and activity of emotional terms, he plots the placement of each emotional expression to a set of precise acoustic descriptions of the musical mechanisms. Thus, sadness is mapped to low sound level, low activity and negative valence. More controversially for musicians, he suggests a quick and dirty shortcut to an expressive performance and suggests that:

> The single cue combination that was most expressive according to listeners had the following characteristics (with cues in order of predictive strength); legato articulation, soft spectrum, slow tempo, high sound level and slow tone attacks. This raised an interesting question; is this how performers should play in order to be judged as expressive by listeners? (Juslin, 2001)

[3] Exceptions may be actors in "The Blair Witch Project" and the current advertisement for Volvo cars "The mystery of Dalarö"

[4] The imaginary wall represented by the apron of the stage, as opposed to the other three walls which are embodied by physical scenery.

While this example is not of immediate relevance to synthetic liveliness, placed as it is at an opposite expressive pole, it suggests that similar shortcuts may exist for a range of emotive effects, perhaps including composite effects such as liveliness. Frick (quoted in Juslin) used the concept of 'prosodic contours' to refer to dynamic patterns of voice cues over time found in vocal expression. How these cues express emotion requires an explanation and Juslin suggests applying a functionalist perspective and concludes that: 'music performers are able to communicate emotions to listeners by using the same acoustic code as is used in vocal expression of emotion.'

Actors instinctively know this to be the case and freed from the requirement to be natural, as they were until the last century, will apply this principle and use musical constructs: intonation, timing, empty bars, syncopation, dissonance and timbre to coldly synthesise emotional expression and liveliness.

2.3 Lively, life-like and animated?

Can a lively embodiment be no more than a synthetic vocal construct or are there other essential characteristics of liveliness? As already stated the term lively is a container term for many others. Liveliness as a container of the notion of life-like or living is clearly problematic in a simulated environment. The majority of the research into believable, virtual actors has been applied to animated embodiments or avatars that present some external lifelike manifestation, usually as a graphical animation. Prendinger makes the point that; 'The concept of "life-likeness" is certainly not restricted to animated agents.' (Prendinger and Ishizuka, 2003). According to Hayes-Roth and Doyle (Hayes-Roth and Doyle, 1998) 'Animate characters share all the features of life like characters except for their embodiment; that is, animate characters are not necessarily animated, but can still be perceived as perfectly life-like'. If concepts of animation and life-like may be disconnected, it may be possible to disconnect concepts of liveliness and life-like. This is keenly demonstrated in the context of live performance where the experience may be lively but the source may, in part be dead.

2.4 Lively live performances

A live performance is presented by living people and experienced in real time by other living people? It is effectively the opposite of a recording experienced by a recorder. It follows that a film can never be live and neither can a computer game. Can a performing dog- show? Most people would say yes. In which case a live performance requires living things as performers and we can assume living things as viewers. Beyond that a live performance can have dead parts. A recent single released by Robbie Williams saw him duet with a film recording of the deceased Frank Sinatra. Sampling recordings by dead (or absent) artists for integration into live performances by living artists is core to modern dance music. This suggests there are degrees of live performance. Any performance will have a combination of live and dead parts.

Researching the performance of classical music, Widmer's objective is 'to develop quantitative models and theories of musical expression'.[5] One method employed is to systematically extract rules from the performance of Mozart sonatas by a great human player. It is hoped this will enable the computer to perform the sonata as if it were a human player. After achieving this very difficult task, the theoretical question that will arise on the inaugural performance by the computer player will be: Is the computer presenting a live performance or a recording? The information is recorded as digital musical data, processed in real time according to rules extracted from live human performance and output in real time to a live human audience. It is both live and recorded but may be as lively as its original exclusively human model.

This ambiguity between that which is quantifiable and that which is uncertain or serendipitous presents one of the most difficult problems for the successful synthesis of liveliness – and the least developed at this stage of this research.

2.5 Serendipity in performance

'[Theatre is] The stimulation to imagination and emotion that is created by carefully crafted uncertainty.' (Laurel, 1991)

Laurel says of computing: 'People experience the requirement for precision as troublesome and often drown in precision because of the complexity and artificiality of its expression.' Of theatre she says 'the imprecision of dramatic representation is the price people pay often quite enthusiastically – in order to gain a kind of life-likeness, including the possibility of surprise and delight.' She qualifies this suggesting that good theatre is 'Effective agency, in worlds in which the causal relations among events are not obscured by the randomness and noise characteristic of open systems' like "real life". (Laurel, 1991)

For synthetic liveliness her final qualification presents a problem. Surprise or serendipity appear to be products of the same open systems that generate the noise that Laurel is keen to avoid. Theories of creativity (Boden, 1995; Perkins, 1996) recognize the importance of randomness in creative systems. Perkins' 'Klondike space' explores the necessity of chance occurrences in developing creative strategies, in this case to find gold. Harnad emphasizes 'cerebral serendipity' as a class of mutation theories that '...emphasize the crucial role of chance in creativity.' (Harnad, 1990)

We may suppose that serendipity, spontaneity and chance have a crucial role in the synthesis of liveliness.

[5] Computer-Based Music Research: Artificial Intelligence Models of Musical Expression. START Research Prize Project, Austrian Research Institute for Artificial Intelligence (ÖFAI).

Stochastic variables have provided quick and dirty shortcuts to 'humanness' in de-quantisation[6] algorithms used in midi sequencing packages such as Cubase[7]. Most musical humanisation algorithms depend upon inserting random fluctuations in timing or touch. By allowing the user to adjust the degree to which notes are shifted against a temporal grid, the imprecision of human playing is simulated.

Developing an appropriate formal model to propagate serendipitous rather than simply random outcomes at runtime is clearly important. Campos and colleagues provide one potential lead. They acknowledge that it may not be possible to program serendipity but that it is quite possible to '...program for serendipity, that is to induce serendipitous insights through the use of computers' (Campos and de Figueredo, 2002). This research is ongoing.

Uncertainty has it own set of research domains. Complexity theory, chaos theory, and emergence evolve formal descriptions of informality and spontaneity. These big theories have a curious seductiveness for many in the performing arts and are frequently misunderstood. Cohen and Stewart (Cohen and Stewart, 1994) suggest that the edge of chaos may represent the point of maximum creativity within a system. Performers are familiar with performing along this edge between the formal constraints of the pre-built system (the score, the script) and a disastrous slip into error and chaos. It manifests itself in making daring choices from the available interpretive options. This may be extending a sung note beyond the notated duration, lengthening a silent bar or a pause, adding an untried variation or cadenza, increasing or decreasing the tempo or adjusting the dynamics. Performers occupy this space when they are at their bravest, most inspired and I would suggest liveliest. It's a risky strategy and there is a high probability of error.

At this stage in the research perhaps we should conclude no more than 'liveliness' is a convenient container term for a number of properties that appears to positively enhance the simulation of other more complex properties of believability, naturalness and expressiveness and even truthfulness.

3 Better acting for synthetic actors

To recap: The task of acting is 'to feign, to simulate, to represent, to impersonate.' (Zarrilli, 2002) To achieve this most actors today apply a technique. The best-known and most influential contemporary acting technique is the Stanislavski method. Core to Stanislavski's method was the notion of believability and truth. The actor's skill is evaluated according to the degree to which they represented truth. Truth was to be found primarily by tapping into the unconscious, psychological motivation, the subtext that informed the external expression (behaviour) of the character and the action of the drama. Hollywood and TV exemplify this approach and it is the method assigned by default to synthetic actors. Kirby (quoted above) proposes a continuum of behaviour from not acting, to acting and it may be useful to position our synthetic actors upon this continuum. At one extreme is the sort of character acting intended by Stanislavski; at the other, Kirby suggests, are the characters assigned by the audience as character actors because of the context in which they are embodied. It is this latter group that may be of most interest to directors of synthetic actors and as significant representatives of this group we may refer to Elizabethan and Jacobean actors.

3.1 Elizabethan and Jacobean verse drama

Because of the circumstances in which the plays were presented, playwrights had to encode instructions for expression, timing, intonation and other musical and vocal mechanisms into the text. Actors could interpret and execute these on the fly with no rehearsal and incomplete knowledge of the creative domain. This may map much more accurately to the abilities and requirements for successful synthetic acting than the more recent acting methods.

The conditions in which Shakespeare's works were first performed were unlike those of today. A strong rhetorical tradition gave the words a pre-eminent position in the spectacle and the majority of the text was in verse. The plays were performed in the open air, during the day and under natural light. The female roles were all taken by boys or men and the costumes were contemporary dress accessorized appropriately to suggest the status of the character. There was no director but a bookkeeper onstage kept the actors on track. Rehearsals were barely sufficient for the play to be run in its entirety before the first performance. Actors would only be supplied with a cue script with their own speeches and cues. To ensure that the actors knew the story and who was playing what role a plaque would be hung backstage providing a cast list and synopsis of the story. Shakespeare had to provide code for his autonomous acting agents that was, simple, robust and versatile. He was also at pains to constrain their extemporizations, particularly those of the lead comic. For a full history and analysis of the acting methods of the time see Tucker (2002).

This is hard for a contemporary audience to imagine, brought up as we are on a very precious and complex approach to the works of the Bard. The quick and dirty system used to present his work seems shockingly unsophisticated. Underlying this construct and of most relevance to this research was the verse.

[6] De-quantisation or humanization are not terms customarily employed within the electronic music industry. For example in the sequencing package Cakewalk (Copyright 2003 Twelve Tone Systems Inc), this property is referred to as swing.

[7] Cubase: Copyright 2003 Steinberg Media Technologies GmbH.

3.2 Verse structure driving voice synthesis

Shakespeare's verse is written in iambic pentameters. Five iambs or weak strong units per line of verse. 'If music be the food of love play on'[8] is an example of a perfect iambic pentameter. The iambic pentameter provided the basic metrical unit from which all of Shakespeare's verse is constructed. The verse structure could carry a number of different instructions rather like a scripting language, including instructions for the synthesis of the voice:

The meaning of the words. Shakespeare's language was complex even for its time, and there is no evidence that the young boys who played the female roles would have the vocabulary to know what it meant. There is evidence that the Elizabethans spoke as plainly as us and that the language used by Shakespeare would have been as difficult for a young actor to interpret as a line of computer code is for the average actor today. In the example from Twelfth Night. The verse instructs the actor, that the line is about music, food and love and is an instruction to another person to continue to play. It is not about 'if', 'the' or 'play'.

The point at which to change tone. Full stops although added by editors can indicate a change in tone and provide an actor with a road map of tonal changes that gives variety and emotional range to the speech.

The tone with which to respond to another actor's cue. The lineage of a line e.g. a line of verse split between more than one actor would require that all actors adopted the same tone and maintained the prosodic contour of the iambic pentameter. In this way the system reflected the natural human tendency to mirror the tone of a conversational partner.

The speed of delivery of a line. A monosyllabic line implied a slower delivery in order to maintain clarity.

When to breathe. In an open theatre it has been found that one human breath is capable of projecting on line of verse at full voice before requiring another breath.

When to leave the stage and when to enter. Scenes would frequently end with a rhyming couplet that indicated to the actor on stage that the scene was over and the next scene was about to begin.

3.3 The text itself

In combination with the verse structure, Shakespeare would use clues in the text itself to carry instruction to the actors.

Help! What's going on or who's he? Narrator figures would be used to summarise the story the action and the state of the character for audience and actors.

Help! Who is the boss? Modes of address. The familiar thou/thee compared to the formal you/yours would indicate the relative status of the character and possibly the emotional state.

[8] From Twelfth Night.

When to syncopate the rhythm. Alliteration, assonance and repeated words would be used to vary the effect and to bend the rules.

When to emphasize a word or phrase. Peculiar rhymes, unusual lineage, or repeated words may all indicate that the actor should indulge himself a little.

Shakespeare tells us in the verse how he expected his actors to act. In Hamlet he seems to lecture them and us on his expectations. The first part of his lecture, in prose, is as follows.

> Speak the speech, I pray you, as I pronounced it to you, trippingly on the tongue. But if you mouth it, as many of our players do, I had as lief the town crier spoke my lines. Nor do not saw the air too much with your hand, thus, by use all gently, for in the very torrent, tempest, and (as I may say) whirlwind of your passion, you must acquire and beget a temperance that may give it smoothness.

Hamlet is instructing a group of traveling actors to act naturally in order that the affectiveness of their performance communicates most powerfully. To do this he urges the actors to 'beget a temperance that may give it smoothness.' and not to indulge themselves; thus constraining the license he may have provided in the verse.

3.4 Uneven distribution of knowledge

As a 21st century user of theatre we generally expect the actors to know more about the performance than we do. This is not universally true, but for the sort of scripted agent that is the subject of this research we may take it to be true. For the Elizabethan dramatist the distribution of knowledge between actor and audience was probabilistic.

- Most of the plays were based on known stories or historical events. Some participants (audience or actor) would be familiar with the narrative to others it would be new.
- Some players may have played the role before, others not.
- Rehearsals of key moment, fights and dances would have taken place but less dangerous moments would be un-rehearsed.
- The player may have learnt the line or would have the bookkeeper say it.

With a star system that would have licensed the lead comedian to extemporize and a fondness for borrowing special effects including real bears and dogs the crucial management role was assigned to the verse. This rough and ready system set against the rigor of the verse must have produced a very lively event. We can imagine a young actor playing Juliet not knowing that the Friar would be a coward or even that Romeo was to die. We may be

fairly sure that actor and audience would be hearing jokes for the first time during the opening performance and that they may have corpsed, dried or fluffed as a result. The structural temporal constraints of the verse held everything together but the system was designed to encourage the emergence of liveliness and programmed for serendipitous opportunism.

Elizabethan and Jacobean verse drama provides us with a potential model for a new type of expressive synthetic voice and points toward a new type of synthetic dialogue: A synthetic dialogue that could be generated in real time, requires only dumb autonomous agents, requires only partial knowledge of the creative domain and has robust error correction capabilities. Its principal virtue in synthesizing liveliness may be in its transparency and indeterminacy.

4 Conclusion: Seeming to be lively

To reassign the problem of creating believable, expressive synthetic characters to an acting problem that an antiquated acting technique will fix would be a significant over-simplification. To tentatively redefine our mutual objectives as the synthesis of liveliness, may provide impetus to further debate toward providing a more realisable framework for training synthetic actors to seem to be. If we can agree that the costuming, language, gesture, expression, behaviour and every other acted projection of notional reality ascribed to synthetic actors is rooted in their version of physical life it will accordingly be unlikely to match with ours. It therefore seems probable that naturalness, believability and expressiveness as aspirational labels will cease to be useful and something like liveliness will provide a better fit. Good acting for humans or synthetics is a challenging discipline. This is evidenced in the range of different approaches currently being researched (Trappl and Petta, 1997; Prendinger and Ishizuka, 2003) and the number of bad actors flowing from drama schools. From the viewpoint of someone who is positioned to one side of the research community in this field, the objective in this paper has been to dispel some illusions about the nature of human acting – which the theatrical community are rather keen to maintain. The hope is that, in return, the scientific community chooses to address one or two of the fundamental issues that continue to vex the performance community:

- What do we mean by expressiveness, believability and naturalness and do we need all of them to achieve one of them?
- Can a lively performance be embodied in a non-human entity? Are we in danger of being replaced?
- Can you quantify comic timing? Some actors think you can.

References

Margaret A. Boden. Creativity and unpredictability. *Stanford Humanities Review*, 4, 1995.

Jose Campos and Antonio Dias de Figueredo. Programming for serendipity. In *AAAI Fall Symposium on Chance Discovery*, 2002.

Jack Cohen and Ian Stewart. *The collapse of chaos: discovering simplicity in a complex world.* New York, Viking, 1994.

Michael F. Green. *The art of coarse acting, or, How to wreck an amateur dramatic society*. London, Samuel French, 1994.

Stevan Harnad. *Creativity: Method or Magic?* Unpublished manuscript, 1990.

Barbara Hayes-Roth. What makes characters seem lifelike? In Helmut Prendinger and Mitsuru Ishizuka, editors, *Life-like Characters. Tools, Affective Functions, and Applications*, pages 447–462. Springer, 2003.

Barbara Hayes-Roth and Patrick Doyle. Animate characters. *Autonomous Agents and Multi-Agent Systems*, 1: 2, 1998.

Patrik N. Juslin. Communicating emotion in music performance: A review and theoretical framework. In Patrik N. Juslin and John A. Sloboda, editors, *Music and emotion : theory and research*, Affective Science, pages 309–340. Oxford University Press, 2001.

Brenda Laurel. *Computers as theatre*. Addison-Wesley, 1991.

D. N. Perkins. Creativity: Beyond the darwinian paradigm. In Margaret A. Boden, editor, *Dimensions of Creativity*, pages 119–142. MIT Press/Bradford Books, 1996.

Helmut Prendinger and Mitsuru Ishizuka. *Life-like Characters. Tools, Affective Functions, and Applications*. Springer, 2003.

K. S. Stanislavskii and E. R. Hapgood. *An actor prepares*. London, G. Bles, 1937.

Lee Strasberg and Evangeline Morphos. *A dream of passion: the development of the method.* Little Brown and Company, 1987.

Robert Trappl and Paolo Petta. *Creating personalities for synthetic actors: towards autonomous personality agents*. Springer, 1997.

Patrick Tucker. *Secrets of acting Shakespeare: the original approach*. Routledge/Theatre Arts, 2002.

Peter D. Wavish and David Connah. Virtual actors that can perform scripts and improvise roles. In *First International Conference on Autonomous Agents*, pages 317–322, 1997.

Phillip B. Zarrilli. *Acting (re)considered: a theoretical and practical guide*. Routledge, 2002.

To tell or not to tell...Building an interactive virtual storyteller

Andre Silva*
*INESC-ID and IST, Tagus Park Av. Prof. Cavaco Silva, 2780-990 Porto Salvo, Portugal
andre.silva@gaips.inesc.pt

Guilherme Raimundo
†INESC-ID and IST, Tagus Park Av. Prof. Cavaco Silva, 2780-990 Porto Salvo, Portugal
guilherme.raimundo@gaips.inesc.pt

Celso de Melo
‡INESC-ID and IST, Tagus Park Av. Prof. Cavaco Silva, 2780-990 Porto Salvo, Portugal
celso.melo@gaips.inesc.pt

Ana Paiva
§INESC-ID and IST, Tagus Park Av. Prof. Cavaco Silva, 2780-990 Porto Salvo, Portugal
amp@inesc-id.pt

Abstract

Storytellers do not always tell the story the same way. They observe their ”audience”, see their reactions and adapt the way, the gestures, the posture and the content of the story, to better respond to the audience’s reactions. Clearly, this adaptation is however not only in the content of the story but also on the way the story is told, thus the facial expressions and the gestures of the storyteller. In this paper we describe a synthetic character that tells stories interactively, discussing how the adaptation to the audience is done through the content of the story and the associated expressions and gestures of the character.

1 Introduction

Any storyteller plays a fundamental role in children’s stories, dragging them into the story, keeping their attention and freeing their imagination. In fact, a storyteller can turn a story into a good or a bad one. The use of the voice, facial expressions, and the appropriate gestures, are basic ingredients for transforming the content of a simple story into the most fascinating narrative we have ever heard. But this need for a storyteller to be expressive, to balance the words and the tone, to use gestures appropriately, poses major research challenges if one aims at building a ”synthetic” storyteller. However, recent developments of embodied agents such as [2], [3], [6] and [8] among others, have shown amazing advances, which allows us to consider the technical challenges for building a virtual storyteller can in fact be overcome and achieve, under limited circumstances, a believable storyteller.

In fact, in [10] and recently in [9] a simple virtual storyteller was presented and its interactivity discussed. Our ultimate goal is for the virtual storyteller to be able to tell the content of a story in a natural way, expressing the proper emotional state as the story progresses and capturing the user’s attention in the same way a human storyteller would.

In the work here presented, we will show how we have adapted the storytelling (both content, gestures and expressions) to the interactivity in a virtual storyteller.

This paper is organised as follows. First we will describe the idea for the storyteller. Then we describe the character, the structure and contents of the stories embedded in knowledge and the gestures of the character. Then we describe how the user will influence the stories being told, and the gestures of the character.

2 Papous, the storyteller: The Idea

Real human storytellers do not always tell the story the same way. They observe their ”audience” and adapt the way they are telling the story, their expressions and gestures to better respond to the audience’s reactions. This means that the storyteller gets feedback from his audience and uses that feedback to shape the way in which the story is delivered.

So, our main research question is: how can we adapt the gestures and expressions of the character to be able to adequately respond to the audience.

To do that, we have built an interactive storyteller that adapts to the user’s input using a tangible interface as shown in Figure 1.

The user supplies the virtual storyteller with certain input (in this case ”postcards”, which will allow the character to decide how the story should be told. For instance, the user may decide he wants to hear a more terrifying version of the story, supplying this simple wish to the virtual storyteller. The virtual storyteller is then responsible for choosing the course of the story that is most suitable for the user’s input and for adapting his visible behaviour to the user’s intentions. The storyteller gestures and expressions change in order to look frightened and scared in order to create the right response from the audience.

To do that, the architecture presented contains a set of modules, in particular:

- *(1) the Input Interface module* is the component re-

sponsible for receiving the input from the user and then handling it by organizing it for future processing by the story engine. Also, since the input can be supplied at any time, it is also this module's responsibility to store it in a coherent way when the Story Engine module isn't ready to process it yet.

- *(2) The Story Engine module* contains the story itself, parsed from a story file. This module is responsible for parsing the story, organizing it and maintaining the necessary information to decide how to tell it, according to the input received from the Input Interface module.

- *(3) The Character Engine* is the component that handles all the processing of the story bits that need to be told, thus guaranteeing that the synthetic character performs the adequate actions, moves and gestures to convey the desired meaning. Moreover, the Story Engine tells the Character Engine how it should set the character's behaviour in order to maintain coherence with the direction that the story will take.

2.1 The Character

The storyteller is a synthetic 3D granddad (inspired in a set of TV granddads such as the Tweenies and "Avô Cantigas" (in Portugal). The character takes advantage of its voice, gestures and facial expressions to convey the story content to be told. The character's behaviour and the way the story is narrated, is influenced by the user's input.

The facial expression engine follows the MPEG-4 standard [7] in which the six universal emotions of joy, sadness, angriness, disgust, surprise and fear [4] are contemplated. According to the present emotional state of the character the emotional facial displays are blended together. In this way it is possible to convey several emotions simultaneously having each contribute to the final output with a specified weight.

At the moment the lip-synch of the model is still very simple. When the voice is heard the facial engine generates random visemes (the visual equivalent of a phoneme). Due to this random nature the visual output you get is cartoon-like where the facial display doesn't match exactly the audio. However, due to a cartoon-like appearance of Papous, although quite simple, this approach, still widely used, leads to quite satisfactory results.

3 Interactivity

As the virtual storyteller progresses through the levels, it must choose which StoryBit to narrate, according to the user's input. There are two underlying problems to solve here: (1) the first one is the user's input. What type of input can we get from the user? The second one is navigation. That is, given the user's input, how will the storyteller decide which bit to pick next, maintaining the coherence of the story?

3.1 User's Input

Concerning the first question, we decided to use a tangible interface that will get the user physically involved with the story. We investigated the use of several different types of input, such as voice or even SenToy [1]. However, due to the fact that we wanted to provide some "story meaning" to the input, and at the same time get the user physically involved, we decided to use the Influencing box [5], as we could associate "images" and meanings to the input. With the box, the user just needs to insert illusive cards (which are tagged with bar codes). The user may choose the most appropriate card for a particular situation (for example, choose a scary sign or a forest) which will then influence the whole story. Figure 1, depicts the use of the Influencing box as input for the system.

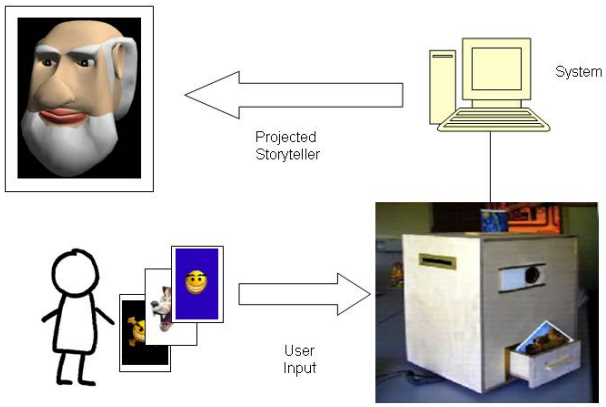


Figure 1: Input using the Influence Box

The user stands in front of the box, which allows him to insert the cards without much effort, while the character itself is projected onto a wall. Before the story commences, the user can insert cards too, and this input information will be used to decide how the story will start, a kind of setting up the scene (this "pre-story" input will only be considered by the system if the human author has decided to provide more than one way for the story to begin).

Inside the Influencing box, these cards are identified and their meaning is sent to the Input Interface module for processing (shown previously in Figure 1).

There are four types of cards: Propp function card's; character cards; mood cards and scene cards. All the cards are available all the time, which provides a large diversity on the possible progressions in the story. The user

can supply these cards at any time, and their meaning is stored by the system until it becomes ready to use that information, which means, when it becomes necessary to choose the next Storybit for narration.

3.2 Adaptation to the Audience

The input supplied by the user is used to guide the virtual storyteller, helping him decide which StoryBits to choose and the gestures and expressions to make.

Although the stories are quite simple in terms of non-linearity, given the knowledge structure associated to each StoryBit (which contains basically a small piece of story), the process of navigating through StoryBits is quite rich and is based on heuristics that rely on the StoryBits properties (the emotion/mood of that bit, the characters, the scenario and the Propp function of the bit). All this process of deciding what to tell next and how to tell is based on a desirability factor that all StoryBits have. Naturally, this desirability factor is not a constant and depends not only on the input received at each moment as a card but also from the previously narrated StoryBits.

Thus, with each level transition, the virtual storyteller calculates the desirability factor for each available StoryBit and then decides which one to choose. The system calculates the desirability factor by trying to match as many available StoryBit properties as it can with the user's input (the chosen StoryBit is the one with the highest desirability factor). For instance, if the user wishes the story to be told in a happier way, then the virtual storyteller will try to accommodate his wish by choosing the appropriate StoryBits in the future.

The user inputs themselves are received and weighed, taking into account the time when in which they arrive and their type. This means that the most recent input has a bigger part in the decision that the storyteller has to face with each level transition.

For instance, let us assume the user decides that he wants the story to be told in a scarier way and provides the necessary input. He can then change his mind and decide he wants the story to be sadder. The user's most recent input is more important for the choice of which StoryBit to narrate and therefore, the next StoryBit will be coherent with his later choice, making the virtual storyteller tell the story in a sadder way.

3.3 Adaptation of the storyteller's gestures and expressions

Emotions are intangible, intimate and personal. They are also the main elements that need to be changed in the course of a narration. However, to express emotions, one requires the adequate and coordinated use of vocal, facial and body expressions. Furthermore, to express emotions in a manner that can be easily perceived by the audience, requires an understanding that artists in Dramatic Expression and Animation seem to dominate. Thus, a good storyteller also needs to express himself dramatically. So, if user input affects Papous' internal emotional model, then one important question is: How will emotional state change reflect on its expression? If the storyteller is sad its gestures should be slow and narrow. However, if it is enthusiastic its gestures should be energetic and wide.

Our approach to this problem relies on two main aspects: (1) high-level parameters that define movement qualities; (2) A theory that correlates emotions to these parameters.

For instance, the user may express that he or she would want the story to be told in a more scary way. Therefore, the character's emotions are affected by the user's choices. Consequently the virtual storyteller's visible behaviour (its facial expression, voice and gestures) is also influenced by the user's input. The character's verbal output is affected gradually, allowing for several levels of emotional change in the voice. This is done by changing the Text-to-Speech system's parameters (the Eloquent TTS system), adjusting them so they convey the emotion the character is trying to express. For instance, the character can be mildly sad or very sad. The same concept was applied to the facial expression of the virtual storyteller.

4 Results and Final Comments

Although still at at the beginning, a preliminary usability test was made with sixteen children, with ages between nine and ten years old, in order to evaluate the text to speech engine, the facial model and facial expressions and the tangible interface, when compared with a classic menu. The test consisted on the narration of the story of Little Red Ridding Hood and a questionnaire that children answered after hearing and interacting with the virtual storyteller.

Based on the given answers we were able to draw the following initial conclusions:

- The application is of easy use
- The main obstacle to the understanding of the story is the quality of the text-to-speech
- The "looks" of the story teller were nice
- Sometimes, the facial expressions were not completely identified by the children
- The tangible interface was a success (fourteen out of sixteen children chose it over the classic menu interface)

As the story teller is still at an early stage, there are obviously several aspects that need to be improved. In particular, the TTS, which we need to consider also the use of a human recorded voice provided by a professional actor.

Secondly, as the final version of the storyteller's body is still being improved, the character's gestures and body

expression will be then more developed. By connecting the input from the user with the utterance of the voice, the emotional state, facial and body expression in an interactive emotional storytelling system, we hope to deliver a pleasing and didactic experience to the user.

5 Acknowledgements

Thanks to Fernando Rebelo for the design of Papous. Thanks to the GAIPS team for their comments and criticisms during the development of this system. Thanks to Phoebe Sengers for allowing us to use the Influencing Machine.

The work on Papous was funded under the Sapiens Program- Funda cão para a Ciência e Tecnologia, project number POSI/SRI/41071/2001.

References

[1] P. A., A. G., H. K., M. D., C. M., and M. C. Sentoy in fantasya: Designing an affective sympathetic interface to a computer game. *Personal and Ubiquitous Computing Journal (Ub Comp)*, (5-6), 2002.

[2] J. Cassell. Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. In J. C. et al., editor, *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 1999.

[3] J. Cassell, T. Bickmore, L. Campbell, H. Vilhjalmsson, and H. Yan. Conversation as a system framework: Designing embodied conversational agents. In e. a. J. Cassell, editor, *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 1999.

[4] P. Ekman. *Emotion in the Face*. New York, Cambridge University Press, 1982.

[5] S. et al. The enigmatics of affect. In *Proceedings of Designing Interactive Systems- DIS'2002*. ACM Press, 2002.

[6] T. Noma, L. Zhao, and N. Badler. Design of a virtual human presenter. *IEEE Computer Graphics and Applications*, 20(4):79–85, 2000.

[7] I. Pandzic and F. R. (editors). *Mpeg-4 Facial Animation - The standard, Implementation and Applications*. John Wiley & Sons, 2002.

[8] C. Pelachaud and I. Poggi. Subtleties of facial expressions in embodied agents. *Journal of Visualization and Computer Animation*, 2002.

[9] A. Silva, G. Raimundo, and A. Paiva. Tell me that bit again... bringing interactivity to a virtual storyteller. In *Virtual Storytelling: Using Virtual Reality Technologies for Storytelling*. Springer, 2004.

[10] A. Silva, M. Vala, and P. A. Papous: The virtual storyteller. In *Intelligent Virtual Agents*. Springer, 2001.

Abstract for Symposium on Speech, Language, and Gesture for Expressive Characters. AISB'04, University of Leeds.

Pragmatics of Body-Speech Coordination in Multi-Modal Expression

Satinder P. Gill and Masahito Kawamori

Body Moves are rhythmic coordinations in communication of at least two people. In performing them, we indicate the state of our connection and understanding, most importantly the degree of our contact and commitment within a communication situation (Gill, Kawamori, Katagiri, Shimojima, 2000). The coordinated action of gesture and speech suggests there is a dialectic in cognition that is essential for effective communication (Kita, 1993, Streeck, 1993, McNeill, 1994). Body Moves indicates the construction, establishment, of mutual ground within a space of action.

Studies of collaborative drawing of conceptual designs (Gill, 1997; Gill, et al. 2000; Gill and Borchers, 2003), reveals a dialogue of body movements. These 'moves' are not representative of the verbal discourse; they are not iconic. Such gestures, movements of the whole body (for example from one side of a table to another), and interactions with pens, paper, whiteboards, SMARTboards, and other materials, constitute part of an overall dialogue and move it forward. Body moves are investigated as a special case of information flow in dialogue, within engagement space. Body language is to be considered both as a form of expression and as communication dynamics. In this paper the empirical work from studies of conceptual design will be presented.

1. Metacommunication

The analysis draws on the concept of meta-communication (Scheflen 1965). Meta-communications convey information about the conversation, as opposed to being about the topic situation of the conversation itself (i.e. content). Katagiri et al., define metacommunication as consisting in cuing facts: "the conveyance of information at the meta-level can be triggered by a number of facts holding in a conversation, and these "cuing" facts convey a variety of information about the conversation situation."

The work described here began with applying the idea of a conversational move to 'body language', and specifically to body movements that occur in response to each other, whether this is related to a verbal utterance or independent of it. Such moves are distinguished from representational gestures of the verbal utterance that serve primarily to illustrate it. Where a conversation move is a verbal action which causes the conversation to move forward (Carletta et al, 1977), the Body Move is a bodily action which initiates or responds either to a bodily action or verbal utterance, or moves with another bodily action (i.e. parallel and coordinated).

In this paper, the work considers Body Moves as conveying information about the conversation situation. In Gill et al. (2000), Body Moves comes under the category of *cuing facts*, as extra linguistic factors and under *cued information,* as they are also about conversation organisation (Katagiri op. cit.). It is from this perspective that they were originally seen as constituting a kind of information flow.

2. Body Moves as Composite Dialogue Acts.

Satinder Gill is an affiliated Researcher with the Topological Media Lab, Georgia Tech. Masahito Kawamori currently heads the NTT Telecommunications and Broadcasting Group's R&D, Tokyo. The authors began this work at NTT's Information Science, and then Communication Science, Research Labs, NTT, Atsugi, Japan, as part of the Cognitive Science and Dialogue Understanding Groups.

From an excerpt of videotaped interaction in our landscape architecture study, seven Body Moves are identified: body-check *(b-check*), acknowledgement *(ack), take-turn*, *attempt-contact*, *dem-ref, focus*, and *parallel-coordinated* move. In developing the categories of Body Moves, we found it helpful to draw upon communicative act theories and those features of communicative acts which bore parallel to the phenomena observed, for example, Carletta et al.'s categories of conversational moves (1997), Traum's work on 'grounding' (1994), Katagiri, Shimojima, Koiso & Swert's work on 'Echoing in Japanese Conversation' (1997), and Allwood, (1995). However some features of communicative acts are specific to speech and are not embodied in Body Moves, such as intonation, and asking questions or making commands. Hence some Body Moves have required the development of new terms in order to either demarkate between them and their CA counterpart, such as 'check', which as a Body Moves becomes *b-check*, or because there appears to be no clear counterpart in CA theory, for instance *dem-ref, attempt-contact* and *focus*.

Further, the BM may be described as a Composite Dialogue Act (CDA) (see Engle, 1998, on 'composite signal') as its composite necessarily lies across more than one person. A CDA can take one of four forms: a BM is accompanied by a CA; a BM is accompanied by no speech; a CA is accompanied by no BM; there is only silence, e.g. as in a pause. In the cases where a BM has an accompanying CA, the nature of the CA is dependent upon the context of the BM. For the seven Body Moves, we presented a particular set of CA for BM, but this is not exhaustive of the set of possible CDAs. The associated CAs covered were *suggest, confirm,* information-reference *(info-ref)* and acknowledgement *(Ack)*. *Suggest* (Traum, 1994) is when the initiator proposes a new item as part of a plan; *confirm* is when either the initiator elicits confirmation and/or the responder confirms they understand; acknowledgment *(Ack)* (Carletta, 1997) shows that the speaker has heard the move to which it responds, and often demonstrates understanding and acceptance; information-reference *(info-ref)* is a newly constructed CA for the case of the CDA, and denotes the content of speech which the BM provides the evidence for.

3. Study (Landscape architecture)

Two landscape architects, one fully qualified and director of the company, the other being trained in the company, to fully qualify in another year's time. They are both familiar with each other and share a mutual respect, despite the difference in status and experience. Their task is to produce a plan for the car park and the site. Some time earlier, they had produced a sketch plan for the client of the site, which is to be transformed from being an old derilict brewery to a headquarters of the client. The client has in turn produced a version of this, largely following their ideas, and wants them to take this further. Part of the discussion between A (senior) and B (junior) is whether to go for something radical or generally remain within the bounds of what they have in front of them. They, or rather, A, decide that changing it would not greatly improve on what they have and would cost more.

There's a great deal of body interactions in this design activity. Their mutual respect means that B is able to express disagreements and produce his own suggestions. However, the discrepancy in status is evident in the take-turns and keep-turns that A performs.

Below are two examples of their Body Moves. The first one *(Ack)* contrasts in its definition with the conversation move (Acknowledge), and the second *(Focus)* is a new category of 'move'. Both Body Moves are composites of movement/gesture and speech actions of more than one person, hence they are Composite Dialogue Acts (Gill, et al, 2000).

{} body movements with speech
[] body movements as turns (i.e. no speech)
| indicates the point at which body actions start

Conversation Act: Acknowledge

'Acknowledge' is a verbal response that minimally shows that the speaker has heard the move to which it responds, and often also demonstrates that the move was understood and

accepted. Carletta et al. Refer to Clark and Schaefer's (1989) five kinds of evidence that an utterance has been accepted: 'continued attention', 'initiating a relevant utterance', 'verbally acknowledging the utterance', 'demonstrating an understanding of the utterance by paraphrasing it', and 'repeating part or all of the utterance verbatim'. Carletta et al. Count only the last three as acknowledge moves, stating that the first leaves no trace in a dialogue transcript to be coded, and the second involves more dialogue moves.

Example:
G : Ehm, if you . . . you're heading southwards.
F : Mmhmm.

Example :
G : Do you have a stone circle at the bottom?
F : No.
G : No, you don't.

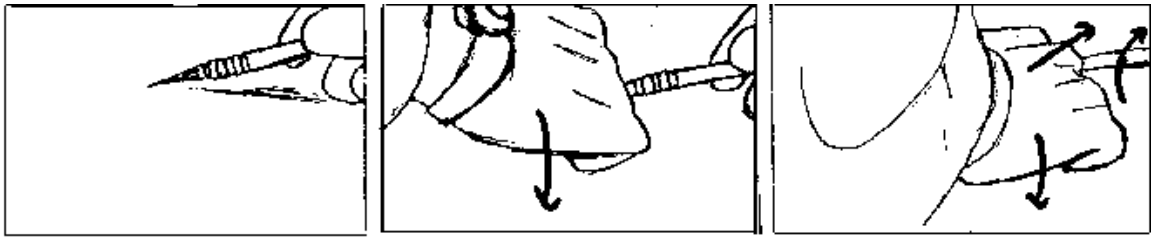
NB: the first of Clark and Schaefer's evidence for acceptance, 'continued attention', is seen to be useful for the case of Body Moves.

Body Move: Acknowledge (Ack)

The acknowledge move *(Ack)* gives an idea of the attitude of the response, i.e. how the person hears and understands and perceives, what is being discussed. It shows continued attention. In the CA 'acknowledge', Carletta et al.(1997) have not included this aspect of acknowledge, which was raised by Clark and Schaefer (1989), because it leaves no trace in a dialogue transcript to be coded. However, it is a part of the Body Moves. The hearer or listener demonstrates, with their gesture, how they are acknowledging the other's proposal or request for agreement. The BM occurs in response to the other's CA of information reference or suggestion, and their body release-turn or bodily place-holder. Its associated DA is the speech act 'ack' or 'accept'. In one of the examples shown below, the magnitude of the gesture and the physical proximity of the hands moving in close to the others', indicates the degree of engagement in the situation. The movement creates a change in the *degree* of contact, which indicates the nature of the aknowledgement or acceptance.

3.1. Example of Body Move: (Ack)

A is talking about the main office which people will be walking towards. The design, so far, has not considered how people are going to move from the various parts of the car park to reach there. Again, the conversation is about signing the route to guide people through the spaces. A is to the right of the picture and B is to the left.



(1) **(2)** **(3)**

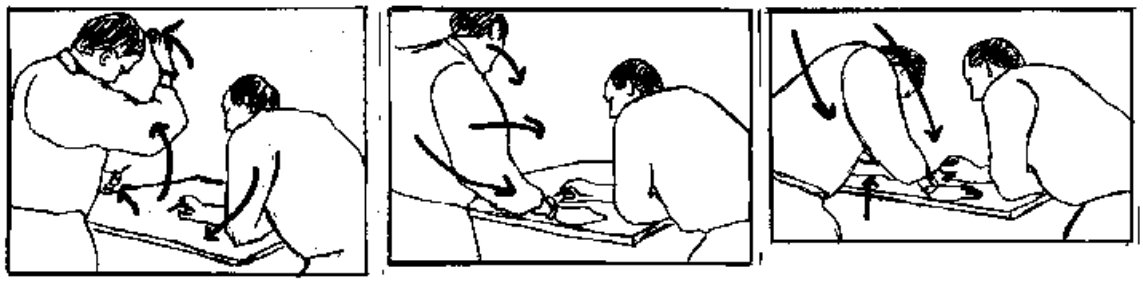
A	... Well it seems we've ignored the main	**A:** *CA: Suggest*(1)
	{A keeps pen pointed on the same spot}	**(1)**
	access of the brew house	
B	Yes	**B:** *CA:Ack*
A	Haven't we	**A:** *CA:Confirm*(1)
B	Totally. Yeh. Cos isn't it fronted by this	**B:** *CA:Accept*
	{B moves hands in very close to A's pointed pen}	**B: BM:Ack(2-3)**

B demonstrates his acknowledgement of the situation by moving both hands very close into the space that A has been holding attention upon. His movement parallels his emphatic 'Totally'. In moving in so close (emphasis) B demonstrates to A that he is engaging in the situation. The contact is very strong as B's hands come right up to A's who has to slightly move his pencil back but keeps his hand in the same position.

3.2. Example of Body Move: Focus

Focus is a signal for a change in the level of focus causing a meta-shift in the discussion. It involves a movement of the body towards the area the speaker is attending to and in response causes the listener or other party to move their body forward towards the same focus. The response move to it may not involve understanding but it does involve a willingness to perceive the message (Allwood et al. 1991)[1]. Maybe this BM mediates recognition on the part of the recipient because it causes them to give their attention to what the person is going to do next by making them move into that person's space of bodily attention, thereby creating increased *contact* for engagement in this shift. The associated CA to the BM is a suggest act.

In this example, A is talking about how people would leave their cars and walk over to their offices. This discussion is about how to move through the car park in the most effective way to best enter the site for your office. There needs to be signing to exit the car park and enter the site. A is to the right of the table and B is to the left.



(1) **(2)** **(3)**

A	if people getting if you get out of the car here how do	A: *CA:Suggest*
	{A keeps elbows on the table}	**A: BM:Focus(1-2)**
	you get to the office |	
	{B bends down}	**B: BM:Focus(2-3)**

[1] Allwood, Nivre and Ahlsen (1991) pay special attention to the type of reaction conveyed by feedback utterances, and context sensitivity of feedback. With regard to context sensitivity, they focus upon the type of speech act (mood), the factual polarity, and how the information status of the preceding utterance influences the interpretation of feedback utterances. The four basic communicative functions are:
1) ***Contact*** - whether the interlocutor is willing and able to continue the interaction
2) ***Perception*** - whether the interlocutor is willing and able to perceive the message
3) ***Understanding*** - whether the interlocutor is willing and able to understand the message
4) ***Attitudinal reactions*** - whether the interlocutor is willing and able to react and (adequately) respond to the message, specifically whether he/she accepts or rejects it.
The case for linguistic feedback is the need to elicit and give information about the basic communicative functions, i.e. continued contact, perception, understanding, and emotional/attitudinal reaction, in a sufficiently unobtrusive way to allow communication to serve as an instrument for pursuing various human activities. Feedback is therefore an essential instrument for successful communication and for the incrementality of communication, i.e. the step by step build up of consensual joint understanding which in turn is a means for pursuing a variety of other human activities. Aspects of Allwood et al's theory were adapted and modified for the body situation, such as 'contact'.

This sequence follows on from a prior take-turn where A had moved his body down onto the table and B had moved his body out of the drawing space **(1).** However, B subsequently moves down into the same space as A's body movement *Focus* engages B in his space of attention (2-3).

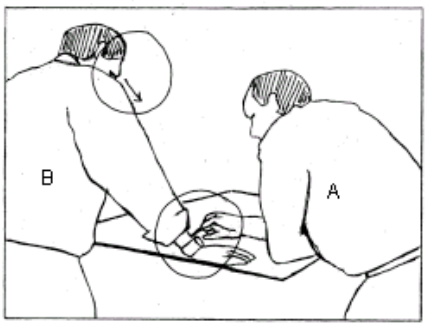
4. Parallel Coordinated Moves

In these examples, Body Moves were seen to have the quality of a cueing fact. However, this posed a challenge for handling rhythmic coordinations that did not fit the sequential structure of a cueing system. In Gill (2000, 2002) and Gill and Borchers (2003), a non-sequential rhythmic coordination, called the Parallel Coordinated Move (PCM) that was identified in Gill et al. (2000) was further analysed by drawing upon joint activity theory (Clark, 1996) and synchronous communication studies of body and speech coordination (Kendon, 1970).

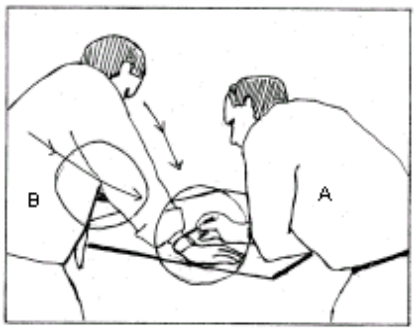
Example:

A wants to modify the client's sketch plan but without doing anything radically different because the client's primary concern is to provide a certain number of car parking spaces, and doing something radical might not help that and so be time consuming. As the discussion opens, A cites an example of a part of the drawing which could be a 'reference point' in a conceptual space of patterns. B says that this space is 'a bit naff'. A says that it is 'arbitrary' and seeks to make that space into something meaningful by making it part of an abstract design, where you have a flow of trees, paving patterns, and lines of blocks, 'patterns blasting through'. B wants to consider the space in question as a point that you walk through, i.e. something that's more functional (as opposed to abstract); an entrance point from the car park onto the site. He thinks the functionality will give it meaning. A, however, thinks that at the level of functionality ('looking at it from a very practical point of view'), a more central position near the main office of the Brewery site would be the appropriate entrance point from the car park ('we're missing the central access to the Brewery site). The Parallel Coordinated Move (PCM) occurs towards the end of this difference in focus level, after passing through three phases. Immediately prior to the PCM, A moves to B's level of 'functionality', and immediately following the PCM he acknowledges B's point by saying 'well you've got to leave some gaps obviously'. B acknowledges A's acknowledgment, and both move on with the dialogue, the matter of their conceptual difference having found some common ground.

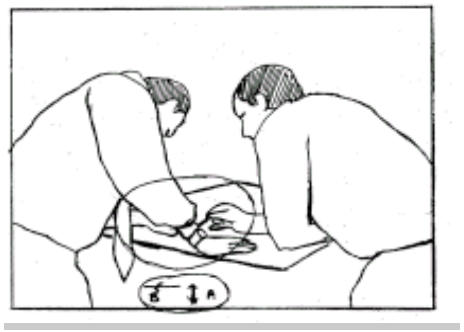
The two architects are working around a table and their movements occur in the space of its surface and perimeter. A is on the right of the table and B is on the left, in the example below. [i]



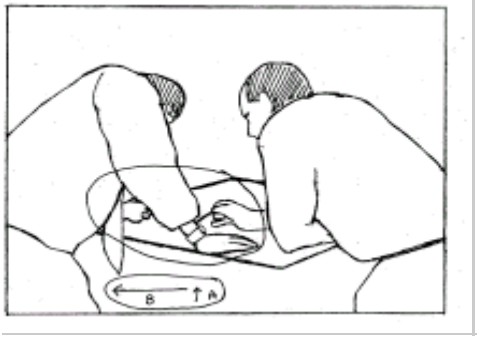
(1)



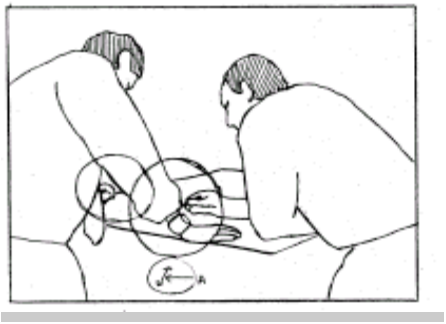
(2)



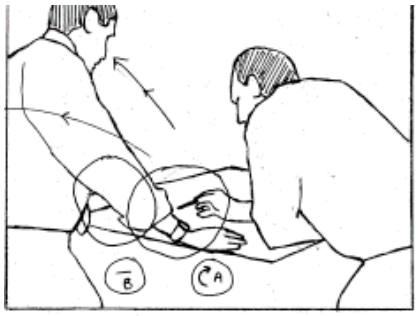
(3)



(4



(5)



(6)

1. A	Looking at it from a very practical point of view,	
2.	if you get out of the car here| how do you	A: *CA:Suggest* **(1)**
3.	{keeps elbows on the table}	A: BM:Focus **(1-2)**
4.	{A's hand moves into a pointing gesture}	**(1-2)**
5.	get to the office,	
6.	{A's pen touches down on a point}	
7.	{B moves to bend down}	B: BM:Focus **(2-3)**
8.	do you **wa | lk, go**	
9.	**{B runs pen along diagram}**	**B: BM:PCM (3-4)**
10.	**{A moves pen from side to side}**	**A: BM:PCM (3-4)**
11.	**through here |**	
12.	Pause	
13.	{B begins subtle move forwards, A's body is still}	
14.	uh |	
	{B's hand has moved forwards, A's body is still}	
15.	and then you've got to	
16.	{A moves hand across}	
17.	leave some gaps [obviously	A: CA: Accept
18.	{A makes 'cutting' action with pencil back and forth}	A: BM: Dem-Ref
19.	{B's body and head swing back}	B: BM: Ack
20. B	[yeah, yeah, yeah	B: CA: Ack
21. A	uh	
22. B	Is this the number of spaces	
23. A	So if you want get, If I want to go here	
24. B	yeh	

Prior to this moment (above) A had performed a bodily take-turn (see Gill at al. op. cit. pp105-106), which had interrupted B who shifted his body out of the space he was acting

within. Here, A moves his hand from one pointing gesture **(1)** ('if you get out of the car here'), into another pointing gesture **(1-2)**, where he brings the point of his pen over one position on the drawing upon uttering 'get' [L5], then touches down on the paper's surface at 'fice' in 'office' **(2)**(L5). This pointing position is a placement holder for the ensueing utterance, 'do you walk…" [L8]. It is not simply a reference to the utterance he is making at the time he moves, but refers to what he is going to say. In previous work, we have described A's body movement in this example as a *Focus* move (Ibid. pp.109-110), illustrated further above, which engages B in his space of bodily attention. At 'office', B's head has moved into the position from which he begins his descent into the conceptual drawing space**(2)**.

A's hand gesture **(1-2)** combined with his inviting question beginning 'how do you …[get to the office]', brings a response from B who moves down and places his pen, held in his left hand, in very close proximity to A's right hand/pen position **(3)**. This is followed by a drawing movement of a straight line to his left, away from A's position **(4)**. A begins drawing just after B does[2]. His gesture is deictic, that is, it marks out the location of his speech, and it is rhythmic or synchronised with his speech (during **(3)**). Yet B is not looking at him or that gesture.[3]

Upon detailed analysis (Gill, 2000, 2002), as B touches down (just after the utterance 'office' is finished), and begins his move, A begins moves his pen back and forth: 'go' - forth, 'through' – back, 'here' – forth [L8,11]. It could be that in making his alteration with his gesture he is clarifying his meaning to himself as well to B[4]. In addition, his motion can seen as both a self-gesture and as a holding action that maintains his position in their parallel body dialogue.

A and B's Parallel Coordinated Move is made up of B's body performing a silent movement that is partly autonomous of the body and speech of A. B closes by moving his hand out of the space of bodily attention of A and out of the space he himself was acting within. His closure is accompanied by A moving over into that space of action. Just as A moves in upon B's space, B's hand comes back down to onto the edge of the table, and their spaces of bodily attention have an intersection **(5)**. B's body is still, his hand is still holding his pen in a closed grasp, and A is marking a specific point along the line that B drew, with a backwards-forwards motion ['and then you've got to leave some gaps...][L15,17]. At 'gaps', B makes a quick movement of his head and torso up and back, away from the table, saying 'yeah, yeah' **(6)**. His left hand (the drawing hand) releases and is flat on the table. A's gesture has both physically and metaphorically acknowledged B's proposal, and is followed by a reciprocal acknowledgement by B of A's acknowledgement.

Prior attempts by B to share his idea with A involve a sequential pattern of action-response motions. Hence when B draws, A responds only when B has finished. A does not move inside or touch B's space of bodily attention, although he is attending to him. This is called a 'passive distance' that involves less commitment. The openness of the parallel move is not dominated by any one view. In the case where people differ, as in this example, it is proposed that simultaneous rhythmic synchrony enables them to negotiate their difference by experiencing it bodily.

4.1 Joint Action

---

[2] In an earlier version of this paper (Gill, 2000), the video analysis, having been made from a poor quality recording without sufficient video analysis tools, showed that A's body was still. However, the digitised version, viewed using video analysis tools, shows A's movements very clearly. The PCM was first identified as such in 1997 whilst the author was at ATR. The reason for using the expression 'parallel' is even more obvious than before. It denotes simultaneous and coordinated autonomous action.

[3] Kendon (1970) proposes that the listener self-regulates his or her movements to fit the speaker's speech stream. Gill and Kawamori (2002) also show that gestures are highly synchronised in non face-to-face communication where speech acts as a feedback link.

[4] Work on non face-to-face gestural coordination (Gill and Kawamori, 2002), indicates that certain synchronous gestures occur with own speech function, enabling the speaker to maintain his/her own communication situation.

The Parallel Coordinated Move is a coordinated multi-gestural activity that involves a high degree of commitment to be engaged For the case of the body, the concept of joint action has been adapted with concepts of 'contact' and 'focus', a consideration of history and meaning of the gesture, and with the idea of extending the other person's space of bodily attention by one's own body motion. The prior sequential Body Moves leading up to the PCM, involved evolution in the specific forms of the gesture.

because it is a replica[5] of the form of the trajectory of the three previous gestures made in the three prior proposal attempts. This may explain its autonomous quality of being 'parallel'. However, having a history of form does not mean that the function of this gesture is the same as in the previous interactions. Mc Neill (1992) has argued for the specific meaning of each gestural movement. Hence, even though the form of the trajectory of hand and pen movement is shared, the meaning of the gesture lies in its interaction, which involves differing postural orientations from those of previous interactions. The historical element may explain their ability to have a high degree of *contact* without the space of bodily attention being disturbed when in close proximity. It also suggests how A can *understand* B's act. There has been some 'grounding' (Clark, 1996) taking place to enable this. That grounding is achieved can be observed in the motion that A performs when he moves over to the space that B had been drawing in. This is a replica of the motion that B performed in the attempted proposal prior to the parallel coordinated move, of a cutting action of making a passage through a line of trees. A has gesturally and in speech, accepted B's suggestion which was bodily ignored or rejected on previous attempts. His acknowledgement is located in history and in the present moment.

5. Simultaneous Autonomy

In an experiment that compared the affordances of two technologies (a SMARTboard and a whiteboard) for collaborative conceptual sketching activities (Gill and Borchers, 2003, Gill, 2004), it was found that simultaneous action had a greater role to play in sustaining the communication. Our motivation behind the experiment was to re-create the situation of the landscape architects, and find more Parallel Coordinated Moves to analyse and better discern their function and pattern in relation to other Body Moves.

The study compared collaborative drawing activity at a SMARTboard with a whiteboard. The users are asked to design a shared dorm living space. In the example below, E is to right of the whiteboard, and F is to its left.

(1,2,3,4,5) tag reference to specific moment of body move in the figures **(1),(2),(3),(4),(5)**

(1) (2) (3) (4) (5)

FIGURE 3

1. E: Do you want us to do like a bed	E: CA:Suggest **(1)**
2. {E moves in to the whiteboard, index finger point touches the	**(1)**
3. surface; F at the surface about to draw}.	
4. E: and then that then	
5. {E's finger traces the outline of a bed	**E: PCM (1-2)**

[5] 'replica' does not denote 'imitation', but 'reproduction'.

6. {F is drawing in a line towards the left, his head nods}	**F: PCM (1-2)**
7. E: one here	E: CA:Suggest **(3)**
8. {E moves his hand down and taps the surface, of the space	E: BM:Dem-Ref **(3)**
10. that he has just traced, with the back of his hand};	
11. {F traces the outline of the bed in the air, moving his hand	F: BM:B-Check **(3-4)**
12. straight to the left and back and then down}.	
13. E: and then | one here	E: CA:Suggest **(4-5)**
14. {E lifts his hand up and taps the board again with the back	E: BM:Dem-Ref **(4-5)**
15. of his hand; at | F moves his hand back to original position	
16. Silence [E lifts hand off from the surface, as F is	
17. about to touch it with his pen]	
20. F: ye	F: CA:Ack
21. {F puts pen back on paper};	**(5)**
22. [E's body begins moving back]	
23. E: and maybe do like a dresser between them	
24. {E's body moves back to rest-reflection position; F is drawing the beds}	

F acknowledges E's proposal, in tracing the proposed idea above the surface of the board (Fig. 3, pictures (3) and (4)) with his pen, whilst E taps a position of one bed with the back of his hand on the surface to locate it. Through gesture *(Ack),* F checks the proposal that E is making through gesture (*Dem-Ref*) and speech (*Suggest*). After tracing, F continues to draw, and his pen touches the surface (5) at the same time as E begins to lift his hand away. There is no break in the fluidity of the rhythm of the coordination between them (of body and speech).

It is interesting to contrast such activity with the SMARTboard, and further understand its function in coordinating interaction. At the time this study was undertaken, the SMARTboard permitted only one source of input at a time. Participants' commitment, politeness, and attention to each other became reduced at a single SMARTboard, showing behaviours that are in marked contrast to those of users at a whiteboard (Gill, Martin, and Sethi, 2001). When a designer at the SmartBoard does not easily give the turn to the other one, we see various strategies to force it. These include, moving close to the board and inside the visual locus of the drawing space in a quick motion, or moving back and forth, or reaching for a pen, or looking at the pen, or simply reaching out and asking for the pen the other person is currently using, or just moving right in front of the body of the person currently drawing, thereby forcing them back, and taking a pen from the pen holder. As either person can act at the whiteboard, there is no need for such strategies.

Acting in parallel, for example, drawing on the surface at the same, involves simultaneous autonomy. Simultaneous motion and touch provides for a certain kind of awareness of and attendance to the states of contact within the engagement. Contact with each other's ideas can be made with gestures as well as speech.

There is a pattern of movement from sequential (e.g. turn taking) to parallel actions, as part of this design activity that suggests that *coordinated autonomous action is part of sustainable collaborative activity* (Gill et al, 2003; Gill, 2004).
In a further related study, when a group of four users has three such SMARTboards available to them, there appears to be a transposition of the patterns of autonomy and cooperation that one finds between a dyad working on a whiteboard (Borchers, Gill, and To, 2002, Gill and Borchers, 2003), although the body moves and parallel actions that constitute them take a very different form. The specific configuration or form of gestures and body movements do not define the function of the Composite Dialogue Acts they constitute. This is defined within the act's context i.e. the particular interaction situation.

6. Body Moves as Metacommunication

Recently, in Gill et al. (2003), a visit back to Scheflen (1974) and Bateson's (1955) formative work has lead to a further understanding of how the various forms of Body Moves (both sequential and parallel) are meta-communicative: the 'movement of the body helps in clarifying meaning by supplementing features of the structure of language', (Scheflen op cit. p.11). Further, Body Moves contributes to the idea that the structure of language lies in its performance.

Body Moves (BM) are meta-communicative spontaneous bodily interactions which communicate information about the communication situation tacitly. Bateson's idea of metacommunication came from observing dogs at play, and suggested there was a signal by which they communicated to each other. The simplest metasignals were seen to specify a type of meaning or qualify the literal nature of the act. Smiling, for example, may metacommunicate that a certain aggressive act is to be considered as friendly and competitive rather than as attacking in behaviour. 'One can also metabehave to someone else's behaviour and thereby alter or influence it' (Scheflen, 1973, p.101).

In using the term 'Body Move' we refer to the 'act' performed, rather than to the specific choregraphy of the body. BMs are rather like moves in a game, however, unlike the strategic nature of moves in a game we cannot say that BMs embody specific intentions. But like the case of play, we could say that they embody an intention for communication itself. In considering the composite nature of the move, we cover the relation between body motion, speech motion, and communicative act, both within the individual and across engaged individuals at the same time. The identification of any Body Move has arisen from the patterns of feedback and contact embodied in salient rhythms.

7. Conclusion

These studies indicate that Body Moves exist. Further work needs to be undertaken to determine the universal nature of the moves that have been identified. We would need to study different contexts for the discovery of other moves, and discern how they may be bound by a specific interaction context (in this case, conceptual drawing tasks), materials/interfaces (e.g. pens and paper, SMARTboards, whiteboards, co-located or distributed), and social factors (e.g. gender, status etc.) and culture.

References

Allwood, J., Nivre, J., & Ahlsen, E. (1991). On the Semantics and Pragmatics of Linguistic Feedback. Gothenburg Papers. *Theoretical Linguistics,* 64.

Bateson, G. (1955). *"The Message. 'This is the Play.'"* In Schaffner, B. (ed.) Group Processes. Vol.II. New York: Macy.

Birdwhistle, R.L. (1970*). Kinesics and Context*. University of Pennsylvania.

Carletta, J., Isard, A., Isard S., Doherty-Sneddon, G, & Anderson, A. 1997. *The Reliability of a Dialogue Structure Coding System.* Paper, Association for Computational Linguistics.

Clark, H.H. and Schaefer, E.F (1989). Contributing to discourse. *Cognitive Science,* 13, 259-294.

Clark, H.H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Engel, R. (1998). Not Channels but composite signals: speech, gesture, diagrams and object demonstrations are integrated in multi-modal explanations. In M.A. Gernsbacher & S.J. Derry (Eds.) *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Gill, S.P. 1997. *Body Language: The uspoken dialogue of Bodies in Rhythm. Unpublished report, ATR.*

Gill, S.P., Kawamori, M., Katagiri, Y., Shimojima, A. (2000). The Role of Body Moves in Dialogue. In *RASK*, Vol.12, pp.89-114.

Gill, S.P. (2002). *The Parallel Coordinated Move: Case of a Conceptual Drawing Task*. Published Working Paper: CKIR (Centre for Knowledge and Innovation Research), Helsinki.

Gill, S.P. and Borchers, (2003). Knowledge in Co-Action: Social Intellige nce in Collaborative Design Activity. *AI & Society, 17.3*.

Gill, S.P. (2004) (in print). Body Moves and Tacit Knowing. To appear in Gorayska, B. and Mey, J.L. (Eds.) *Cognition and Technology: Co-existence, Convergence, Co-evolution*. John Benjamin. Publication due in August 2004.

Katagiri, Y., Koiso, H., Shimojima, A. 1997. *Scorekeeping for Conversation-Construction*. Proceedings of the Munich Workshop on Semantics anda Pragmatics of Dialogue. The University of Munich, Germany, 1997.

Kawamori, M., Kawabata, T., Shimazu, A. (1998). Discourse Markers in Spontaneous Dialogue: A corpus based study of Japanese and English. In *Proceedings of 17th International Conference on Computational Linguistics* (COLING-ACL98).

Kendon, A. (1970). Movement Coordination in Social Interaction: Some examples described. In *Acta Psychologia*, 32:100-125.

Kita, S. 1993. *Language and thought interface: A study of spontaneous gestures and Japanese mimetics.* Ph.D. thesis, University of Chicago, Chicago, Illinois.

Scheflen, A.E. (1973). *Communicational Structure: Analysis of a Psychotherapy Transaction.* Indiana University Press.

Scheflen, A.E. (1974). *How Behaviour Means. Exploring the contexts of speech and meaning: Kinesics, posture, interaction, setting, and culture.* New York: Anchor Press/Doubleday.

Traum, D.T. 1994. *A Computational Theory of Grounding in Natural Language Conversation*, Ph.D thesis, The University of Rochester, Rochester, NY.

[i] Transcription Coding Scheme

The following conventions are used to encode Body Moves (BM) and Communicative Acts (CA) in this example. For a fuller coding scheme for BMs, see (Gill et al. 2000).

{}	body movements with speech
|	indicates the point at which the body action starts
// //	comment
_	denoted speech aligned to body movements
(1,2,3)	tag reference to specific moment of BM in the figures (1), (2), (3)