Workshop Programme

"Multimodal Corpora
From Multimodal Behaviour Theories to Usable Models"

Saturday 27 May 2006 http://wwwhomes.uni-bielefeld.de/mmc06/

9:00 Welcome & Introduction
J.-C. Martin, P. Küehnlein, P. Paggio, R. Stiefelhagen, F. Pianesi

20 mn per presentation + 10 mn of questions

9:15 MEETING & METADATA

9h15 - 9h45
Working with Very Sparse Data to Detect Speaker and Listener Participation in a Meetings Corpus
Nick Campbell & Noriko Suzuki

9h45 - 10h15
Multimodal Annotated Corpora of Consensus Decision Making Meetings
Fabio Pianesi, Chiara Leonardi, Massimo Zancanaro

10h15 - 10h45
Language Resource Archiving supporting Multimodality Research
Peter Wittenburg, Daan Broeder, Peter Berck, Han Sloetjes, Alex Klassmann

10:45 Discussion
11:00 Coffee Break

11:30 HAND GESTURES

11h30 - 12h
Analysis of gesture expressivity modulations
Nicolas Ech Chafai, Catherine Pelachaud, Danielle Pelé

12h - 12h30
Synthesizing Gesture Expressivity Based on Real Sequences
George Caridakis, Amaryllis Raouzaiou, Kostas Karpouzis, Stefanos Kollias

12h30 - 13h
An Annotation Scheme for Conversational Gestures: How to economically capture timing and form
Michael Kipp, Michael Neff, Irene Albrecht

13h - 13h30
Using FORM Gesture Data to Predict Phase Labels
Craig Martell, Joshua Kroll

13:30 Lunch

14:45 MULTIMODALITY DURING CONVERSATION

14h45 - 15h15
Degrees of freedom of facial movements in face-to-face conversational speech
Gérard Bailly, Frédéric Elisei, Pierre Badin & Christophe Savariaux

15h15 - 15h45
A coding scheme for the annotation of feedback, turn management and sequencing phenomena
Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta and Patrizia Paggio

15h45 - 16h15
A framework for analyzing embodied communicative feedback in multimodal corpora
Jens Allwood, Stefan Kopp, Karl Grammer, Elisabeth Ahlsén

16:15 Discussion
16:30 Coffee break

17:00 MULTIMODAL HCI

17h - 17h30
A Study into Multimodal Behaviour in Error Correction
Marie-Luce Bourguet

17h30 - 18h
Making a Case for Spatial Prompting in Human-Robot Communication
Anders Green, Helge Hüttenrauch

18h - 18h30
The MIMUS Corpus
Pilar Manchón Portillo, Carmen del Solar Valdés, Gabriel de Amores Carredano, Guillermo Pérez García

18:30 Discussion
19:00 End of Workshop

Workshop Organiser(s)

Jean-Claude MARTIN, CNRS-LIMSI / Univ. Paris 8-LINC

Peter KÜEHNLEIN, Bielefeld University

Patrizia PAGGIO, Centre for Language Technology (CST), University of Copenhagen

Rainer STIEFELHAGEN, Universitaet Karlsruhe (TH), Interactive Systems Labs, ITI

Fabio PIANESI, Istituto Trentino di Cultura - Centro per la Ricerca Scientifica e Tecnologica (ITC-irst)

Workshop Programme Committee

Table of Contents

Author Index

Introduction

'Multimodal Corpora' target the recording and annotation of several communication modalities such as speech, hand gesture, facial expression, body posture, etc. Theoretical issues are also addressed, given their importance to the design of multimodal corpora.

This workshop follows similar events held at LREC'2000, LREC'2002 and LREC'2004. There is an increasing interest in multimodal communication and multimodal corpora as visible by recently launched European Networks of Excellence and integrated projects such as HUMAINE, SIMILAR, CHIL and AMI, and similar efforts in the USA and in Asia. Furthermore, the success of recent conferences dedicated to multimodal communication (ICMI'2005, IVA'2005, Interacting Bodies'2005, Nordic Symposium on Multimodal Communication 2005) also testifies the growing interest in this area, and the general need for data on multimodal behaviours.

The focus of this LREC'2006 workshop on multimodal corpora is on non-verbal communication studies and their contribution to the definition of collection protocols, coding schemes, inter-coder agreement measures and reliable models of multimodal behaviour that can be built from corpora and compared to results that can be found in the literature.

Topics to be addressed in the workshop include, but are not limited to:

- Studies of multimodal behaviour

- Multimodal interaction in groups and meetings

- Building models of behaviour from multiple sources of knowledge : manual annotation, image processing, motion capture, literature studies

- Coding schemes for the annotation of multimodal video corpora

- Validation of multimodal annotations

- Exploitation of multimodal corpora in different types of applications (information extraction, information retrieval, meeting transcription, multi-modal interfaces, translation, summarisation, www services, communication and clinical studies, HCI design)

- Methods, tools, and best practices for the acquisition, creation, management, access, distribution, and use of multimedia and multimodal corpora

- Metadata descriptions of multimodal corpora

- Benchmarking of systems and products; use of multimodal corpora for the evaluation of real systems

- Automated multimodal fusion and/or generation (e.g., coordinated speech, gaze, gesture, facial expressions)

We expect the output of this one day workshop to be 1) a deeper understanding of the theoretical issues and research questions related to verbal and non-verbal communication that multimodal corpora should address, 2) how such corpora should be built in order to provide useful and usable answers to research questions, 3) an updated view of state-of-the-art research on multimodal corpora.

17 papers were submitted, of which 13 were accepted as oral presentations and covering several areas such as meeting analysis, metadata, hand gestures, multimodality during conversation and multimodal Human-Computer Interaction.

Looking forward to an exciting workshop !

Jean-Claude MARTIN

Peter KÜEHNLEIN

Patrizia PAGGIO

Rainer STIEFELHAGEN

Fabio PIANESI

Working with Very Sparse Data to Detect Speaker and Listener Participation in a Meetings Corpus

Nick Campbell & Noriko Suzuki
Department of Cognitive Media Informatics,
ATR Media Information Science Labs, Keihanna Science City, Kyoto, 619-0288, Japan,
fnick,norikog@atr.jp

ABSTRACT
At ATR, we are collecting and analysing `meetings' data using a table-top sensor device consisting of a small 360-degree camera surrounded by an array of high-quality directional microphones. This equipment provides a stream of information about the audio and visual events of the meeting which is then processed to form a representation of the verbal and non-verbal interpersonal activity, or discourse flow, during the meeting. In this paper we show that simple primitives can provide a rich source of information.

INTRODUCTION
Several laboratories around the world are now collecting and analysing "meetings data" in an effort to automate some of the transcription, search, and information-retrieval processes that are currently very time-consuming, and to produce a technology capable of tracking a meeting in realtime and recording and annotating its main events. One key area of this research is devoted to identifying and tracking the active participants in a meeting in order to maximise efficiency in data collection by processing inactive or nonparticipating members differently. [1, 2, 3, 4, 5, 6, 7, 8]. At ATR we are now completing the second year of a threeyear SCOPE funded project to collect and analyse such data. This paper reports an analysis of material collected from one such meeting in terms of speaker overlaps and conflicting speech turns. Our goal is to determine whether it is necessary to track multiple participants, or whether processing can be constrained by identifying the dominant member(s) alone. The results show that in a clear majority of the cases, only one speaker is active at any time, and that the number of overlapping turns, when two or more participants are actively engaged in speaking at the same time, amount to less than 15% of the meeting. This encourages us to pursue future research by focussing our resources on identifying the single main speaker at any given time, rather than attempting to monitor all of the speech activity throughout the meeting. The second part of the paper shows that a change in speaker might be predicted from the amount and types of body movement. These movements are speaker-specific and not uniform, but systematically increase in the time immediately before onset of speech. By observing the bodily movements of the participants, we can form an estimate of who is going to speak next, and prepare to focus our attention (i.e., the recording devices) accordingly.

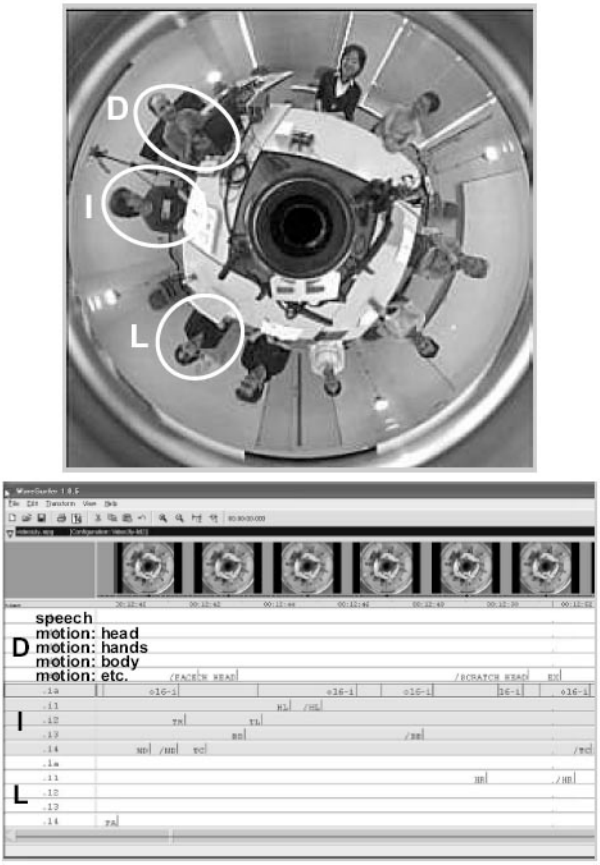


Figure 1. The camera's-eye view of a meeting (top), showing the annotated movement data for three participants (D,I,L) using the wavesurfer video plugin (bottom)

CATEGORIES OF SPEECH ACTIVITY
We have regularly been recording our monthly project meetings, where research results and project planning are discussed, to provide a database of natural (non-acted/no role-playing) speech and interaction information. The number of members attending each monthly project meeting can vary between four and twelve. Participation is voluntary, but since the research is being carried out by three teams at different locations (ATR, NAIST, and

Kobe University) the meetings provide an essential focus-point for coordinating the research activities. All meetings are recorded on both video and audio, using purpose-built equipment that has been described elsewhere [9, 10, 11]. All visible body-movements of the participants (head, hands, and torso) are annotated from observation of the video recordings, topic changes are noted, and the categories of speech activity are tagged by human labellers working interactively with the data.

id	topic	seconds
t-o2	progress-update(s1)	45
t-o9	progress-update(s2)	205
t-o15	progress-update(s3)	64
t-o23	progress-update(s8)	76
t-o12	self-introduction(s5)	191
sub-total		(738)
t-o6	data-tagging results	15
t-o8	data-preparation	157
t-o10	tanktops-and-skin-tones	142
t-o14	equipment-settings	82
t-o16	NAIST responsibilities	119
t-o18	reporting procedures	160
t-o20	Kobe Uni. responsibilities	58
t-o24	kinematics	148
t-o29	chameleon-eye-lens	564
t-o22	translation	11
t-o27	choice-of-camera	7
sub-total		(1306)
total		2044

Table 1. Topics that arose during the July meeting, with durations, showing the division between researcher-centred and technology-centred discussions

The speech is not yet being transcribed verbatim, but tags are assigned per topic and per activity type. We consider it necessary to distinguish (i) "on-topic" speech from (ii) "personal" speech, and also (iii) "backchannel utterances" and (iv) "laughter". We had also proposed (v) "yes" and (vi) "no" as relevant categories, but our experience with annotating these further two types of speech event suggests that they will not be easily recognisable using automatic processing, and we currently limit our tagging of speech activity to types i-iv above.

OVERLAPPING SPEECH

This paper reports the results of an analysis of one such meeting. Eight members were present at the meeting, which was held at NAIST in July 2005. They included the research director (s1), two team leaders (s3,s8), two researchers (s2,s4) two administrative assistants (s6,s7) and a guest researcher visiting from Ireland (s5). An observer was also present to monitor the recordings. The statistics of speech activity reported below clearly reflect the different roles of the participants, and the importance (in terms of time devoted to each) of the various topics.

Topics of discussion (see Table 1) included (a) progressupdates (approx. 36%) where one speaker tended to dominate, with the others listening and asking occasional questions, and (b) technical topics (approx. 64%), where more members became involved in the discussions. There were 2513 different "speech events" in the meeting, which lasted approximately 45 minutes altogether. Here, a speech-event is defined as a block of continuous speech, bounded by a cessation of speech activity, from one speaker, as indicated by `+' = start and `-' = end markers in the columns of figure 2. A brief silence after a burst of speech is marked by the `-' label.

```
0.72    .|.|.|.|.|.|.|.|-p      938.49
0.78    .|+o.|.|.|.|.|.|.|      939.27
0.23    .|=o.|.|.|.|.|.|.|+p    939.51
0.47    .|-o.|.|.|.|.|.|.|=p    939.99
1.4     .|.|.|+p.|.|.|.|-p      941.39
0.75    .|.|.|-p.|.|.|.|.|      942.15
0.04    .|.|.|+p.|.|.|.|.|      942.2
0.82    .|.|.|-p.|.|.|.|.|      943.02
0.55    .|.|.|+p.|.|.|.|.|      943.58
0.58    .|.|.|-p.|.|.|.|.|      944.16
1.04    .|.|.|.|.|.|.|.|+b      945.21
0.15    .|.|.|.|.|.|.|.|-b      945.37
0.18    +o.|.|.|.|.|.|.|.|      945.55
0.92    =o.|.|.|.|.|.|.|.|+p    946.48
1.32    -o.|.|.|.|.|.|.|.|=p    947.8
0.01    +o.|.|.|.|.|.|.|.|-p    947.82
0.39    -o.|.|.|.|.|.|.|.|+p    948.21
0.21    .|.|.|.|.|.|.|+b.|=p    948.42
1.25    .|.|.|.|.|.|.|-b.|-p    949.68
0.03    +o.|.|.|.|.|.|.|.|+p    949.72
2.3     -o.|.|.|.|.|.|.|.|=p    952.03
0.95    .|.|.|.|.|.|.|+o.|-p    952.98
0.29    .|.|.|+p.|.|.|-o.|.|    953.27
0.89    .|.|.|-p.|.|.|.|.|      954.17
0.65    +o.|.|.|.|.|.|.|.|      954.82
0.41    =o.|.|+p.|.|.|.|.|      955.24
1       -o.|.|=p.|.|.|.|+p      956.24
1.2     +o.|.|-p.|.|.|.|=p      957.45
0.03    =o.|.|.|.|.|.|.|.|-p    957.49
1.75    -o.|.|.|.|.|.|.|.|      959.24
0.07    +o.|.|.|.|.|.|.|.|      959.32
0.08    -o.|.|.|.|.|.|.|.|      959.4
0.41    +o.|.|.|.|.|.|.|.|      959.82
```

Figure 2. A sample of the audio labelling, showing three categories of speech sctivity: o=opinion, or public speech, p=private or personal speech, and b=backchannel utterances. A `+' indicates onset of speaking, `=' continuation, and `-' cessation of speech. The time in seconds of each event is shown on the left, and absolute time on the right

s1	s2	s3	s4	s5	s6	s7	s8
759	587	106	127	522	64	75	138

Table 2. Counts of speech events per participant

The distribution of events per speaker is shown in Table 2. Tables 3 and 4 detail the types of speech activity and times spent on each per speaker. Mean event duration is 0.7 seconds (sd=0.78), with the longest recorded event being 17 seconds. The 25th quantile of event durations is at a quarter of a second, and the 75th at 1 second. There were in addition 1730 points throughout the meeting during which no-one spoke. Both total utterance counts and overall speaking times indicate that s1 (the project leader), and s2 (a guest researcher expert in graphics processing) dominated the meeting. It is also evident from tables 2 & 3 that s5, the observer, also took an active part in the discussion. The administrative assistants spoke least at this research-based meeting.

	s1	s2	s3	s4	s5	s6	s7	s8
o	344	32	49	47	212	27	22	44
p	5	7	-	4	14	5	30	2
b	64	11	2	10	17	3	2	1

Table 3. Utterance timings for each participant for three categories of activity: O; on-topic talk, P: private talk, B: backchannel utterances. All timings are rounded to whole seconds.

on-topic	backchannel	private	laugh
2110	207	196	406

Table 4. Number of events for each speaking type on-topic backchannel private laugh

The count of participants actively speaking during each turn is given in Table 5. It shows that by far the majority of turns are single-speaker events. It is 6.5 times more likely that any given utterance will be single-speaker, and only 15% likely that more than one speaker will be active. There is only a 7% chance of more than 2 people speaking at any time in this meeting of 8 researchers. These figures may of course be culture-specific, and even meeting-specific. It might be supposed that backchannels contribute to the majority of overlapping utterances, but a count of singlespeaker backchannel utterances (n=134) versus a count of multi-speaker, overlapping backchannel utterances (n=74) shows this not to be the case. If we exclude from this s1's overlapping backchannels to s2 (n=19) then the ratio becomes 55:134, and it is 2.5 times more likely that a bakchannel utterance will be spoken without overlap.

silent	solo	two	three	four
1730	2000	291	15	1

Table 5. Number of participants active at each turn silent solo two three four

SPEECH & MOVEMENT

It has often been observed (e.g., [12, 13, 14]) that people move more when they speak. To determine whether these two types of activity had any useful correlation, we also examined the physical activity of all participants that was visible to the camera. We looked both at activity prior to speaking, and at activity while speaking. Since all were seated around a table, this study is limited to upper-body movement. Figure 1 shows the multi-tiered annotation that we use for labelling body movements which are apparent to a human observer when viewing the 360-degree camera output. In addition to a speech-related tier, separate tiers are available for “head”, “hands”, “body”, and “other”, where the last can be used for complex gestures such as “play with pencil”, “scratch head”, “fix glasses”, “stroke beard”, etc. For this paper, we simply counted the number of active labels at each moment of time and categorised them as follows: “Motion 1: only one body part is moving (e.g., the head or a hand), “Motion 2”: two body parts moving (e.g., head and hand, or two hands), “Motion 3”: three body parts moving (e.g., head and hand and body), and “Motion 4”: four or more body parts moving. The data from three speakers (those circled in the figure) were then compared for the periods immediately prior to onset of speech. Figure 3 clearly shows a rise in the amount of activity as the person prepares to speak. However, we can see individual differences, and it appears that two speakers reach a peak of activity shortly before speaking, while the third continues to increase up to the onset of speech. We can also note differences in parts of the body moved: Participant I, for example (the centre portion of figure 3), appears to move his head much more than the others (as indicated by the white portion of the bars). Figure 4 provides a breakdown of the types of activity per participant. It shows that for all speakers, the occurrance of speech having no overlap with body movement accounts for less than 20% of the total speaking time. It also shows that speakers behave differently; with all speakers moving 2 or more body parts at least 50% of the time, but one speaker (the centre column) moving 3 or more body parts more than 50% of the time while speaking.

DISCUSSION

The above analysis of the audio data shows that in a clear majority of the cases, only one speaker is active in any given turn. This implies that we will only lose a small amount of relevant information if we limit our processing to the single most dominant member at any one time. This will considerably reduce the work-load of the processing.

Furthermore, from an examination of the video data, we confirmed that people do tend to move more when they speak, and found that there is a steady rise in the amount of movement of all participants particularly in the 10 to 15 seconds preceding the onset of speech.

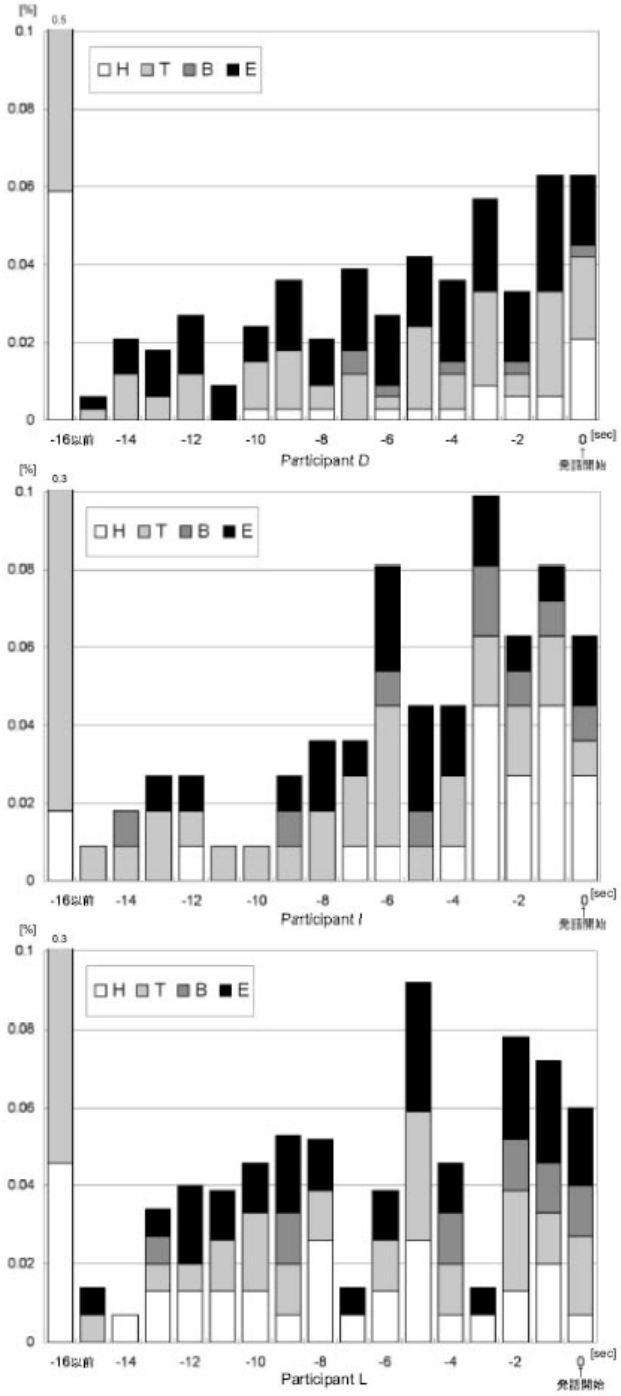


Figure 3. Rising amounts of bodily movement for 3 participants across a period of 16 seconds prior to onset of speech. Two speakers show a peak of activity a few seconds before speaking. Here "H" represents head movement, "T represents hand movement, "B" represents body movement, and "E" represents particular gestures (see text for details). The rightmost column shows onset of speech, and the leftmost sums all movements since last speech event.

From the two above findings, we conclude that it is feasible to design technology, based on the very simple presence or absence of speech noise and movement in the video signal, that will be able to detect and track the speakers in such a meeting situation. However, it will require development of separate technology to be able to determine the reactions of the other participants to any particular utterance or topic. This remains as future work.

ACKNOWLEDGEMENTS

This work is supported as part of the Strategic Information and Communications R&D Promotion Programme (SCOPE) by the Ministry of Internal Affairs and Communications, Japan and is being carried out as collaborative research between members of ATR, Kobe University, and the Nara Institute of Science & Technology.

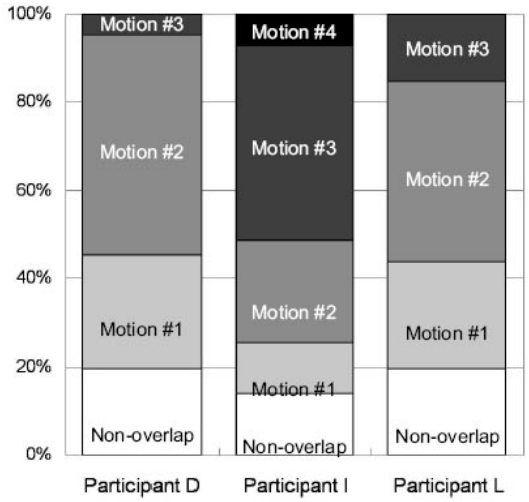


Figure 4. Number of major body parts that move while speaking. Non-overlap indicates that the speaker spoke while remaining relatively still. Participant I (centre) differs in moving more than the other two.

REFERENCES

[1] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: The impact of meeting type on speech style", in Proc. International Conference on Spoken Language Processing (ICSLP), Denver, Sept. 2002.

[2] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings", IEEE Trans. On Pattern Analysis and Machine Intelligence, vol. 27, no. 3, pp. 305317, Mar. 2005.

[3] W. N. Campbell, "A multimedia database of meetings and informal interactions for tracking participant involvement and discourse flow, in Proc LREC 2006, Lisbon.

[4] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus", in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Hong-Kong, Apr. 2003.

[5] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus", in Proc. HLT-NAACL SIGDIAL Workshop, Boston, Apr. 2004.

[6] V. Stanford, J. Garofolo, and M. Michel, "The NIST smart space and meeting room projects: Signals, acquisition, annotation, and metrics", in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Hong Kong, 2003.

[7] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement", in Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI), Edinburgh, Jul. 2005.

[8] M. Katoh, K. Yamamoto, J. Ogata, T. Yoshimura, F. Asano, H. Asoh, N. Kitawaki, "State Estimation of Meetings by Information Fusion using Bayesian Network", Proc Eurospeech, pp. 113-116, Lisbon, 2005.

[9] W. N. Campbell, "Non-Verbal Speech Processing for a Communicative Agent", Proc Eurospeech, pp. 769-772, Lisbon, 2005.

[10] W. N. Campbell, "A Multi-media Database for Meetings Research", pp 77-82 in Proc Oriental COCOSDA, 2006, Jakarta, Indonesia.

[11] Project Homepage: http://feast.atr/non-verbal

[12] Zhang, D.,et al., "Multimodal group action clustering in meetings"., VSSN'04, 54-62, 2004.

[13] Katoh, M., et al., "State estimation of meetings by information fusion using bayesian network", INTERSPEECH2005, 113-116, 2005.

[14] W. S. Condon, "Communication: Rhythm and Structure. Rhythm in Psychological, Linguistic and Musical Processes", J. R. Evans and M. Clynes. Springfield, Illinois, Charles C Thomas Publisher: 55-78. 1986.

Multimodal Annotated Corpora of Consensus Decision Making Meetings

Fabio Pianesi, Massimo Zancanaro and Chiara Leonardi
ITC-Irst
Via Sommarive - 38050 Povo-Trento, Italy
{pianesi, zancana, cleonardi }@itc.it
tel. +39-0461314570

ABSTRACT
In this paper we present an annotated audio-video corpus of multi-party meetings. The multimodal corpus provides for each subject involved in the experimental sessions 6 annotation dimensions referring to group dynamics; speech activity and body activity. The corpus is based on the audio and video recorded eleven sessions which took place in a lab setting appropriately equipped with cameras and microphones. Our main concern in collecting this multimodal corpus was to explore the possibility of providing feed-back services to facilitate group processes and to enhance self awareness among small groups engaged in meetings.

Author Keywords
Meetings, multimodal corpus, annotation.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

In this paper we present an annotated audio-video corpus of meeting activities. The present annotations focus on information that can be used to monitor and understand group dynamics and personal behaviors. The multimodal corpus was developed with two goals in mind. First, the collection of the data has to support the design and development of systems and tools capable of monitoring and tracking individual and group behavior, to foster better interaction styles, group satisfaction and productivity. Second, the multimodal data collected are meant to enhance the corpus of empirical data in human interaction as well to improve our understanding in the multiple social, psychological and emotional aspects involved in multiparty meetings. The former goal has been discussed in [7, 11] where we a study of the users acceptance of a simulation of an automatic relational report aiming at enhancing reflexive thinking and self-awareness among small groups. This study demonstrated an high degree of acceptance of the aspects concerning automatic multimodal coaching.

Several multimodal corpora have been already developed to analyse meetings. In particular the MM4 corpus [10] and the VACE corpus [5] are closed the one proposed here since they annotated low-level cues, such as speech, gesture, posture, and gaze in order to interpret high level meeting events. Brdiczka and colleagues [4] proposes a fusion algorithm to detect subgroup activities in a meeting. Our research aims at a step further, namely the automatic annotation of group dynamics.

METHODS AND PROCEDURES FOR DATA COLLECTION

The multimodal annotated corpus is based on the audio and video recorded during eleven meetings, which took place in a lab setting appropriately equipped with cameras and microphones (see below).

In order to provide for as much uniform context as possible, our groups were engaged in the solution of one of two versions of the Survival Task.

Interaction context – the Survival Task

The Survival task is frequently used in experimental and social psychology to elicit decision-making processes in small groups. Originally designed by National Aeronautics and Space Administration (NASA) to train astronauts before the first Moon landing– the Survival Task proved to be a good indicator of group decision making processes [8]. The exercise consists in promoting group discussion by asking participants to reach a consensus on how to survive in a disaster scenario, like moon landing or a plane crashing in Canada. The group has to rank a number (usually 15) of items according to their importance for crew members to survive

Consensus decision making scenario was chosen for the purpose of meeting dynamics analysis mainly because of the intensive engagement requested to groups in order to reach a mutual agreement, thus offering the possibility to observe a large set of social dynamics and attitudes. In consensus decision making processes, each participant is asked to express her/his opinion and the group is encouraged to discuss each individual proposal through weighing and evaluation of decision quality.

In our setting, we retained the basic structure of the Survival Task. In particular, a) the task was competitive across groups/team, with a price being awarded to the group providing the best survival kit. b) The task was collaborative and based on consensus within the group, meaning that a participant's proposal became part of the

common sorted list only if he/she managed to convince the other of the validity of his/her proposal.

Experimental protocol

Before starting each recording sessions, participants were given general information about the task filled a consent form. Each participant was equipped with close-talk microphones and asked to sit around a round-shaped table without restrictions concerning their positions and movements around the table (see Figure 1).

Figure 1. Experimental setting

A document was given to the group containing the items that were the objects of the discussion, and the instructions concerning the task. The experimenter sat in the room away from the table, without participating to the discussion, and collecting information and observations on an experimental sheet.

All the participants (40% males and 60% females) involved in the study were clerical people working at ITC-irst. In all cases they knew each other, and had often been involved in common group activities in the past. The average age was 35 years.

Setting and recording procedure

Each session was recorded in the specially-equipped CHIL room at ITC-irst (see Figure 2), by means of five Firewire cameras (AVT MARLIN), four placed on the four corners of the room while one lied the table. Four Web (SONY SNC-RZ30P) were installed on the walls surrounding the table.

Speech activity was recorded using four closed-talk microphones, six tabletop microphones and seven T-shaped microphone arrays, each consisting of four omni directional microphones installed on the four walls in order to obtain an optimal coverage of the environment for speaker localization and tracking.

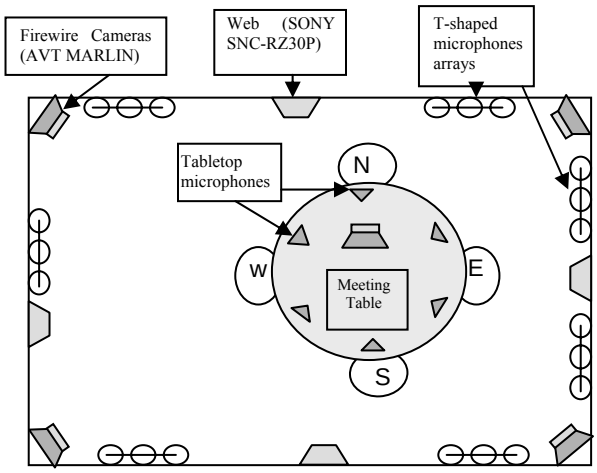


Figure 2. The experimental setting in the CHIL room

THE CORPUS

Eleven groups of four people each were recorded. The average duration was 25 minutes, the range being 0.13.08'' - 0.30.06". The total length of the audio-video corpus is 3.44.55" hours. See Table 1 for details.

Groups	Sessions length
1	0.29.00"
2	0.18.24"
3	0.26.10"
4	0.15.49"
5	0.19.06"
6	0.22.15"
7	0.30.06"
8	0.18.04"
9	0.13.08"
10	0.17.23"
11	0.15.30"
Total	3.44.55"

Table 1. Durations of the sessions

Data annotation

Each subject is identified according to cardinal points (N, S, E, W).

Currently, the following annotations are available for the data: functional relational roles (task roles and socio-emotional roles), which address facets of the group dynamics; speech activity; body activity (head position, head orientation and fidgeting activity).

In the following, we describe for each category the procedures and the annotation output.

Functional role annotation

As stated before, our main concern in developing a multimodal corpus was training a system able to automatically recognize group behaviour for the shake of the implementation of a automatic facilitator. Categories developed by Benne and Sheats concerning functional roles in small groups [2] and the two dimensional approach developed by Bales [1] revealed suitable both for a) mapping onto constellations of low-level observable

patterns that can be detected through vision and speech and b) for presenting individual profiles to participants through the relational report. The Functional Role Coding Scheme (FRCS), consisting of five labels for the Task Area and five labels for the Socio Emotional Area. The Task Area includes functional roles related to the facilitation and coordination of the tasks the group is involved in, as well as to the technical skills of the members as they are deployed in the course of the meeting. The Socio Emotional Area involves roles oriented toward the functioning of the team as a group. The reliability of the scheme was assessed on a subset of the corpus consisting of 130 minutes for the Socio-Emotional Area and 126 minutes for the Task Area. Two trained annotators coded five participants on the Socio-Emotional Area and five in the Task Area. The Cohen's K (Cohen, 1960) was used to assess inter-annotator agreement. Following [9] the agreement on the roles of the Task Area is good ($0.6 < k < 0.8$) while the agreement on the roles of Socio-Emotional Area is on the borderline between being good and moderate ($0.4 < k < 0.6$). For a more detailed description of the coding scheme, including information about its reliability, see [7, 11].

A synthetic description of FRCS follows.

The Task Area Functional Roles

Orienteer (o). S/He orients the group by introducing the agenda, defining goals and procedures, keeping the group focused and on track and summarizing the most important arguments and the group decisions.

Giver (g). S/He provides factual information, states his/her beliefs and attitudes about an idea and answers to questions.

Seeker (s). S/He requests suggestions and information, as well as clarifications, to promote effective group decisions..

Procedural technician (PT). S/He uses the resources available to the group, managing them for the sake of the group.

Follower (f). S/He listens, does not participate actively to the interaction.

The Socio-Emotional Functional Roles

Attacker (a). S/He deflates the status of others, expresses disapproval, attacks the group or the problem.

Gate-keeper (gk). S/He is the moderator within the group, who mediates the communicative relations: s/he encourages and facilitates the participation and regulated the flow of communication.

Protagonist (p). S/He takes the floor without need to be consulted driving the conversation, assuming a personal perspective and asserting his/her authority.

Supporter (su). S/He shows a cooperative attitude demonstrating understanding, attention and acceptance as well as providing technical and relational support.

Neutral (n). S/He passively accepts the idea of others, serving as an audience in group discussion.

Functional role annotations consists of tuples ⟨*role-type*; *participant-code*; *role-code*; start: *start-time*; end: *end-time*; duration: *duration*⟩ (Table 2).

task; w; o; start:621.466; end:645.965; duration:24.499023
task; w; g; start:647.769; end:806.186; duration:158.41699
task; w; g; start:831.091; end:835.619; duration:4.528015
task; w; o; start:855.022; end:1783.348; duration:928.32605
task; w; g; start:1784.843; end:1878.873; duration:94.03003
……

Table 2. Sample annotation of meeting video recordings

For instance, the tuple ⟨task: w; o; start: 621.466; end: 645.965; duration: 24.499023⟩ refers to the role of orienteer ('o') belonging to the 'task' area, as played by participant *w* from time 621.466 till time 645.965, for a duration of seconds 24.499023.

Speech activity

Speech activity here refers to the identification of the presence or absence of human speech, without distinguishing between verbal and non-verbal activity.

Each session was segmented by first automatically labeling the speech activity recorded by the close-talk microphones. The voice activity detector (VAD) is based on the time energy of the signal [3]. For each speaker, VAD identifies the amount of speech activity, and produces an output such as ⟨*participant-code*, *start time*, *end time*, *label*⟩, where *label* takes on the value 'speech' and 'no-speech'.

VAD's output was then manually checked and improved. In the first place, errors of the automatic annotation were removed; in particular, since subjects were close to each other, the speech activity of a subject often entered the close-talk microphone of the subject sitting nearby, giving raise to a wrong assignment.

Secondly, VAD is based on time energy, and it is not able to distinguish between verbal activity and other acoustic non-verbal events. Manual annotation purified the VAD annotation from breaths, yawns, coughing, and noises caused by the subjects when touching the microphones. Laughs were retained and annotated by means of the additional label *la*.

3D tracking of body activity

Visual cues were employed to derive head position and orientation as well as body activity.

Head position

The subjects' position in the room is tracked through head position identification. All of the 3D positions have an absolute timestamp and are referenced to an origin which is on the floor under the centre of the table. The 3D co-ordinate system for the room is oriented in the following way: X axis represents a Westerly direction, Z axis represents a Northerly direction, Y is the height from the floor. For each participant the 3D tracking produces a tuple ⟨*timestamp*; *x axis*; *z axis*; *y axis*⟩, where an absolute timestamp is followed by the cardinal point which identifies head position in the room. An example of the output is presented in table 3.

Timestamp [µs]	X	Y	Z
1124351961 746271	-1179.47	1086.32	-128.697
1124351961 839697	-1200.04	1131.63	-165.695
1124351961 935088	-1170.6	1064.99	156.321

Table 3. Sample of head position tuple

Head orientation

Stiefelhagen and colleagues [12] estimated the potential of head orientation in detecting who is looking at whom in around-a-table setting. Starting from head position detection, color and hedge features were used to track head orientation and to estimate focus of attention.

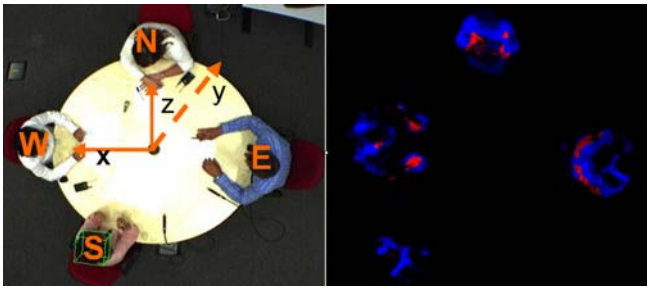


Figure 3. Head orientation and head position detection

The output from the 3D tracking consists for each subject of tuples such as ⟨*timestamp*; *head orientation*⟩. Head orientation can take on one of the following values: "down", when subject head is oriented toward the table, "S", "N", "W", "E", when the head is oriented toward South, North, West or East, each of them referring to one of the other participants (see Figure 3).

Fidgeting

Fidgeting refers to localised repetitive motions such as when the hand remains stationary while the fingers are tapping the table, or playing with glasses, etc.

Fidgeting has been tracked by using skin region features and an MHI of the convex skin polygons and temporal motion as the trigger is used. For a more detailed description see [6].

For each subject, the output of the analysis consists in the tuples ⟨*timestamp; fidgeting energy; hand/arm activity*⟩. An example of the output is the following: ⟨1124358961419507; 16; 1⟩, in which an absolute timestamp is followed by two normalised fidgeting values. The first ('16') represents the fidgeting energy of the person's body and the second ('1') represents his hand/arm activity. The normalised values are referenced to that person's most vigorous fidgeting during the entire recorded sequence, hence they are person specific.

CONCLUSION

We presented in this paper a multimodal corpus of annotated consensus decision making meetings. The corpus provides for each subject six annotation levels: manual annotation of the participants functional role and speech activity and automatic annotation of body activity during meetings, head position and orientation and fidgeting activity.

BIBLIOGRAPHY

1. Bales, R.F. *Personality and interpersonal behavior*. New York: Holt, Rinehart and Winston, 1970.
2. Benne, K.D., Sheats, P. Functional Roles of Group Members, *Journal of Social Issues* 4, (1948), 41-49.
3. Carli, G., Gretter, G. A Start-End Point Detection Algorithm for a Real-Time Acoustic Front-End based on DSP32C VME Board. In *Proc. ICSPAT*, (1992).
4. Brdiczka, O., Maisonnasse, J., and Reignier, P. Automatic detection of interaction groups. *Proc. of the 7th international Conference on Multimodal interface*. Trento, Italy. October, 2005.
5. Chen, L., Rose, R.T., Parrill, F., Han, X., Tu, J., Huang, Z., Harper, M., Quek, F., McNeill, D.,Tuttle, R., Huang, T.: VACE multimodal meeting corpus. *Proc. of Multimodal Interaction and Related Machine Learning Algorithms*. (2005).
6. Chippendale, P. Towards Automatic Body Language Annotation. Talk delivered at the *International Conference on Automatic Face and Gesture Recognition* - FG2006 (IEEE) Southampton, UK (2006).
7. Falcon V., Leonardi C., Pianesi F., Zancanaro M., Annotation of Group Behaviour: a Proposal for a Coding Scheme. *Proc. of Workshop on Multimodal Multiparty Multimodal Processing* at ICMI 2005, 39-46.
8. Hall, J. W., Watson, W. H. The Effects of a normative intervention on group decision-making performance. In *Human Relations*, 23(4), (1970), 299-317.
9. Landis J.R, Koch G.G. The measurement of observer agreement for categorical data. *Biometrics* 33: 159-174, 1977.
10. McCowan, D. Gatica-Perez, S. Bengio, D. Moore and H. Bourlard. Towards Computer Understanding of Human Interactions. In: Ambient Intelligence, E. Aarts, R. Collier, E. van Loenen & B. de Ruyter (eds.), *Lecture Notes in Computer Science*, Springer-Verlag Heidelberg, 2004, pp. 235-251.
11. Pianesi F., Zancanaro M., Falcon V., Not E. Toward supporting group dynamics. In Proceedings of AIAI'06. Athens, June 2006.
12. Stiefelhagen, R., Zhan, J., Waibel, A. Modeling focus of attention for meeting indexing. In *CHI '02 extended abstracts on Human factors in computing systems* (2002).

Language Resource Archiving supporting Multimodality Research

Peter Wittenburg, Daan Broeder, Peter Berck, Han Sloetjes, Alex Klassmann
Max-Planck-Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
peter.wittenburg@mpi.nl

ABSTRACT
At the MPI multimodal research has a long history. An increasing amount of resources is created to test scientific hypothesis. This requires proper methods and technologies to manage these resources. During the last five years mature tools[1] were developed for these purposes that guide the resources during their whole life-cycle; ELAN can be used to create accurate and complex annotations; IMDI helps the user to create useful metadata descriptions, to model the underlying relations between the resources and to search for suitable resources; LAMUS is used to upload and manage large language resource repositories and finally ANNEX and LEXUS can be used to access multimodal resources via the web.

INTRODUCTION
Investigating multimodal behavior was and is one of the key pillars in psycholinguistic research to get a deeper understanding of the mental processes underlying speech production and speech comprehension, and to better understand the relation between language and cognition. Therefore, at the Max-Planck-Institute for Psycholinguistics many studies were and are carried out using a number of modalities such as speech, prosody, gestures, signs, eye movements and body movements [1-11]. Different recording techniques such as video, audio, eye trackers, data glove, motion trackers, and ultrasonic and infrared marker devices were and are used to gather multimodal data.

As a result of the various research projects carried out at the MPI for Psycholinguistics its language resource archive now covers about 150.000 objects most of which are sessions that are linguistically meaningful units such as interviews, route descriptions, narratives etc. This is covered in about 15 Terabytes of data, a large amount since much data is digitized video. Only video (including audio) signals, sometimes taken from different perspectives, carry the rich information that is necessary to analyze and annotate human communicative behavior. In general the annotation is a manual process since these signals are often recorded in natural environments and contain utterances in minority languages or spoken by children, second language learners etc, i.e. there are no proper language models, the corpora are in general too small to estimate parameters for stochastic recognition machines and the signals contain too rich information. Signals such as eye tracking data is normally not annotated, but just used to determine relevant points in time where for example the eyes fixate a certain pattern.

A large part of this archived data is well-organized and described with the help of the IMDI (ISLE Metadata Initiative) metadata infrastructure. Although the IMDI set[2] contains elements that are typical to describe multimodality it is difficult to guess how much of the data in the archive is actually used for multimodality research. In principle, any video recording can be used to carry out such studies and researchers often forget to mark multimodality in the metadata descriptions. Therefore, we can only refer to institute projects that are started purposefully to include multimodal analysis (see annual reports[3]).

This paper will focus on the aspects of managing the technical complexity that naturally evolves when doing multimodal research, i.e. during annotation, during resource management and during analysis. It will present a framework that allows researchers to carry out multimodality work with a high accuracy and efficiently. It will not focus on either scientific results, models of multimodal behavior in production and comprehension, and encoding schemes that are used to encode human behavior. For further information about the more scientific aspects we refer to the annual reports of the MPI.

ANNOTATION SCHEMES
Despite some research projects in the area of iconic gestures, stereotypical tasks such as “route description”

[1] All tools are available or will soon be available under Open Source license. For details we refer to the following web-site: www.mpi.nl/tools

[2] www.mpi.nl/IMDI

[3] www.mpi.nl/research/publications/AnnualReports

where speech and gestures are recorded and annotated according to a more general schema [11,12] we cannot speak about the emergence and broad usage of generic schemas for the encoding of linguistic phenomena. It is understood that a bottom up description of multimodal streams starting with articulator movements is enormously complex and therefore not tractable. Instead of that researchers are looking for encoding schemes at the semantic level that allow them directly to test their scientific hypothesis. Therefore, almost all studies invent new schemes to do the linguistic encoding.

However, it was of great importance to define an annotation scheme at the structural level which is powerful enough to represent the linguistically interesting phenomena with the required flexibility and time granularity. Therefore, a.o. the EAF XML (ELAN Annotation Format) schema was developed and improved over time. It allows the researcher to define (and modify) his/her own tier setup to encode the behavior at any linguistically relevant level, to encode dependency relationships between them and to connect them where necessary to the time axis. Although all meaningful behavior is generated by mental processes we cannot speak about dependent streams, multimodal streams such as for example speech, eye movements and gestures have to be treated as completely independent, i.e., all types of timing relationships can occur [13]. Therefore, EAF has the notion of "time references" so that any single annotation can be associated with a period of time on the axis[4]. On the other hand, there will be hierarchical relations such as between the movement of the whole arm and that one of the hand in gestures or between the spoken words included in a verbal utterance. To cope with these phenomena EAF has the notion of "hierarchical relations" that can evolve to trees of different depth during encoding. Often phenomena are not linked to the time axis, but refer to an element on another tier such as in interlinearized representations. To cope with this EAF introduced the notion of "symbolic references". In linguistic encoding often phenomena are related that are on the same tier but non-adjacent time periods or that are on different tiers. An annotation format therefore has to support the encoding of such phenomena as well.

To normalize the timing encoding we should add that points in time that are used to anchor annotations are shared and stored as ordered sequence. This is in accordance with the Annotation Graph model from Bird and Liberman [14].

[4] It is assumed that preprocessing is used to unify the time axis underlying different recordings. The ELAN tool for example has a few operators to carry out this unification.

The archive still contains many multimodal annotations that were created with the MediaTagger tool [15] which was one of the first supporting flexible multimodal annotation. However, the underlying model had limitations [16] and it was not compliant with the Annotation Graph model. Due to changes in the internal Quicktime representation and due to operating system peculiarities[5] the conversion process to the more generic and XML-based EAF format turns out to be a time consuming process. Some limited multimodal annotations are also done by making use of CHAT [17]. Import modules are available to easily convert them into EAF format. With the exception of the MediaTagger resources we can argue that all multimodal annotations are available in archivable formats.

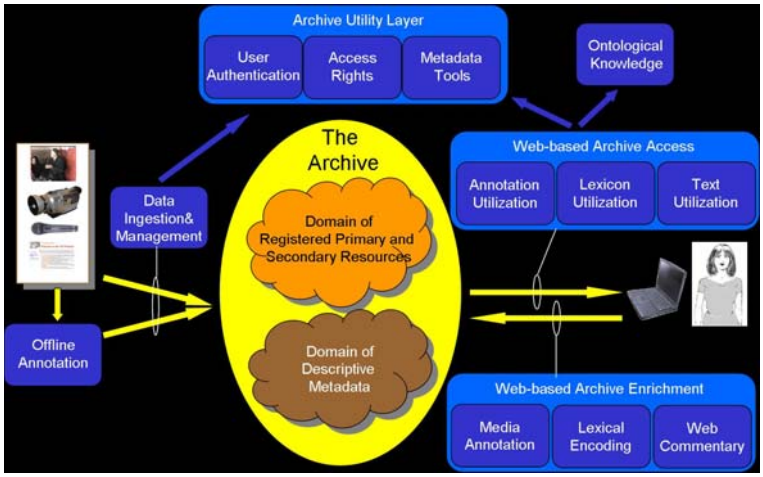
MANAGEMENT AND ACCESS ARCHITECTURE

Given the large amount of (multimedia/multimodal) resources created and stored at the MPI we had to work out an architecture that supports their whole life cycle from creation to usage and long-term preservation. The following figure gives an overview of the architecture and except the web-based annotation creation and commentary all components have been implemented and are in operation.

In the following we will briefly describe the architecture and then explain some components in more detail. The user can:

- off-line annotate and analyze recordings by ELAN (not included in the figure)
- describe them with metadata using the IMDI Editor (Metadata Tools)
- upload them into the archive with LAMUS (Data Ingestion and Management)
- define suitable access policies with AMS (User Authentication and Access Rights)
- search and browse for suitable resources with the IMDI tools (local and web-based, Metadata Tools)
- download one or complete sub-archives with the IMDI tools (Metadata Tools)
- carry out content searches on the annotations and visualize the annotated media recordings with ANNEX (Annotation Utilization, Media Annotation)
- manipulate lexica with LEXUS if applicable (Lexical Utilization, Lexical Encoding).

[5] Former MAC-OS version made a difference between data and resource fork information being both crucial for a correct interpretation. However, copying activities were carried out without awareness of this relation.



In addition, services take care that several instances of the recordings are stored at different locations in the Netherlands and Germany to ensure long-term survival.

CREATING MULTIMODAL RESOURCES

The latest version of ELAN offers many advanced features that facilitate the time-consuming manual annotation work. It not only supports the flexible annotation model described above, but it also deals with different types of media streams or time series as they occur in multimodal observations. Video signals resulting from several cameras can be displayed and analyzed together with time series data created by the many channels of for example a data glove device. All streams and the created annotations are time synchronized, i.e. selecting a time fragment in one viewer will directly update the position in the other viewers. Different options for visualizing the complex annotations that easily can contain more than 20 layers help users while navigating, comparing instances of similar phenomena etc. In many studies of multimodal interaction precise time accuracy in the order of video-frames is of greatest importance. This is the reason why we asked SPEX, the Dutch center for evaluations, to carry out measurements about the accuracy of ELAN. While earlier versions of ELAN made use of JavaMediaFramework the later versions make use of the native media libraries on the Windows platform. Together with well-chosen MPEG codecs it was shown that this solution offers the required frame accuracy in annotation and in playing. For other research projects where time accuracy is not that important, but where efficiency is the primary criterion ELAN offers a fast tagging mode.

For a detailed description of the features of ELAN we refer to the manual which is available on the web. Now ELAN has reached a level of maturity that it is a tool widely used for multimodal and sign language studies. All annotations are represented in XML which makes it a suitable candidate as well for data that has to be archived.

MANAGING RESOURCES

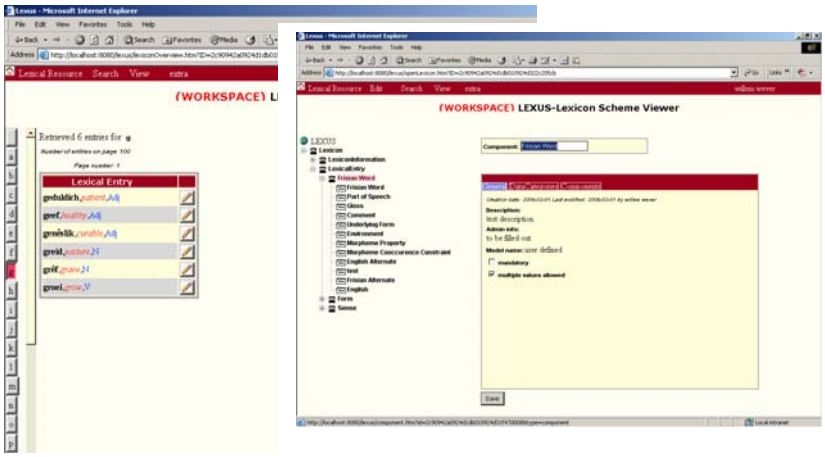
Multimodal research is accompanied in general by a large amount of resources that are related in various ways: media recordings from different devices are related since they share the same time axis, annotations are linked with specific recording channels, recording sessions are embedded in experimental setups etc. It is very important to store this relation information. In the MPI setup this is done by using the IMDI framework. IMDI allows users not only to carefully describe the sessions, but also to express the different type of relations. In doing so metadata descriptions are supporting the user in creating a well-organized browsable archive that can be accessed by searching as well as by browsing. IMDI therefore is the basis for managing a large amount of closely related resources as they are typical for multimodality research. The possibility of fine-grained metadata descriptions can be used to formulate scientifically interesting queries, in particular in conjunction with queries about the content of annotations. Also IMDI files adhere to a publicly available XML schema and therefore are in an archivable format.

The LAMUS system (Language Archive Management and Upload System) is used as a gate keeper for the language resource archive at the MPI. It allows depositors and managers to add new resources to the archive and to update existing ones in a way that the archive remains coherent and consistent. A user can request a workspace for some period of time, define an uplink in the archive as anchor point for new resources, create corpus structure, add metadata descriptions and integrate resources. Once all manipulations in the workspace have been carried out to the satisfaction of the user or after a validation procedure has been taken place, the workspace content can be uploaded to become part of the archive. LAMUS, therefore, controls the consistency of the archive and with the help of configurable resource type files its coherence. The configuration file specifies the accepted formats and when parsers are available checks the formal correctness of the ingested files.

LAMUS has an access management component that allows depositors or managers to define access policies and rights with powerful commands such as “make all audio resources in this sub-corpus available to all”. Policies can be defined that determine the kind of declarations (code of conducts etc) users have to accept before getting access to resources that are protected. Another important aspect is that LAMUS initiates the creation and updating of indexes for fast searching in metadata and annotations whenever a new resource is uploaded.

ACCESSING RESOURCES

The MPI archive provides a number of access methods knowing that researchers have different wishes. The most simple is to browse and search in the metadata domain to find useful resources. Once found they can be downloaded or viewed with a normal plug-in. Many researchers, however, want to carry out analysis on their computer by using own software and therefore want to access and operate on a number of files. They are offered a Tree-Copier option in the metadata browser that allows them to specify a sub-corpus in the archive and download all resources (or only those of specific type). The metadata and corpus structure information is also copied so that the user has a complete local copy of this sub-archive, that can also be browsed and searched using local tools just as the mother archive. Tools can then be used to carry out some manipulations off-line and later, with the help of LAMUS, that part could be uploaded again.



The right part shows a view on the structure and the information about the data categories used in a typical LMF-based lexicon with LEXUS. The left part gives one of the possible views of the content which can be easily browsed for example by selecting the begin character.

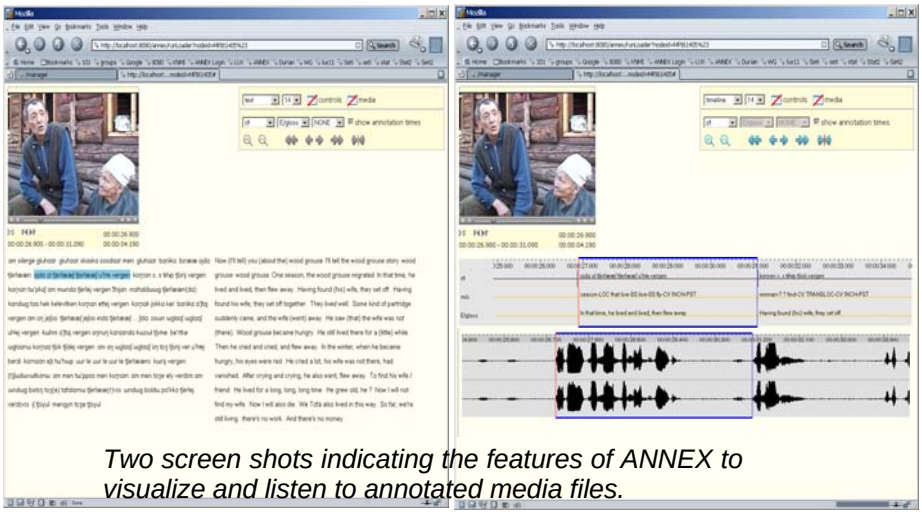
More interesting, however, are web-based applications that allow the user to immediately visualize complex objects such as multimedia lexica or multimodal annotations. ANNEX is a flexible tool that allows to work on annotations in a similar way as ELAN does it. Due to the functioning of the web, no guarantees can be made for the smoothness of the media presentations. This is one of the reasons that ANNEX does not yet allow to support annotating. Also ANNEX comes along with synchronized viewers, different viewers for annotations and it offers search capabilities on the annotation content. The following figures give two views of the look and feel of ANNEX. Again the further details can be found in the manual.

Another web-based tool that can be mentioned is the LEXUS lexicon tool which allows the creation and modification of lexical information, i.e. information that is abstracted from the individual occurrence in annotations. Since LEXUS also can incorporate or link to multimedia signals (photos, audios, videos) it can be used for example to store typical signs or gestures. LEXUS is based on LMF (Lexical Markup Framework) which is a generic model currently being worked out in ISO TC37/SC4. Therefore, LEXUS can operate with a wide range of different lexical structures and content. Also LEXUS has advanced search functionality and many other features and a first interaction with ANNEX has been implemented. The following figures may give an impression of the look and feel of LEXUS. The program offers too many features amongst which is web-based structure and content manipulation and collaboration so that we like to refer to the manual for details.

CONCLUSIONS

For the many multimodal research projects carried out at the MPI and in collaboration with other institutions a complete set of technologies was developed that can support the whole life-cycle of multimodal annotations. The lack of agreed linguistic annotation schemes in multimodal research due to the specificity of the research questions increased the necessity to define and apply a flexible annotation format such as EAF and to apply a flexible lexicon format such as LMF (for the rare cases where lexical abstractions are required in multimodality projects). In ever growing corpora with many interrelated resources IMDI is an excellent way to not only create a meaningful organize, but to carry out scientifically relevant searches in combination with searches on the annotation content.

The language resource archive serves as a reliable repository that can be accessed in several ways leaving enough flexibility for the individual researcher. Its dynamic nature and the move towards web-based applications require the introduction of Unique Resource Identifiers and a smart linguistically motivated versioning. Both will be included in the next LAMUS versions. Still many checks have to be added to ensure consistency of the representations at the structural, format and metadata level. But this has to be balanced with the requirement of



Two screen shots indicating the features of ANNEX to visualize and listen to annotated media files.

flexibility. Still many tools generate formats that are not schema-based and therefore difficult to validate.

The MPI will continue to develop its technology and continue to make it available to other interested institutions under Open Source licenses. A first external installation was finished successfully at Lund university, other external setups will follow.

REFERENCES

[1] W.J.M. Levelt (1980). Online processing contraints on the properties of signed and spoken language. In Biological Constraints on linguistic form. U. Bellugi, M. Studdert-Kennedy (eds.). Vgl. Chemie, Weinheim.

[2] G. Richardson (1984). Word recognition under spatial transformation in retarded and normal readers. Journal of Experimental Child Psychology 38, 220-240.

[3] S. Kita, J. Essegbey (to appear). Pointing left in Ghana: How a taboo on the use of the left hand influences gestural practice. Gesture.

[4] S. Kita (1998). Expressing a turn at an invisable location in route direction. In Ernest Hess-Lüttich, J.E. Müller & A. vanZoest (eds.), Signs & SPace. 159-172. Tübingen: Narr.

[5] A. Özyürek, S. Kita (1999). Expressing manner and path in English and Turkish: Differences in speech, gestures, and conceptualization. In M. Hahn and C. Stones (eds.), Proceedings of the 21 st Annual Meeting of the Coginitive Science Society. 507-512. Amsterdam.

[6] M. Gullberg, K. Holmqvist (2001). Eye tracking and the perception of gestures in face-to-face interaction vs. on screen. In C. Cave, I. Guaitella, S. Santi (Eds.), Oralite et gesturalite: Interactions et comportemetns multimodaux dans la communication (pp. 381-384). Paris: L'Harmattan.

[7] H. Lausberg, S. Kita (2001). Hemispheric specialization in spontaneous gesticulation investigated in split-brain patients. In C. Cave, I. Guaitella, S. Santi (Eds.), Oralite et gesturalite: Interactions et comportemetns multimodaux dans la communication (pp. 431-434). Paris: L'Harmattan.

[8] M. Seyfeddinipur, S. Kita (2001). Gesture and dysfluency in speech. In C. Cave, I. Guaitella, S. Santi (Eds.), Oralite et gesturalite: Interactions et comportemetns multimodaux dans la communication (pp. 266-270). Paris: L'Harmattan.

[9] N. Enfield. 2002. Hand pointing in Laos: form and function in a locality description task. MPI Annual Report 2002. Nijmegen.

[10] U. Zeshan. 2004. Sign Language Typology Project. MPI Annual Report 2004. Nijmegen.

[11] S. Kita, I. v. Gijn, H. vd. Hulst (1998). Movement Phases in Signs and Co-speech Gestures, and their Transcription by Human Coders. In I. Wachsmuth and Martin Frühlich (eds.), Gesture and Sign Language in Human-Computer Interaction, Vol. 1371: 23-35. Proceedings of the International Gesture Workshop Bielefeld, Lecture Notes in Artificial Intelligence. Berlin: Springer Verlag.

[12] S. Kita, I. v. Gijn, H. vd. Hulst (2000). Gesture Encoding. MPI Internal Report.

[13] H. Brugman, P. Wittenburg, St. Levinson, S. Kita (2002), Multimodal Annotations in Gesture and Sign Language Studies. LREC 2002 Conference. Las Palma, Mai

[14] S. Bird and M. Liberman. 2001. A formal framework for linguistic annotation. Speech Communication, 33(1,2):23-60.

[15] www.mpi.nl/world/tg/CAVA/mt

[16] H. Brugman and P. Wittenburg. 2001. The application of annotation models for the construction of databases and tools. IRCS Workshop on Linguistic Databases, University of Pennsylvania. Philadelphia.

[17] http://childes.pst.cmu.edu

Analysis of gesture expressivity modulations from cartoons animations

Nicolas Ech Chafai
University of Paris 8 / France Télécom R&D
n.chafai@iut.univ-paris8.fr

Catherine Pelachaud
University of Paris 8
c.pelachaud@iut.univ-paris8.fr

Danielle Pelé
France Télécom R&D
danielle.pele@francetelecom.com

ABSTRACT
In this paper, we describe a gesture expressivity analysis of a character in a conversational interaction. To determine gesture properties that attract and maintain perceptual attention, we study if there exists some effects in the field of gesture expressivity modulations. We study this role at a low level (gesture phase) and at a higher level (discourse structure) based on a corpus annotation of Tex Avery (MGM) cartoons. First results of the analysis point out synchronization properties between modulations and gesture phase or discourse structure.

Author Keywords
Gesture expressivity, gestures and discourse, corpus annotation, visual attention, 2D and 3D animations.

INTRODUCTION
We are currently developing an embodied conversational agent ECA endowed with an expressive and communicative behavior (Hartmann *et al.*, 2005; Poggi & Pelachaud, 2000). Our aim is to endow an ECA animation system with the capability to attract the attention of a user at specific points of the ECA animation. At the gesture level we have to set which gestures provide semantic information (Kendon, 2004), and which gestures attract the gaze of the other interactant: as Cosnier said in preface of Calbris (2003), there are gestures that carry a meaning, and gestures that manage communication and have pragmatic functions.

We look at how gesture expressivity varies in the animation. We also investigate how gesture expressivity properties can act as a pragmatic tool. Our hypothesis is that gesture expressivity modulations could partly play this role. These modulations could provide, by a sudden change in the perceived behavior of the speaker, some of her intentions to the listener.

Our approach is based on the analysis of traditional animations: animators have developed sharp skills over decades in elicitating empathy and in regulating attentional behavior of spectators through character's movement and expressions; we aim at taking into account theses skills to develop our application.

To get precise data on gesture expressivity modulations, we annotate each expressivity parameter defined in Hartmann *et al.* (*op. cit.*), not at a gesture unit or phrase level, but at a gesture *phase* level. We describe our choices for each of the annotated parameters.

In the remaining of this paper, we first present a state of the art of the works that apply traditional animation features to 3D animation, and that study more specifically gesture abilities to attract listener's gaze attention. Then we describe our corpus we used to analyze gesture expressivity in cartoons, and we precise our annotation methodology. Finally we describe the results that were observed in the analysis process, and that give a first view on the role played by gesture expressivity modulations during a conversational interaction.

STATE OF THE ART
Some previous works already tried to produce 3D animations based on traditional animation. Several fondamental principles of traditional animation (Thomas & Johnston, 1981) have been applied to 3D animation: Choi *et al.* (2004) proposed a system able to computationaly apply the principle of anticipation on a 3D animation: through the production of a backward movement over the following movements, this principle leads to direct the spectator's attention towards the place of the action. Lance *et al.* (2004) studied animators' abilities to express emotion and empathy in cartoon characters, and built up a system able to generate an expressive gaze for a virtual character. Bregler *et al.* (2002) captured the animations of 2D objects (deformable or not) by following some feature points; this follow-up allows one to animate in the same way different kinds of 2D or even 3D objects. Not only the movement is identically produced, but it also preserves the same expressivity. But theses works do not resolve the question whether imitating 2D animation onto 3D animation is perceptually acceptable by a spectator or not. Lasseter (1987) pointed out how the *principles* from 2D animation

could be successfully applied in 3D animation; however the perception that the spectator has could change if we limit 3D animation to a 2D *imitation* and if we do not look at finding to which extent the 2D animation principles could be interpreted. In our work, we try to find some new rules of 2D animation that could be applied in a gestural animation of 3D characters.

In the domain of human gesture study, there exists works dealing more specifically with gesture ability to attract listener's attention. Eye tracking techniques allow researchers to follow where and when a listener gazes at, and in particular on which gestures he gazes at. This type of disposal was adopted by Gullberg & Holmqvist (1999) to study which are the elements that lead to gaze at a particular gesture; laterality seems to play a preponderant role, as opposed to self centred gestures. With the same kind of disposal, Barrier *et al.* (2005) have determined that through the use of deictic signals, a speaker is able to redirect listener's focal attention toward his gestures, or toward a virtual space built by his gestures. In cartoons (Thomas & Johnston, *op. cit.*) noticed how efficient an animation that could be understood from its silhouette is; this observation complements results from Gullberg & Holmqvist by adding a notion of point of view: a same body gesture can change silhouette type depending from where we are looking at it. Our work aims at determining new criteria that could attract spectator's gaze attention through some kinds of gesture expressivity properties, and to implement these criteria in an ECA.

CORPUS

We base our corpus on two videos from Tex Avery cartoons (MGM). Each of theses videos lasts about ten seconds. Our choice of a low level analysis (described later) leads to a corpus with little data. In regard of our aim to animate conversational agents, we chose sequence showing a conversational interaction between characters; the first one serves as basis for our analysis, the second has been used to verify the results from the first one. One of theses videos comes from the cartoon *Blitz Wolf* (1942): it displays a pig character trying to convince two other pigs to protect their selves against a wolf's threat[6]. The other video comes from *Henpecked Hoboes* (1946): in this cartoon, the main two characters are George and Junior[7] who are trying to catch an hen to feed themselves; in the sequence that we are interested in, George explains to Junior the set of actions they will have to perform to reach their goal. Theses two sequences exhibit two different discourse goal: in the first one the pig aims to incitate and advice; in the second one George aims to communicate informations.

ANNOTATION DESCRIPTION

To get precise data on the modulations of gesture expressivity, we annotate the expressivity on a gesture phase level. Kendon defines gesture *unit*, gesture *phrase*, and gesture *phase*, as three different levels in the gesture production (2004, chap. 7). There are different kinds of gesture phase; Kendon organizes them around the phase of stroke recognized as the expressive part of the gesture: preparation, stroke, post-stroke-hold, and recovery. Kita *et al*. (1997) refine theses phases and distinguish: preparation, stroke, hold and independent hold, retraction, and partial retraction.

In our analysis, we are using most of the phases described by Kita *et al*. For sake of simplicity we consider 'independent hold' as having the same function as 'hold'; no distinction in both terminologies is made. And we add the phase of anticipation: it refers directly to one of the fundamental principles of animation as described in Thomas & Johnston (1981); from our point of view it seems necessary to add this phase in the analysis. Thus, we consider the following set of gesture phases (Kita *et al.*, *op. cit.*, Kendon, *op. cit.*, Kipp, 2003):

- Anticipation: preceding a gesture phase, the arm may produce a backward movement. This happens due to motor constraints, but also to get spectator's attention focusing on the following movement;
- Preparation: the arm moves to the location where the speaker wants to produce his stroke;
- Stroke: expressive phase of gesture, it is produced synchronously or anticipates the verbal referent;
- Hold: the stroke may be hold for a while;
- Recoil: following the stroke, the arm may recoil to emphasize this stroke;
- Retraction: the arm moves to a rest position;
- Partial retraction: before the arm finishes moving to a rest position, another gesture starts and thus ends up the retraction.

The expressivity parameters we chose for our annotation are those implemented by Hartmann *et al.* (2005) in their conversational agent Greta. They correspond to: *fluidity* is the smoothness and continuity of overall movement (*e.g.* smooth, graceful versus sudden, jerky); *power* is the dynamic properties of the movement (*e.g.* weak/relaxed versus strong/tense); *spatial extend* is the amplitude of movements (*e.g.* amont of space taken up by body);

[6] Produced right in the middle of WW2, this cartoon is a short propaganda film: the animators are displaying Big Bad Wolf under A. Hitler's features and are warning how dangerous he is. The main pig figures the judgement value of the American state. Animators are figuring this pig to display to American people what kind of behavior they have to adopt towards WW2: they have to support war effort. Obviously, the title of *Blitz Wolf* directly refers to the "Blitz Krieg" practiced by Hitler.

[7] Refers to George and Lennie characters from J. Steinbeck's novel "Of Mice and Men" (1937).

repetitivity is the tendency to rhythmic repeats of specific movements along specific modalities.

Three values are available for each parameter: positive, neutral, negative. And we define parameters with a set of criteria:

- Fluidity: it corresponds either to the level of continuity between successives phases, or to the movement curvature, or even to presence of an anticipation phase;
- Power: it stands for the shape opening (opened / closed), or the acceleration of the arm, or even for the continuity in tension at the end of the movement;
- Spatial expansion: we define it as the gestural space, or the swivel angle, or even as the point of view from which the gesture is seen *ie.* with a high or a low silhouette (Thomas & Johnston, *op. cit.*);
- Repetitivity: repetition of the gesture stroke.

Both analyzed videos are annotated using Anvil tool (Kipp, *op. cit.*), which allows us to precise each value of expressivity parameter for each of the gestural phases. Then, we observe and notice the modulations in gesture expressivity: that is, we analyze the variation over time of each expressivity parameter. We are not interested in finding out which particular parameter varies. Rather we concentrate on the variation itself of the parameters.

On the one hand we try to find some kinds of correlations between theses modulations and the production of the corresponding gesture. On the other hand we try to find some correlations between theses modulations and the structure of the verbal utterance, in order to observe if there is any regularity in it.

OBSERVATIONS

We have annotated for each gesture of our corpus the value of each expressivity parameter. When analysing the data we do not consider the annotated value of each parameter as such but we look at the variation over time of these values. This analysis is based on one of the two annotated videos; the second is used to verify the results we obtain. We observe two types of variations that are found over each expressivity parameter. We, now, consider no more the value of each expressivity parameter but these variations which are:

- Irregularities: it corresponds to a brief time period (a single gesture phase) in which the annotated modality has a sudden change of value, and then comes back at the original one just after this phase. For example, it happens when a character produces a powerful sequence of movements, except for a single phase that is produced with a low power (Figure 1a);
- Discontinuities: it corresponds to a sudden change in the annotated modality. For example, it happens when a character of the animation produces a sequence of movements with a low power, succeeding to a sequence with powerful movements (Figure 1b).

That is, each time a sudden variation in gesture expressivity occurs, it is defined as a discontinuity; but if this variation directly precedes another sudden variation we will speak in terms of irregularities. Figure 1 illustrates graphically these concepts.

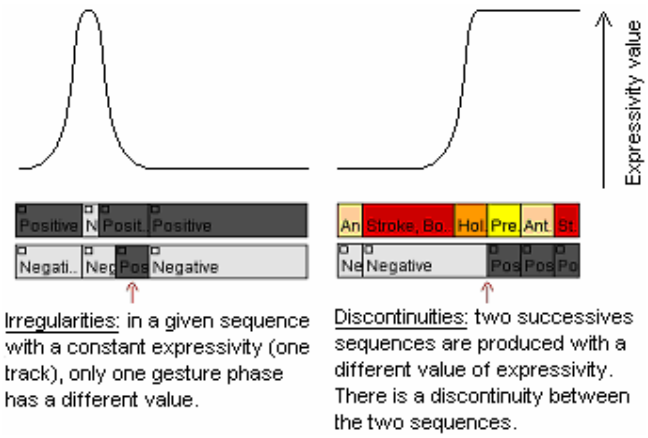


Figure 1a and 1b: Irregularities and Discontinuities

Each occurrence of these two modulation types have been noticed [8]: (4; 8) for irregularities, and (10; 6) for discontinuities. Some invariance appear to inform on their role in a conversational interaction in a cartoon, as described in the following sections. There is differences in results quantity of the two videos; this is partly due to a difference in the quantity of gesture repetitions for each video and the structure of their utterances.

THE FUNCTIONS OF IRREGULARITIES

From the annotation, we observe that irregularities seem to play a role of anticipation by linking similar elements of the enunciative structure as: occurrences of gesture repetitions (2; 7), performatives of a same general class (Poggi & Pelachaud, 2000) (1; 1), gesture phrase (1, 1).

By linking similar structures, irregularities are able to perform the role of an AND connector that allows the spectator to anticipate the behavior the character will display. Following the principle of anticipation (Thomas & Johnston, 1981), this property should enhance the visibility of gesture, *ie.* to enhance our propensity to gaze at this particular gesture.

[8] In the form (2; 3), we indicate that in the analyzed video there were two occurrences of a modulation type involved in a particular property, and three occurrences in the video used to verified results.

THE FUNCTIONS OF DISCONTINUITIES

We also observe that discontinuities may perform a relation of contrast. This relation may take diverse forms. It could enhance the emphasis on a specific gesture by contrasting it from the others (6; 1): over a whole sequence produced with low fluidity, only a single gesture phrase (and not phase, that would have lead to an irregularity) has been produced with high fluidity. That leads to an isolation effect of this gesture phrase. It could also contrast the action verbs of the utterance when they are gesturally illustrated (3; 2): each occurrence of these gestures is produced with a specific expressivity. Another form of discontinuity was noticed when the speaker enunciates a new type of general class of performative (1; 2), he changes his expressivity. Thus, discontinuities are a way to oppose different kernels of the enunciative structure.

Theses different functions of discontinuities seem to be closely linked to a relation of contrast between each of the levels they are referring to. This relation is defined as the speaker's intention that the addressee recognizes, by comparison, similarities and differences of the kernels of the enunciative structure (Mann & Thompson, 1988).

THE FUNCTIONS OF THE MODULATIONS

By summarizing the functions performed by the two kinds of modulations in gesture expressivity, it appears that they act at the different levels of the enunciative structure, and that they do not depend on the kind of performative act the speaker enunciates.

Modulations appear as a pragmatical tool. We noticed that irregularities could affect the spectator's attention through theirs anticipation properties. Discontinuities perform a relation of contrast that suggests an other attentional effect: as Feyereisen (1997) noticed "*communication supposes to perform contrasts. A signal is perceived with more clarity if it is distinguishable from noise or other signals*" (p. 39).

CONCLUSION

We have presented an annotation schema to study gesture expressivity modulations. We annotated a corpus of 2D cartoons and analyzed at the gesture phases level. This analysis leads us to extract some functions of the modulations in gesture expressivity that seem to act as a relation of similarity (irregularities), and as a relation of contrast (discontinuities). We are aware that our study shows some limitations (size of corpus, analysis by one annotator) and its results should be used with precaution.

Nevertheless it appears that the modulations are a kind of pragmatical resource that could have an interest in the animation of ECA, and that could act on the attention that the spectator bears on a gesture. In the near future, an evaluation process will simulate the previous results in an ECA application to test whether they are efficient or not for attracting one's gaze attention and interest.

REFERENCES

Barrier, G., Caelen, J. & Meillon, B. (2005). La visibilité des gestes: Paramètres directionnels, intentionnalité du signe et attribution de pertinence. *Workshop Francophone sur les Agents Conversationnels Animés*. Grenoble, France (2005), 113-123.

Bregler, C., Loeb, L., Chuang, E. & Desphande, H. Turning to masters: Motion capturing cartoons. *SIGGRAPH 2002* (2002).

Calbris, G. *L'expresion gestuelle de la pensée d'un homme politique*. Paris, CNRS Editions, 2003

Choi, J., Kim, D. & Lee, I. Anticipation for facial animation. *CASA'04*, Geneva, Switzerland, CGS (2004).

Feyereisen, P. La compréhension des gestes référentiels. *Geste, cognition et communication*, PULIM (1997), 20-48.

Gullberg, M. and Holmqvist, K. Keeping an eye on gestures: Visual perception of gestures in face-to-face communication. *Pragmatics and Cognition* 7 (1999), 35-63.

Hartmann, B., Mancini, M. & Pelachaud, C. Implementing expressive gesture synthesis for Embodied Conversational Agents. *Gesture Workshop* (2005)

Kendon, A. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

Kipp, M. *Gesture generation by imitation: From human behaviour to computer character animation*. Faculties of Natural Sciences and Technology, Boca Raton, Florida (2004).

Kita, S., Van Gijn, I. & Van der Hulst, H. Movement phases in signs and co-speech gestures, and their transcription by human coders. *Gesture Workshop*, Bielefeld, Germany, Springer-Verlag (1997).

Lance, B., Marsella, S. & Koizumi, D. Towards expressive gaze manner in embodied virtual agents. *AAMAS Workshop on Empathic Agents*, New-York (2004).

Lasseter, J. Principles of traditional animation applied to 3D computer animation. *ACM Computer Graphics 21*, 4 (1987).

Mann, W. and Thompson, S. Rhetorical Structure Theory. Toward a functional theory of text organization. *Text 8*, 3 (1988), 243-281.

Poggi, I. & Pelachaud, C. Performative facial expressions in animated faces. *Embodied Conversational Agents*. J. Cassell, S. Prevost, E. Churchill. Cambridge, Mass., MIT Press (2000), 155-188.

Thomas, F. and Johnston, O. *Disney animation, The illusion of life*. New-York, USA, Abbeville Press, 1981.

Synthesizing Gesture Expressivity Based on Real Sequences

G. Caridakis, A. Raouzaiou, K. Karpouzis, S. Kollias
Image, Video and Multimedia Systems Laboratory, National Technical University of Athens
9, Heroon Politechniou str., 15780, Athens, Greece
{gcari, araouz, kkarpou}@image.ece.ntua.gr, stefanos@cs.ntua.gr
+302107723037

ABSTRACT
In this paper we describe an approach to synthesize gestures via the tools provided in the MPEG-4 standard, using the output of the analysis and taking into account the extracted values of expressivity parameters. We animate emotional gestures, using a symbolic representation of human emotion, based on real video sequences and we extract conclusions regarding the performance of every gesture. The results of the synthetic process can then be applied to emotional ECAs.

Author Keywords
Gesture analysis, MPEG-4, expressivity parameters

INTRODUCTION
Both analysis and synthesis of hand gestures constitute an important part of human computer interaction (HCI) [1]. Sometimes, a simple hand action, such as placing a person's hands over his ears, can pass on the message that he has had enough of what he is hearing; this is conveyed more expressively than with any other spoken phrase. To benefit from the use of gestures in HCI it is necessary to provide the means by which they can be interpreted by computers.

Since the processing of visual information provides strong cues in order to infer the states of a moving object through time, vision-based techniques provide at least adequate, alternatives to capture and interpret human hand motion. At the same time, applications can benefit from the fact that vision systems can be very cost efficient and do not affect the natural interaction with the user. Analyzing hand gestures is a comprehensive task involving motion modeling, motion analysis, pattern recognition, machine learning, and even psycholinguistic studies.

Our system uses as input image sequences and tracks the head and the hands of the actor. Following, we can estimate the MPEG-4 BAP (Body Animation Parameters) for every gesture and extract some important expressivity features. All the results are used to the synthetic and lifelike reconstruction of every gesture.

The presented system of the synthetic gesture reconstruction is illustrated in Figure 1:

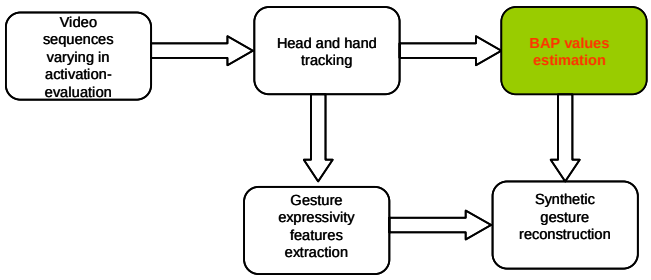


Figure 1: Synthetic Gesture Reconstruction

GESTURES ANALYSIS USING EXPRESSIVITY PARAMETERS

The System Input
The input image sequences of the presented system are videos captured at an acted session including 7 actors, every one of them performing 7 gestures. Each gesture was performed several times with the student-actor impersonating a different situation. Namely the gestures performed are: "*explain", "oh my god"* (both hands over head)*, "leave me alone", "raise hand"* (draw attention)*, "bored"* (one hand under chin)*, "wave", "clap"*.

The different acted situations-emotions are illustrated in Table 1:

Gesture class	quadrant of Whissel's wheel [2]
explain	**(0,0), (+, +), (-, +), (-, -)**
oh my god	**(+, +), (-, +)**
leave me alone	**(-, +), (-, -)**
raise hand	**(0,0), (+, +), (-, -)**
bored	**(-, -)**
wave	**(0,0), (+, +), (-, +), (-, -)**
clap	**(0,0), (+, +), (-, +), (-, -)**

Table 1: Acted Emotions

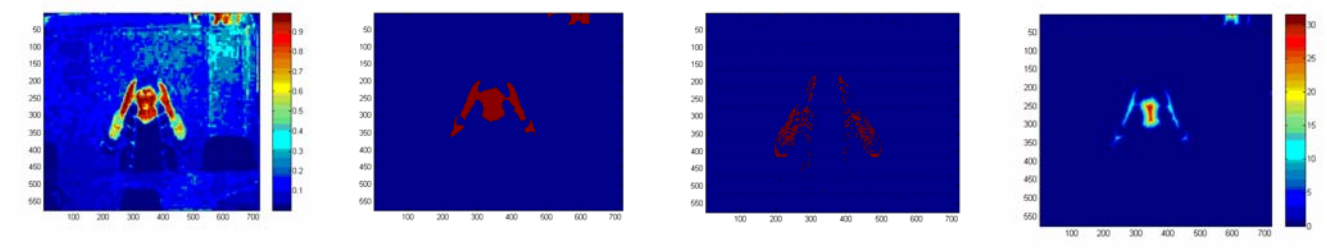
Some of the gesture-emotion combinations were not performed since it did not make much sense reproducing, for example, a "bored" gesture expressing joy. That led us to a whole of 7 actors x 20 variations (Table 1) of the 7 basic gestures =140 image sequences.

Head and Hand tracking

Several approaches have been reviewed for the head-hand tracking module. The major factors taken under consideration are computational cost and robustness, resulting in an accurate near real-time skin detection and tracking module. The general process involves the creation of moving skin masks, namely skin color areas that are tracked between subsequent frames [9]. By tracking the centroid of those skin masks, we produce an estimate of the user's movements. A priori knowledge concerning the human body and the circumstances when filming the gestures was incorporated into the module indicating the different body parts (head, right hand, left hand). For each frame (*Figure 2*) a skin color probability matrix is computed by calculating the joint probability of the Cr/Cb image values (*Figure 3a*). The skin color mask is then obtained from the skin probability matrix using thresholding (*Figure 3b*). Possible moving areas are found by thresholding the pixels' difference between the current frame and the next, resulting in the possible-motion mask (*Figure 3c*). This mask does not contain information about the direction or the magnitude of the movement, but is only indicative of the motion and is used to accelerate the algorithm by concentrating tracking only in moving image areas. Both color and motion masks contain a large number of small objects due to the presence of noise and objects with color similar to the skin. To overcome this, morphological filtering is employed on both masks to remove small objects. All described morphological operations are carried out with a disk-structuring element with a radius of 1% of the image width. The distance transform of the color mask is first calculated (*Figure 3d*) and only objects above the desired size are retained. These objects are used as markers for the morphological reconstruction of the initial color mask. The color mask is then closed to provide better centroid calculation. For the next frame, a new moving skin mask is created, and a one-to-one object correspondence is performed. Object correspondence between two frames is performed on the color mask and is based on object centroid distance for objects of similar (at least 50%) area. In the case of hand object merging and splitting, e.g., in the case of clapping, we establish a new matching of the left-most candidate object to the user's right hand and the right-most object to the left hand. The described algorithm is lightweight, allowing a rate of around 12 fps on a usual PC during our experiments, which is enough for continuous gesture tracking. The object correspondence heuristic makes it possible to individually track the hand segments correctly, at least during usual meaningful gesture sequences. In addition, the fusion of color and motion information eliminates any background noise or artifacts, thus reinforcing the robustness of the proposed approach.

The tracking algorithm is responsible for classifying the skin regions in the image sequence of the examined gesture based on the skin regions extracted from the described method. Skin region size, distance wrt the previous classified position of the region, flow alignment and spatial constraints. These criteria ensure that the next region selected to replace the current one is approximately the same size, close to the last position and moves along the same direction as the previous one as long as the instantaneous speed is above a certain threshold. As a result each candidate region is being awarded a bonus for satisfying these criteria or is being penalized for failing to comply with the restrictions applied. The winner region is appointed as the reference region for the next frame. The criteria don't have an eliminating effect, meaning that if a region fails to satisfy one of them is not being excluded from the process, and the bonus or penalty given to the region is relative to the score achieved in every criterion test. The finally selected region's score is thresholded so that poor scoring winning regions are excluded. In this case the position of the body part is unchanged wrt that in the previous frame. This feature is especially useful in occlusion cases when the position of the body part remains the same as just before occlusion occurs. After a certain number of frames the whole process is reinitialized so that a possible misclassification is not propagated.

Figure 2



(a) **(b)** **(c)** **(d)**

Figure 3

Gesture Expressivity Features Extraction

To define the expressivity parameters we searched through the literature of perception studies to see which parameters were investigated [3, 4]. Six dimensions representing behavior expressivity are defined. The expressivity dimensions have been designed for communicative behaviors only. Each dimension acts differently for each modality. For an arm gesture, expressivity works at the level of the phases of the gesture: for example the preparation phase, the stroke, the hold as well as on the way two gestures are co-articulated [5, 6]. We consider six dimensions of expressivity:

- Overall activation
- Spatial extent
- Temporal
- Fluidity
- Power/Energy
- Repetitivity

Overall activation is considered as the quantity of movement during a conversational turn. In our case it is computed as the sum of the motion vectors' norm: $OA = \sum_{i=0}^{n} |\vec{r}(i)| + |\vec{l}(i)|$. Spatial extent is modeled by expanding or condensing the entire space in front of the agent that is used for gesturing and is calculated as the maximum Euclidean distance of the position of the two hands: $SE = \max(|d(\vec{r}(i) - \vec{l}(i))|)$. The average spatial extent is also calculated for normalization reasons. The temporal parameter of the gesture determines the speed of the arm movement of a gesture's meaning carrying stroke phase and also signifies the duration of movements (e.g., quick versus sustained actions). Fluidity differentiates smooth/graceful from sudden/jerky ones. This concept seeks to capture the continuity between movements, as such, it seems appropriate to modify the continuity of the arms' trajectory paths as well as the acceleration and deceleration of the limbs. To extract this feature from the input image sequences we calculate the sum of the variance of the norms of the motion vectors. The power actually is identical with the first derivative of the motion vectors calculated in the first steps.

The testbed used for comparing the emotionally enriched gestures is GRETA [7]. The mechanisms employed to animate all the expressivity features described above are partly based on the attributes of the TCB Splines used to animate the virtual character. Details about the actual implementation can be found in [8].

EXPERIMENTAL RESULTS

Figure 4 illustrates the gesture *"oh!my god"*. The values for the six dimensions for two different subjects are presented in the diagram of Figure 5. The values shown are normalized.

The values of the results for the different gestures for a) overall activation, b) spatial extent, c) fluidity, d) power/energy are illustrated in Figures 6(a-d), while Figure 7 illustrates the mean values of the six expressivity parameters for three actors and Figures 8(a) and (b) illustrate respectively the mean values of *Overall Activation* and *Power* for positive and negative values of activation. As expected, the values of gestures lying in first and second quadrants (positive activation) are higher.

Some of the frames of the synthesized gesture are illustrated in Figure 9. The tool used for the synthesis is GretaPlayer [7].

Figure 4 Frames from the video of subject 21

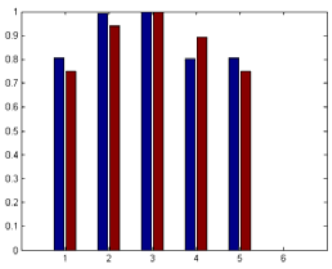
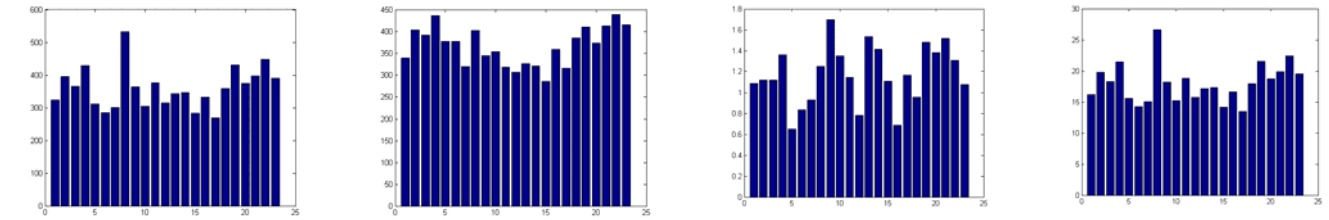


Figure 5



(a) overall activation **(b)** spatial extent **(c)** fluidity **(d)** power/energy

Figure 6

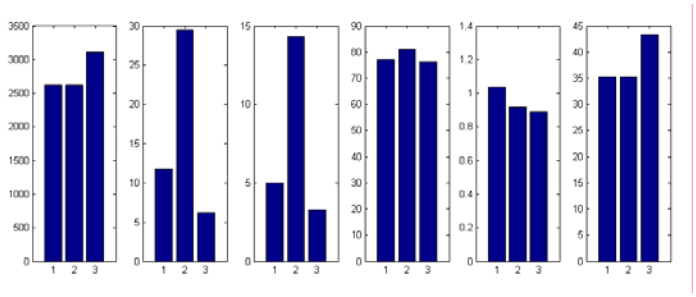
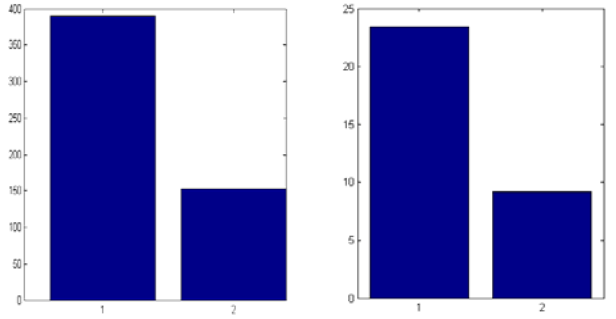


Figure 7: Mean values of the six expressivity parameters for three actors



(a) mean values of *Overall Activation* **(b)** mean values of *Power*

Figure 8

Figure 9

CONCLUSIONS

Analysis and expressivity features extraction of a broader set of gestures are necessary in order to evaluate our results. The conclusions concerning the gestures belonging to different quadrants are very useful to further analysis but also to the synthesis of these gestures. The results of the synthetic process can then be applied to emotional ECAs and make the interaction more lifelike.

REFERENCES

1. Wu, Y. and Huang, T.S., "Hand modeling, analysis, and recognition for vision-based human computer interaction", *IEEE Signal Processing Magazine*, 18(3): 51-60, May 2001.
2. Whissel, C.M., The dictionary of affect in language, Emotion: Theory, Research and Experience: vol. 4, The Measurement of Emotions, R. Plutchik and H. Kellerman, Eds., New York: Academic, 1989.

3. Hartmann, B., Mancini, M. and Pelachaud, C., Implementing Expressive Gesture Synthesis for Embodied Conversational Agents. Gesture Workshop (2005) Vannes
4. Wallbott, H.G, Bodily expression of emotion. European Journal of Social Psychology, 28:879–896, 1998.
5. Harrigan, J.A., Listener's body movements and speaking turns. Communication Research, 12(2):233–250, 1985.
6. Gallaher, P., Individual differences in nonverbal behavior: Dimensions of style. Journal of Personality and Social Psychology 63 (1992)
7. de Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V. and De Carolis, B., From Greta's mind to her face: modeling the dynamics of affective states in a Conversational Embodied Agent. *International Journal of Human-Computer Studies*, 59, 81-118, 2003.
8. Maurizio Mancini, Bjoern Hartmann, Catherine Pelachaud, Non-verbal behaviors expressivity and their representation, PF-star report 3.
9. Martin, J.-C., Caridakis, G., Devillers, L., Karpouzis, K., Abrilian, S., Manual Annotation and Image Processing of Multimodal Emotional Behaviours in TV Interviews, accepted for publication to LREC06.

An Annotation Scheme for Conversational Gestures: How to economically capture timing and form

Michael Kipp
DFKI, Germany
kipp@dfki.de

Michael Neff
MPI Informatik, Germany
neff@mpi-inf.mpg.de

Irene Albrecht
MPI Informatik, Germany
albrecht@mpi-sb.mpg.de

Figure 1. Selected frames of a source video (top) and a kinematic animation (bottom). The animation re-created the motion of the gesturing arm/hand of the original video from a manual annotation of the video which is based on our annotation scheme.

ABSTRACT
The empirical investigation of human gesture stands at the center of multiple research disciplines, and various gesture annotation schemes exist, with varying degrees of precision and annotation effort. We present a gesture annotation scheme for the specific purpose of automatically generating and animating character-specific hand/arm gestures, but with potential general value. We focus on how to capture temporal structure and locational information with relatively little annotation effort. The scheme is evaluated in terms of how accurately it captures the original gestures by re-creating those gestures on an animated character using the annotated data. This paper presents our scheme in detail and compares it to other approaches.

Author Keywords
Embodied Agents, Gesture Generation, Multimodal Interfaces

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
Animated characters are useful in a wide range of applications like interfaces, games and movies. Generating nonverbal behavior for artificial bodies remains a challenging research task. One important technique for reproducing human-like gestures is to analyze original human behavior [7,9]. This can be done using motion capture or by manually annotating video data. While motion capture has unequalled precision, the video annotation approach has other advantages: it is an indirect observation method where people are less aware or unware of the observation, and arbitrary material (e.g. TV shows) can be analyzed, even of people otherwise unavailable. Moreover, the acquired data is usually encoded on an abstract level that can be understood and analyzed by conversational analysts, linguists, ethologists and computer animators alike, whereas motion captured data can only be interpreted with significant computational and human effort.

If the annotated data is to be used with an animation system that can create arbitrary motions for a humanoid character, the need for precise positional data becomes highly important, especially if you want to capture the specific style of a speaker. Speakers do not only differ in what and when they gesture, but also *where* they gesture. For instance, the "raised index finger" can be displayed quite shyly near the chest or dominantly above the head. We believe that such locational variation is integral to personal style. When encoding positional information, the question arises as to how faithfully that encoding reflects the original movement. Successfully re-creating the original motion from the encoded data would prove that something essential must have been captured by the annotation (see Figure 1).

Annotation schemes for human movement can be classified according to the amount of detail they capture, where high detail seems to be proportional to high annotation cost and a low level of abstraction. On one side of the spectrum lies the Bern system [2,3], where a large number of degrees of freedom are manually annotated, thus resembling modern motion capture techniques. While it results in fine grained, purely descriptive and reliably coded data which can be reproduced easily with a synthetic character, annotation effort is immense. In addition, the resulting data is hard to

interpret. It does not abstract away from even minor variations and the amount of data is so massive that it is hard to put it in relation to the accumulated knowledge about gesture structure and form found in the literature. On the other end of the spectrum, lies Conversational Analysis, where the written speech transcription is used as a basis and gestures are annotated by inserting brackets in the text for beginning and end of the gesture [4]. Gesture form is captured by either a free-form written account or by gestural categories which describe one prototypical form of the gesture. Such information would be too informal or too imprecise for automatic character animation. Thus, a key decision in annotation is: how much do you abstract? Or, how large are your equivalence classes?

We propose a scheme that makes a conscious compromise between purely descriptive, high-resolution approaches and abstract interpretative approaches. We restrict ourselves to hand/arm movement to identify the most essential features of a gesture before moving to other body parts. Our scheme encodes positional data but relies on an intelligent "time slicing", based on the concept of movement phases, to determine the most relevant time points for position encoding. It is based on the observation that transition points between phases correspond to key frames in traditional animation. Moreover, we use the concept of a gesture lexicon, well known in Conversational Analysis, where each lexeme contains some generalized information about form. Lexemes can be taken as prototypes of recurring gesture patterns. When encoding lexeme type for an annotated gesture in the video material all this general data is implicitly encoded as well.

TARGET SCENARIO

Our annotation scheme aims at the specific application of gesture generation for an animated character. However, we think that the annotation scheme will be of general interest in the interdisciplinary fields of multimodal and gesture research. The needs that arise from *animating* gestures on the basis of manual annotation provide good guidance on the essential descriptive parameters of human gestures.

The generation approach we aim at "imitates" a human speaker's gesture behavior using statistical models and a database of sample gestures, both extracted from video annotations [7]. For this application, the annotation scheme must capture the temporal and spatial structure of a gesture, and its relation to speech. Since original gesture samples are re-used in generation, the annotation should make it possible to re-create original gestures in synthetic animation. On the other hand, the annotation should be as economical as possible in terms of annotation effort.

Our video corpus consists of selected video clips from two TV talk shows, featuring two different speakers.

ANNOTATION SCHEME

While gestures appear to be quite arbitrary in form at first glance various researchers found them to have fairly stable form, even if they are not clear emblems [5]. Conversational gestures have no clear meaning and may even be a byproduct of speech. However, there seem to be shared lexica or inventories of conversational gestures [12]. For instance, the metaphoric gesture "progressive" [11], where a speaker makes a circular movement with the hands, seems to occur when talking about progress, movement or the future [1]. Another universal gesture is the "open hand" where the speaker holds the open hand in front of the body, showing the palm [4,11]. While such forms appear to be universal, there is still much inter-speaker and intra-speaker variation in terms of the *exact* position of the hands and their ensuing trajectory. To investigate and capture these variations was one driving force of our work.

We use the Anvil video annotation tool [6] for our purposes, which allows annotation on multiple tracks. Coding consists of adding annotation elements which can be complex attribute-value objects.

Capturing Temporal Structure

We capture the temporal structure of a gesture by first identifying the basic movement phases [4,8,11]:

preparation > hold > stroke > hold > retraction

where the stroke is the most energetic part of the gesture while the preparation moves to the stroke's starting position. Holds are optional still phases which can occur before and/or after the stroke. Kita et al. [8] identified *independent holds* which can occur instead of a stroke. The retraction returns to a rest pose (e.g. arms hanging down, resting in lap, or arms folded). Kita et al. refined the notion of stroke by defining a multiple stroke that includes small beat-like movements that follow the first stroke, but seem to belong to the same gesture. In our scheme, a stroke contains a "number" attribute to capture the number of within-stroke movements.

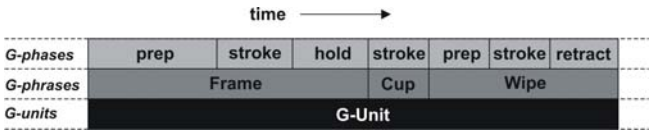


Figure 2. Gesture Annotation on Three Anvil Tracks.

To annotate phases in Anvil, the coder specifies beginning and end times of a phase as well as phase type (prep, stroke, etc.) and stroke number. On a second track, the coder combines phases into gestures, also called gesture phrases (Figure 2). In this way, we store the gesture's internal temporal structure, most importantly begin/end times of the stroke or independent hold. On a third track, we combine gestures into gesture units. A gesture unit is a sequence of contiguous gestures in which the hands do not return to a rest pose until the end of the last gesture [4,11]. This allows us to examine a speaker's g-unit structure. For instance, the average number of gestures, patterns of recurring lexeme sequences etc.

Capturing Spatial Form

In order to capture the spatial form we aimed at the best compromise between exactness and economy. For the sake of economy we make two important assumptions: (1) the most "interesting" configurations occur exactly at the beginning and at the end of a stroke, and (2) bihanded gestures are symmetrical. Although many gestures are actually asymetrical, most of them can be approximated quite well with symmetrical versions.

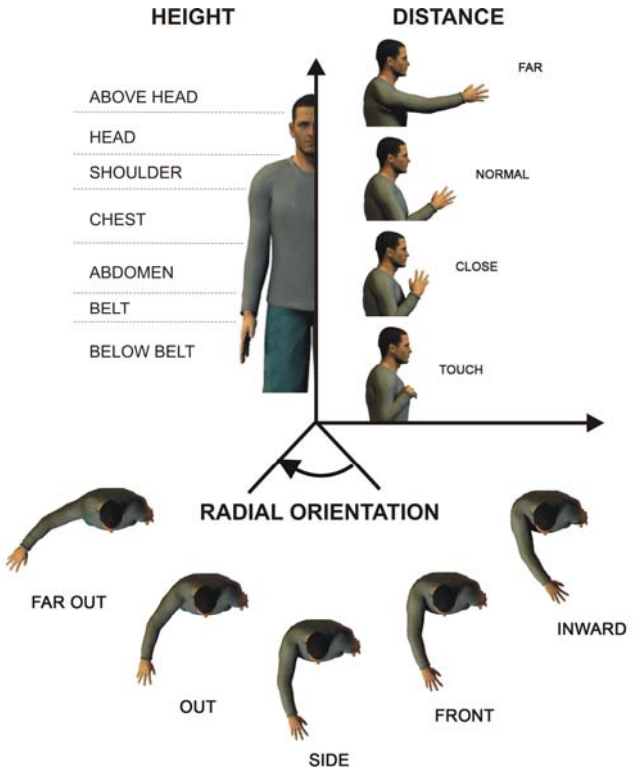


Figure 3. Our three dimensions for hand position.

The first two parameters encoded are handedness and whether the trajectory of the hand(s) in the stroke phase is straight or curved. Next, we have to capture the start and end positions of the hands/arms for the stroke. For a single position we encode three dimensions for hand location (Figure 3) and encode elbow inclination in a fourth dimension (Figure 4). The dimensions were chosen such that (1) we have sufficient granularity for later animation and (2) it is quick and reliable to annotate video, which explains the selection of landmarks like "shoulder", "belt" and intuitive terms like "normal". For bihanded gestures, we also encode hand-to-hand distance for added precision by marking the hands on the video screen; we extended Anvil to handle this new kind of "spatial annotation" (Figure 5). The hand-to-hand distance is normalized by dividing it by the shoulder width which must be encoded each time the size of the displayed speaker changes due to camera movement.

Figure 4. A fourth dimension encodes elbow inclination.

In summary, for each stroke based gesture we encode 2 positions where each position is expressed by 5 attributes. Adding handedness and trajectory gives us 12 attributes to code for the spatial form of a gesture. Independent holds only require 1 position, for the beginning of the hold.

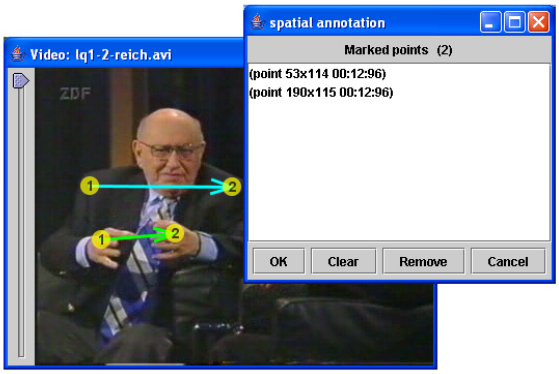


Figure 5. Annotating 2D points in Anvil: Shoulder width (top arrow) and hand-to-hand distance (bottom arrow).

Capturing Membership to Lexical Category

A number of parameters are determined by the gesture's lexeme, including: handshape, palm orientation and exact trajectory. For each lexeme, these parameters can be either fixed (definitional parameter), restricted to a range of values, or arbitrary. To annotate lexemes on the phrase track, we rely on a simplified version of the gesture lexicon collected in [7] where 79% agreement in lexeme coding experiments is reported. Typical lexemes include: RaisedIndexfinger, Cup (open hand), FingerRing (thumb touches index finger) and Progressive (circular movement). We found 31 and 35 different lexemes for our two speakers with an overlap of 27 lexemes between the two.

Capturing the Relationship To Speech

Once shape and lexeme are determined, the gesture must be connected to speech. When annotating real data, we found that the claim that gesture stroke and lexical affiliate always co-occur [11] is often wrong. Therefore, we encode co-occurrence and lexical affiliate in different attributes. Co-occurrence is not trivial. The gesture stroke has a temporal extension and may overlap with many co-occurring words. Choosing every overlapping word does not reflect our intuition of gesture-word co-occurrence. We use the following heuristics to automatically annotate co-occurrence: From the words overlapping with the stroke, choose (1) the word carrying the emphasis, if present, or else (2) the last word. Lexical affiliation is a more difficult task. We rely on the gesture literature and sometimes intuition when it comes to connecting gestures to the

speech's semantics (cf. [7]). The lexeme usually gives some direction: for pointing gestures look for personal pronouns like "you", "his" etc., for the metaphoric "Cup" gesture look for the closest noun, for the metaphoric "Progressive" gesture look for the closest verb or noun that expresses movement or temporal relation.

EVALUATION BY RE-CREATION

Any transcription scheme must be measured by two factors. First, how well the annotation reflects the original motion (usually dependent on application or experiment). Second, how reliably the annotation can be performed by human coders. While we have not yet tested reliability, we propose a method for the first criterion: re-creating the gestures with an animated agent [2]. Using only the pure annotation information already produced satisfying results. Adding information that had been manually collected for specific gesture lexemes (hand shape/orientation, trajectory) produced animations that very precisely matched the original motions. See Figure 1 for an impression of our re-creation experiments.

RELATED WORK

In this section we focus on two highly related schemes (for a general overview see [13]). The Bern scheme [2,3] is an early, purely descriptive scheme which is reliable to code (90-95% agreement) but has high annotation costs. For a gesture of, say, 3 seconds duration, the Bern system encodes 7 time points with 9 dimensions each (counting only the gesture relevant ones), resulting in 63 attributes to code. In comparison, our scheme needs a maximum of 12 attributes for a gesture's positional information. FORM is a more recent descriptive gesture annotation scheme [10]. It encodes positions by body part (left/right upper/lower arm, left/right hand) and has two tracks for each part, one for static locations and one for motions. For each position change of each body part the start/end configurations are annotated. Coding reliability appears to be satisfactory but, like with the Bern system, coding effort is very high: 20 hours coding per minute of video. By contrast, we measured an average effort of only 1 hour per minute of video for our scheme. We explain this stark difference by our very focused approach to gesture annotation. While FORM encodes every movement of independent body parts, we hypothesize that the stroke (or independent hold) alone carries the definitional part of the gesture. Of course, both FORM and the Bern System also encode other body data (head, torso, legs, shoulders etc.) that we do not consider. However, since annotation effort for descriptive schemes is generally very high, we argue that annotation schemes must be targeted at this point to be manageable and have research impact in the desired area.

CONCLUSION

We presented an effective gesture annotation scheme for gesture generation that appears to be a good compromise between detail and economy. Re-creating animations showed that the scheme captures the original motions quite well. We consciously restricted the project to arm/hand movement, ignoring the rest of the body for the sake of simplicity. However, other body parts should be included in the future. Another future issue is to test coding reliability.

We think that the main reason why our annotation so successfully captures gestures in an economic way is that it consciously focuses the annotation effort by exploiting the concept of gesture phases. The coder first identifies those time points most worth investing annotation work in and only then encodes the time-consuming positional data. Another "trick" is to move recurring patterns to a lexicon of gestures. By identifying the lexeme of a gesture, the coder specifies a number of features that need not be transcribed. While our annotation scheme has obvious drawbacks in what it does not capture (handshape, asymmetry, etc.) it is straightforward to extend if necessary. However, part of our intent in creating this scheme was to find the most economical solution for descriptive gesture annotation.

REFERENCES

1. Calbris, G. *Semiotics of French Gesture*. Indiana University Press, Bloomington, Indiana, 1990.
2. Frey, S. *Die Macht des Bildes*. Verlag Hans Huber, Bern, 1999.
3. Frey, S., Hirsbrunner, H. P., Florin, A., Daw, W. and Crawford, R. A Unified Approach to the investigation of Nonverbal and Verbal Behavior in Communication Research. In: *Current Issues in European Social Psychology*, Doise, W. and Moscovici, S. (eds.), Cambridge University Press (1983), 143-199.
4. Kendon, A. *Gesture: Visible Action as Utterance.* Cambridge University Press, 2004.
5. Kendon, A. An Agenda for Gesture Studies. In: *The Semiotic Review of Books* **7** (3), 1996, 8-12.
6. Kipp, M. Anvil - A Generic Annotation Tool for Multimodal Dialogue. In *Proc. Eurospeech 2001*, 1367-1370.
7. Kipp, M. *Gesture Generation by Imitation: From Human Behavior to Computer Character Animation.* Dissertation.com, Boca Raton, FL, USA, 2004.
8. Kita, S., van Gijn, I. and van der Hulst, H. Movement Phases in Signs and Co-speech Gestures, and Their Transcription by Human Coders. In *Gesture and Sign Language in Human-Computer Interaction*, Wachsmuth, I. and Fröhlich, M. (eds.). Springer (1998), 23-35.
9. Kopp, S., Tepper, P. and Cassell, J. Towards integrated microplanning of language and iconic gesture for multimodal output. In: *Proc. Intl. Conf. Multimodal Interfaces 2004*, 97-104.
10. Martell, C. FORM: An Extensible, Kinematically-Based Gesture Annotation Scheme. In *Proc. ICSLP 2002*, 353-356.
11. McNeill, D. *Hand and Mind: What Gestures Reveal about Thought*, University of Chicago Press, 1992.
12. Webb, R. *Linguistic Properties of Metaphoric Gestures*. PhD thesis, University of Rochester, New York, 1997.

13. Wegener Knudsen, M., Martin, J.-C., Dybkjær, L., Machuca Ayuso, M., Bernsen, N.O., Carletta, J., Heid, U., Kita, S., Llisterri, J., Pelachaud, C., Poggi, I., Reithinger, N., van Elswijk, G., Wittenburg, P. *Survey of Multimodal Annotation Schemes and Best Practice*. ISLE Deliverable D9.1, 2002.

Using FORM Data to Predict Phase Labels

Craig Martell
Computer Science Department
Naval Postgraduate School
One University Circle
Monterey, California 93943
cmartell@nps.edu

Joshua Kroll
Mathematics Department
Harvard University
One Oxford Street
Cambridge, Massachusetts 02138
jkroll@fas.harvard.edu

ABSTRACT
In this paper we present an augmentation of the FORM gesture corpus and describe experiments using FORM to predict gesture phase, i.e. preparation, stroke, and retraction. We compare these results to experiments using motion-captured data to predict the same. Interestingly, the FORM data, which is gathered via annotation, does significantly better than the motion-captured data.

Author Keywords
Multimodal Corpora, Gesture, Machine Learning

ACM Classification Keywords
H.5.2 Information Interfaces & Presentation: User Interfaces

INTRODUCTION
FORM was developed as a fine-grained, gesture coding scheme that allows annotators to capture exhaustively the constituent parts of the gestures of video-recorded speakers.

FORM represents gesture data as a collection of 4-tuples, <*startTime, endTime, attribute, value*>. The attribute/value pair represents some change during the specified interval. For example, if there was upper-arm rotation during an interval, the attribute would be Upper Arm: Rotation, and the value would be the degree of rotation. All of the possible attribute/value pairs are described extensively in [5]. It is useful to think of these 4-tuples as labeled arcs in a graph, the nodes of which are the timestamps. In FORM, gestural movement is segmented visually. That is, the annotators would focus first on one attribute in order to mark the timestamps of changes, and then replay the video to focus on the next attribute.

The total FORM dataset is approximately 22 minutes long. There are approximately 3500 arcs/minute, for a total of roughly 77000 arcs.

In [4], we presented preliminary results and discussed future research directions. In this paper, we describe refinements to the FORM annotation scheme and present the results of new inter– annotator-agreement studies and machine-learning experiments using the FORM dataset to predict gesture phases.

OVERCOMING AMBIGUITIES IN FORM
There are known ambiguities in the FORM system as described in [4] and in greater detail in [5]. One concerns the *Upper Arm*: *Location* attributes that specify biceps direction. While anatomically it seems accurate to describe the upper arm rotation by degrees of rotation rather than by the direction of the biceps in free space—as is done in FORM—a problem arises when defining the neutral position of the arm rotation.

In light of this ambiguity, we have extended FORM to include additional attributes and values for wrist location. These allow us to specify in a $5 \times 5 \times 5$ grid the x, y, and z coordinates of the wrist, with (3, 3, 3) being the speaker's sternum. For some purposes the full description of location and movement will be desired, e.g., an experiment concerning how change in elbow flexion correlates with some aspect of pragmatics. However, for other purposes, we need simply specify the location of the wrist—along with the upper-arm lift—at key points along the movement. This should suffice to recreate the motion.

INTER-ANNOTATOR AGREEMENT: THE BAG OF ARCS METHOD
Our experiments with FORM-annotation show that with sufficient training, agreement among the annotators can be very high. Table 1 shows inter-annotator agreement results for two annotators annotating a file of four gesture excursions. The results were generated by the bag-of-arcs algorithm, as given in [5]. Essentially, given an annotation graph, we combine all the arcs for each annotator into a bag. Then all the bags are combined and the intersection is extracted. This intersection constitutes the overlap in annotation, i.e., where the annotators agreed. The percentage of the intersection to the whole is then calculated to get the scores presented.

We calculate the intersection with tolerances for time and value chosen, as described below. Each of the annotators agreed that there were four gesture excursions. The *Precision* column gives the number of frames (at 29.97045 fps) that the annotators can be off from one another by and still be counted as having agreed. A precision of 0 frames

Gesture Excursion	Precision	Exact Match	Off-by-one-or-less
1	0 Frames	44.78	46.77
	7 Frames	64.68	68.66
	15 Frames	74.63	**80.60**
2	0 Frames	29.05	33.94
	7 Frames	61.47	70.64
	15 Frames	70.03	**80.43**
3	0 Frames	41.42	47.34
	7 Frames	47.34	56.81
	15 Frames	63.91	**79.19**
4	0 Frames	40.65	43.23
	7 Frames	59.35	64.51
	15 Frames	64.52	**71.62**

Table 1. Inter-Annotator Agreement on Jan24-09.mov

means that the two annotators had to agree on the exact start and end times of an arc in order to be counted as agreeing. Given that it is vague as to exactly where a gesture phase starts and ends, we first loosened this restriction to within 7 frames (or approximately ± .25 seconds) and then to within 15 frames (or approximately ± .5 seconds). Anything over 15 frames was deemed too tolerant. We also relaxed the algorithm by looking not only at exact matches on the value of an attribute, but also counted as matching any values that were off by no more than one. This is given in the *Off-by-one-or-less* column. As examples, let *arc1* be (428, 446, ForearmRotation, 1(and *arc2* be (427, 451, ForearmRotation, 2). Then *arc1* will match *arc2* if the tolerances are set to *frames* = 15 and *Off-by-one-or-less* = *True*. However, they will not match if *frames* = 0 or if, instead of *Off-by-one-or-less*, *Exact Match* = *True*. The Bag-of-Arcs method is similar to the one used by the IBM BLEU project to judge quality of a machine translation [7]. The FACS project also used a similar metric. They let two facial encodings match along a particular dimension if the first choice of one annotator was the first or second choice of another annotator [1].

The most tolerant measure, then, is given by the (15 *frames, Off-by-one-or-less*) cells. For inter-annotator agreement the first three excursions have agreement of approximately 80%. The fourth excursion had an agreement of 71.62%. (Note, however, that this is not so far off from intra-annotator agreement results. The average for inter-annotator agreement was 77.96%, while the average for intra-annotator agreement was 81.29%).

GESTURE SEGMENTATION: PHASES

In this section, we present experiments which use the FORM representation of gesture—which is fairly *low-level*—to predict the *medium-level* phenomenon of gesture-phase. We have purposively avoided defining what constitute an individual gesture in this project, as it is very difficult to clearly pin down the beginning and end of the constituent movements that make up a gesture excursion. Further, there is not yet a theory to describe in what ways these "kinetic" simples should combine to create a gesture. So far, in this work, we have simply picked out the beginning and end of the gesture excursion—viz., rest position to rest position. This is done with surprising consistency. Similarly, to pick out the phases of an excursion, we do not need to explain which "gesture" they make up. Instead, we only need to segment the excursion and label these segments. It is methodologically much cleaner; and, as we shall see, people do it fairly consistently.

To do this experiment, we added a *Phase* track to both the LeftArm and RightArm Groups of FORM. The annotators segmented the gesture excursion into gesture phases and labeled the phases [2]. Phases were initially of four types: Preparation, Stroke, Retraction, and Hold. Interestingly, though, the annotators were often comfortable claiming there was a phase change, while they were, at the same time, uncomfortable with classifying the new phase. For these cases, we added a fifth type: Unsure. We call the sequence of phases that describe a gesture excursion the *PSR-theory description* of that gesture, and *PSR theory* the theory that says excursions can be so divided.

Inter-Annotator Agreement: Phases

Our inter-annotator agreement study for PSR theory was done differently than the general FORM agreement study. The reason for this concerns the Unsures. Most of the time, annotators placed an Unsure in the space transitioning between two clear-cut phases. By this, we mean that Unsure served as a way to mark the penumbra between the two phases. In these cases, agreement judged using Bag-of-Arcs would return very low results. This is because the penumbra between two phases is often larger than 15 frames. This would prevent a match even under the most relaxed conditions. To counter-act this, we divided the gesture excursion into frames—each one equivalent in length to the frames of the original video—and labeled each of the frames according to the phase of which it was a part. We then simply judged the degree of agreement on the labels of the frames. So, even if one annotator had a large Unsure between a Preparation and a Stroke while the second annotator had the Preparation directly adjacent to the Stroke

	P	S	R	H	U
P	701	90	36	30	4
S	57	739	0	0	16
R	0	0	288	3	0
H	5	0	21	313	30
U	169	136	138	290	165

Table 2. Agreement 68.28%

	P	S	R
P	701	90	36
S	57	739	0
R	0	0	288

Table 3. Agreement 90.42%

the agreement score would be accurate. Tables 2 and 3 present the results of these experiments.

Table 2 is particularly interesting. This presents the result of judging agreement over *all* phase categories, including unsures. Note that the total agreement over all frames was only 68.28%. This low number is largely explained by how Unsures are used, as described above. The annotator represented by the row labels used Unsure much more often. However, we can see that—although there was strong consistency for Preparations, Strokes, and Retractions—there was also more confusion concerning Holds. In particular, the row annotator almost equally divided the column annotator's Holds between Hold and Unsure. In other words, the column annotator was more comfortable saying that there was a Hold in between two other phases than the row annotator was. Inspection of the video reveals that in many of these cases the speaker's hand are performing what we call "incidental movement." Incidental movement is movement during a phase that seems cognitively to be a Hold, even though there is some bouncing or jittery movement of the hand. Some annotators paid attention to the arm as a whole, while others concentrate on the particular part of the body. The latter method could lead to calling this incidental movement an Unsure rather than a Hold.

Thus, we ran the agreement study again, but only judged agreement on Preparations, Strokes, and Retractions. Overall agreement across these three phases was 90.42% (Table 3). As Holds are presumably important for understanding human gesturing, more work is warranted so that we can consistently annotate Hold phases.

AUTOMATIC PHASE PREDICTION: FORM VS. MOCAP

In this section we describe the results of using hidden Markov models (HMMs) to predict phase labels from the underlying kinetic representation in FORM. We conducted a number of experiments which are described extensively in [5]. In addition, for some experiments, the subject in the video was connected to a ReActor2 infrared motion-capture system. This was done so that we could compare FORM and motion-capture as different methods of gathering human gestural-movement information. Motion capture was chosen for comparison because it is considered "ground truth" for capturing bodily movement information. The best results from each of FORM and motion capture are presented below.

Experimental Overview

As mentioned above, we overcame ambiguity in FORM by adding the end-effector position. This position was given as (x, y, z) coordinates in a $5 \times 5 \times 5$ grid. If we combine these coordinates with the value of the *upperArm-Lift* parameter, we get a vector in $\mathbf{R}^4$ which describes the position of an arm at a particular frame. So, a sequence of these vectors encode the movement of an arm throughout a gesture excursion. By dividing the excursion into subsequences of these vectors such that they are co-extensive with the phase segmentation described above, we created a set of labeled data.

However, FORM annotators only put *Location* markers at critical points in the gesture. The goal was to approximate zero-crossings in the first and second derivatives. In order to create the requisite interpolated vectors, we took the R4 vectors for each *Location* point in the gesture excursion and used cubic splines to fill in the values for the intervening frames. This generated a large matrix in R4, the number of columns of which is determined by the number of frames—at 29.97045 fps—in the excursion. We then divided this large matrix in accordance with the phase segmentation to generate bins of matrices representing the different phases. Thus, we produced a bin of preparations, a bin of strokes, and a bin of retractions.

For the motion-capture experiments, we generated vectors with the same parameterization as the FORM vectors from data given by the motion-capture system. However, as the motion-capture system generated vectors for all frames of an excursion, no interpolation was necessary. We simply segmented the sequence of frames according to the human-annotated phase labels to create analogous matrices[9].

For each of these methods, we then ran the following HMM experiment. It is a version of a cross-validation method known as *Leaving-one-out* [6]. For each iteration of the experiment the training set is of size $N - 1$, while one data point, *i*, is used as held-out testing data. This process is repeated N times so each data point gets left out once. Our particular algorithm works as follows. Of the combined set of *all* phase matrices—which we will call *observations*—choose one, observation_i, at each iteration and remove it from the set of observations. Then, for each of the sets of phases Preparation, Stroke, and Retraction, generate an HMM representing that phase and train with all the samples for that phase only. Label observation_i after the hidden Markov model, *M*, which maximizes $P(\text{observation}_i|M)$. If

[9] Other methods tried included automatic smoothing of MoCap data (SimulatedFORM) [5]. However, these results were inferior to those presented here.

	Preparation			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Prep	0.35	1.00	0.52	
FORM	0.67	0.50	0.57	+5.8%
MoCap	0.61	0.49	0.54	+3.8%
	Stroke			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Stroke	0.45	1.00	0.62	
FORM	0.72	0.73	0.72	+16%
MoCap	0.69	0.60	0.64	+3.22%
	Retraction			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Retraction	0.22	1.00	0.33	
FORM	0.61	0.86	0.71	+115%
MoCap	0.46	0.76	0.57	+73%

Table 4. Precision, Recall, and F-Score Results for Various HMM Methods Using the *Craig* Data Set

the label generated for observation$_i$ matches the actual label of observation$_i$, call it a match. Finally, return observation$_i$ to the set of observations. We do this for all *i*. Our total percentage of matches is computed as 100 × (total matched/total number of *observations*).

Baseline: Call-all-*X*

The baseline used for these experiments was Call-all-*x*. Actually, Call-all-*x* is a combination of multiple baselines—one per phase—that produces particularly conservative results. For each of the phases, *x*, in the experiment, we assumed an algorithm that labels all observations as *x*. For example, the Call-all-Prep baseline labels every observation as a preparation. Precision is calculated simply as the proportion of actual preparations in the dataset. Recall will always be 1. *Mutatis mutandis* for all other phases.

Results

Table 4 presents the results of these experiments. For all three phases, both FORM and MoCap did better than baseline. Interestingly, though, FORM did significantly better than MoCap, with a *p*-value of 0.05 using a two-tailed McNemar's test. While this obviously satisfied our original goal of being at least almost as good as motion-capture, it begs the question as to why the FORM data produced better results. Further research is needed here, but we believe that the smoothing of the movement curve imposed by FORM removes much of the incidental movement that MoCap faithfully captures. The result of this smoothing is a curve with coarse-grained features which are more easily classifiable.

REFERENCES

1. Ekman, P., Friesen, W. V., and Tomkins, S. Facial affect scoring technique: A first validity study. In *Nonverbal Communication: Readings with Commentary*, Shirley Weitz, (ed.). Oxford University Press, New York, (1974), 34–50.
2. Kendon, A.. *Gesture: Visible Action as Utterance*. Cambridge University Press, Cambridge, (2004).
3. Kita, S., van Gijn, I. and van der Hulst, H. Movement Phases in Signs and Co-speech Gestures, and Their Transcription by Human Coders. In *Gesture and Sign Language in Human-Computer Interaction*, Wachsmuth, I. and Fröhlich, M. (eds.). Springer, (1998), 23-35.
4. Martell, C. Form: An extensible, kinematically-based gesture annotation scheme. In *Proceedings of the International Conference on Language Resources and Evaluation,* (2002).
5. Martell, C.. FORM: *An Experiment in the Annotation of the Kinematics of Gesture*. Ph.D. thesis, University of Pennsylvania, (2005).
6. Ney, H., Martin, S, and Vessel, F. Statistical language modeling using leaving-one-out. In *Corpus-Based Methods in Language and Speech Processing*, Steve Young and Gerrit Bloothooft (eds.). Kluwer Academic, Dordrecht, (1997), 174–207.
7. Papineni, K., Roukos, S, Ward, T, and Zhu, W. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, (2002), 311–318.

Degrees of freedom of facial movements in face-to-face conversational speech

Gérard Bailly, Frédéric Elisei, Pierre Badin & Christophe Savariaux
Institut de la Communication Parlée, UMR CNRS n°5009, INPG/Univ. Stendhal
46, av. Félix Viallet, 38031 Grenoble CEDEX, France
{gerard.bailly,frederic.elisei,pierre.badin,christophe.savariaux}@icp.inpg.fr

ABSTRACT
In this paper we analyze the degrees of freedom (DoF) of facial movements in face-to-face conversation. We propose here a method for automatically selecting expressive frames in a large fine-grained motion capture corpus that best complement an initial shape model built using neutral speech. Using conversational data from one speaker, we extract 11 DoF that reconstruct facial deformations with a average precision less than a millimeter. Gestural scores are then built that gather movements and discursive labels. This modeling framework offers a productive analysis of conversational speech that seeks in the multimodal signals the rendering of given communicative functions and linguistic events.
Author Keywords: Facial movements, model-based face tracking, expressive audiovisual speech

INTRODUCTION
When we interact with each other and even in absence of the interlocutor (e.g. when phoning), facial movements due to speech articulation are often accompanied by head movements, facial expressions and gestures, used by the speaker for underlining the meaning of the speech acts, involving the listener or elements of the environment in the discourse as well as maintaining mutual attention by back channeling. These facial movements can aid the understanding of the message, but also convey a lot of additional information about the speaker, such as his emotional or mental state. Nonverbal components in face-to-face communication have been studied extensively, mainly by psychologists. Studies typically link head and facial movements or gestures qualitatively to speech acts. Many of the more prominent movements are clearly related to the discourse content or to the situation at hand. For example, if the sets and releases of eye contact are of most importance for face-to-face interaction, much of the body language in conversations is used to facilitate turn-taking. Movements also emphasize a point of view. Some movements serve biological needs, e.g. blinking to wet the eyes. Few quantitative results have been published that clearly describe what are the basic components of the facial movements, what are their precise region of action and how they combine, and finally how such head and facial movements correlate with elements of the discourse. Eckman and Friesen studied extensively emotional expressions of faces [10] and also describe non-emotional facial movements that mark syntactic elements of sentences, in particular endings. The appropriate generation of face, hand and body movements is of most importance for Embodied Conversational Agents [5, 6] as well as for Sociable Robots [4]. The rules governing the firing of mimics and the implementation of that mimics are however often set in a very ad hoc way and results generally from intensive labeling of videos recordings with no special focus on fine-grained motion capture.
Face detection, identification and tracking as well as facial movement tracking use generally model-based approaches where speaker-specific appearance and shape models should be learned from training data [8, 9, 12]. The number of free dimensions of these models heavily influences the system's performance: this number should offer a compact search space without sacrificing a good fit with observed movements. Initial models are often trained off-line using limited hand-labeled data. Most models consider either facial expressions [16] or speech-related facial gestures [15], with few attempts that treat the global problem [3].
The work below presents our first effort in characterizing the DoF of the facial deformation of one speaker when involved in face-to-face conversation. So-called expressemes – for expressive visemes – are extracted from life interaction videos for building shape models with minimal training data. A methodology is proposed to incrementally refine these models by automatically selecting pertinent expressemes.

THE CORPORA AND RECORDING PROCEDURE
For six years we have developed a procedure for building speaker-specific fine-grained shape [17] and appearance models [11] for the face and the lips: we glue more than 200 colored beads on the speaker's face to have access to fleshpoints. We also fit generic teeth, eyes and lips models to photogrammetric video data to regularize geometric data of these important but smaller organs. Corpora were generally dedicated to the study of speech coarticulation and limited to read material as our target application was multimodal text-to-speech synthesis [2]. The speech-related facial movements of seven speakers (2 females, 5 males) with different mother languages (Arabic, English, German, French) have been "cloned". Shape models of all speakers are controlled with the 6 articulatory parameters controlling the jaw, lips and laryngeal positioning.
More recently we extended the recorded material to basic

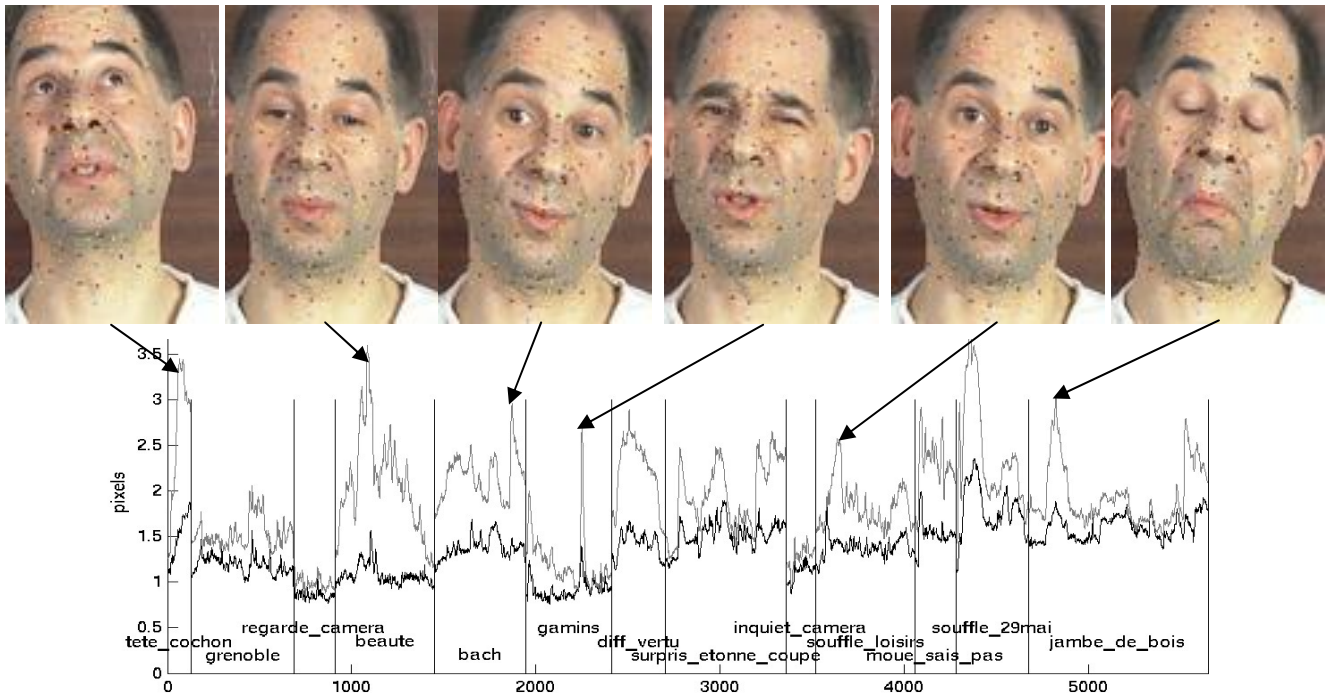


Figure 2: Comparing prediction errors of facial shapes using a model built using 52 speech visemes (light gray) with one incorporating 102 additional expressemes (dark gray), for a series of selected video sequences. The mean error lowers from 1.7 to 1.3 pixels. Frames shown at the top are generating the most important prediction errors of the speech-only model.

acted expressive speech (i.e. smiling, disgust) that most influence lip shape and to free conversation where subjects were asked first to answer to the Proust's questionnaire and then to recall and tell the most enjoyable, the most frightening and the most surprising personal experiences to the experimenter. We study here a corpus of free conversation from one subject lasting approximately 30 minutes. The speaker is filmed with three calibrated PAL cameras (front + both sides). The resulting images have a definition of almost 2 pixels per mm.

MODELLING SPEECH-RELATED MOVEMENTS

Data-driven shape models are classically built using principal component analysis (PCA). Usually a generic mesh is fitted by hand on a few dozen of representative frames. Shape parameters emerging from a PCA performed on these frames are often very difficult to interpret: they often mirror fortuitous correlations observed in the limited set of training material. Key frames should thus be chosen carefully so as to represent the diversity of facial movements usually involved in the task with maximum statistical coverage. We propose here to combine automatic feature point tracking, frame hand-labeling and statistical modeling to gather these key frames. Furthermore our shape models are built using a so-called guided PCA where a priori knowledge is introduced during the linear decomposition. We in fact compute and iteratively subtract predictors using carefully chosen data subsets [1]. For speech movements, this methodology enable us to extract six components directly related to jaw, proper lip movements and clear movements of the throat linked with underlying movements of the larynx and hyoid bone. We added to these six components two additional "expressive" components involved in our acted corpus of expressions: "smile" and "disgust" gesture that emerge from the analysis of our set of "smiling" and "disgust" visemes respectively.

TRACKING LOCAL FACIAL DEFORMATIONS

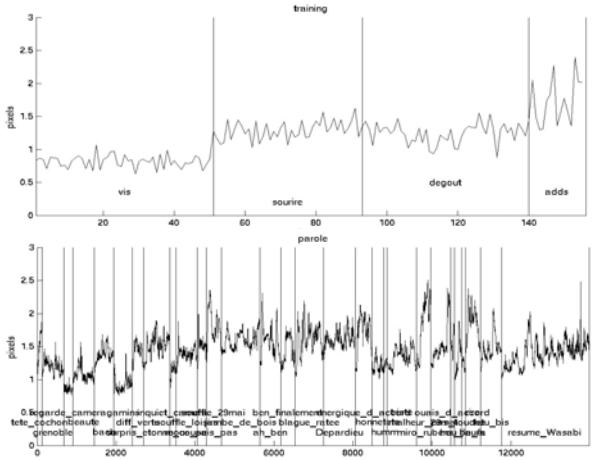
The speech-related shape model of the facial movements is then used to guide a multi-view tracker of the beads using correlation-based techniques [14]. The initial shape model only helps us to constrain the search space within proper regions of interest for each vertex of the facial mesh. The entire corpus of free conversation is then tracked. While most beads are tracked using at least two views, which enable 3D constraints to be applied, some beads are only tracked on one view, notably those located on the speaker's profile or in regions with high curvature.

The beads are tracked as patterns of 13x11 pixels. We track usually around 600 patterns per frame (compared to 250 beads on 3 views). The processing time for each frame is typically 2 seconds on a standard 2Ghz PC. We then interpret the reconstruction error of the beads summed up on all views (see Figure 2). We select automatically discourse units that have the most important reconstruction errors. We then retain the most salient frames which are precisely marked by hand, adding here the untracked beads.

ADDING SOME EXPRESSIVE FACIAL MOVEMENTS

The final objective of this selection process is to whiten the error structure by identifying and adding necessary DoF of

Figure 3: Non photorealistic synthetic views showing the effect on shape of each elementary expressive action. The face is rendered using a unique texture. From left to right: neutral, raising eyebrows, lip corners raiser (obtained from the smiling visemes), nose wrinkler (obtained from the visemes uttered with disgust), un-frowning and chin raiser. Note that lip corners raiser, nose wrinkler and chin raiser do affect lip shape.



(a)

(b)

Figure 4: Modeling and tracking errors with the final model. (a) modeling errors of training data (visemes and expressemes). (b) tracking errors for selected conversational speech sequences.

unexplained facial movements. The analysis of the selected frames reveals for example an important residual error in the region of the forehead. Three basic components have been identified and added using first principal components of given regions of selected frames: eyebrows raising/lowering, forehead frowning, and chin raising/lowering. These elementary gestures combine to control facial shapes. The effects of single elementary gestures with reference to the neutral face are shown in Figure 3. The shape model that includes speech-related and expression-related facial movements has finally 11 DoF.

ANALYSING THE EXPRESSIVE CORPUS

The beads positions for each frame of the set of read and conversational speech data has been estimated using the beads tracker. The facial movements should now be explained by the DoF of the final shape model. The Figure 4 presents the modeling and tracking errors for several thousands of frames. The modeling error for the 154 training frames is less than 1 pixel for visemes and around 1,5 pixels for expressemes (Figure 4a). Note that all these frames have been manually marked. The tracking of beads on the entire sequences from which the visemes and expressemes have been extracted reaches almost the same precision. Note that this tracking error is computed using only the positions of tracked beads (usually 75% of the beads set). Despite the fact that the full model reduces the mean tracking error at around 1,3 mm (see Figure 2 and Figure 4b) and tends to decrease the number of error bursts, significant modeling errors still remain that claim for extra DoF to be added to the final shape model (see Figure 4b).

Reliable gestural scores can be built (see Figure 5) that gather the time evolution of the shape parameters together with the speech signal and discourse labels. These gestural scores provide very valuable data on synchronization of multimodal events that participate to the encoding of distinctive communicative functions: these scores provide the necessary receptacle of ground-truth bottom-up events and theory-specific top-down interpretations.

A CASE STUDY: EYEBROWS MOVEMENTS

Eyebrows movements are known to contribute to discourse structuring [7] and are often used as redundant markers of emphasis [13]. Our preliminary analysis of 20 turns of our conversational speech data evidences two distinct eyebrows gestures as displayed in Figure 6: bursts associated with words on emphasis that co-occur with pitch accents and more global gestures coextensive with dialog acts.

COMMENTS

MPEG4/SNHC identifies 64 Facial Animation Parameters. Similarly the well-known FACS individualizes 28 facial elementary gestures – not including eyes, eyelids and head movements - that combine to produce facial mimics. It is still an open question to determine (a) what are the basic synergies between these elementary gestures that are required to encode the complex repertoire of facial mimics; (b) how they effectively combine and how they are controlled; and (c) how speaker-specific strategies implement universal or culture-specific facial attitudes.

We claim here that this repertoire may be learnt using limited resources i.e. recording a limited set of visemes and expressemes and that a dozen of

basic gestures is sufficient to reach a prediction error of about one millimeter uniformly distributed all over the face.

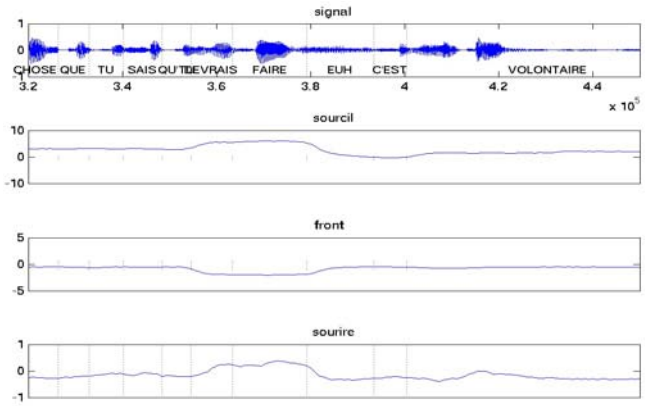


Figure 5: Gestural score for a selected speech act showing a burst of smiling ("sourire" score) during the uttering of "devrais faire". The following hesitation "euh" is also associated with lower eyebrows ("sourcil" score).

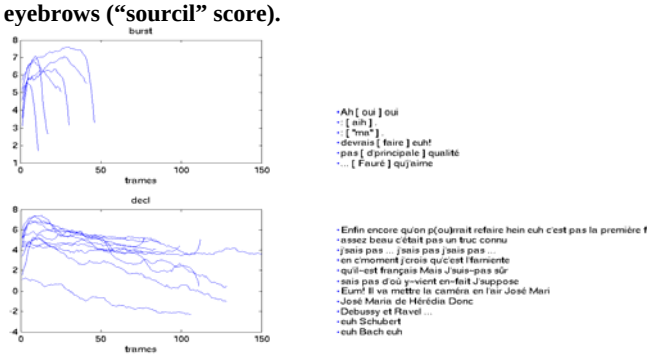


Figure 6: Time course of the eyebrow parameter. Top: bursts associated with words on emphasis. Bottom: initial burst+ declination associated with entire dialog acts.

CONCLUSIONS AND PERSPECTIVES

A productive analysis of conversational speech should combine two complementary approaches: a top-down approach that seeks in the multimodal signals the rendering of given communicative functions and linguistic events; and a bottom-up approach that reveals multimodal events that emerge from the observation of human partners in action. Combining both approaches will avoid the observer's biases and opens the route towards proper quantitative models of control and negotiation between overlapping scopes of communicative functions. The analysis of multimodal events should also be driven by entropy constraints i.e. implement coherently co-occurring communicative functions and not only result/emerge from global energy-based analysis such as PCA.

We will label gestural scores produced by our model-based gesture-aware tracker with communicative functions in order to study the scope and dynamics of their multimodal gestural instances.

ACKNOWLEDGMENTS

We thank Alain Arnal for his technical help and Ralf Baumbach for his preliminary work on the data. The paper benefited from the comments of 4 reviewers. This work has been financed by a PROCOPE grant between ICP and TUB, the GIS PEGASUS and the Rhône-Alpes region.

REFERENCES

[1] Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., and Savariaux, C. (2002) *Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images.* Journal of Phonetics, **30**(3): p.533-553.

[2] Bailly, G., Bérar, M., Elisei, F., and Odisio, M. (2003) *Audiovisual speech synthesis.* International Journal of Speech Technology, **6**: p.331-346.

[3] Beskow, J. and Nordenberg, M. (2005) *Data-driven synthesis of expressive visual speech using an MPEG-4 talking head*. in *Interspeech*. Lisbon, Portugal. p.793-796.

[4] Breazeal, C. (2000) *Sociable machines: expressive social exchange between humans and robots*. Sc.D. dissertation, in *Department of Electrical Engineering and Computer Science*. MIT: Boston, MA.

[5] Buisine, S., Abrilian, S., and Martin, J.-C. (2004) *Evaluation of multimodal behaviour of embodied agents*, in *From brows to trust: evaluating embodied conversational agents*, Z. Ruttkay and C. Pelachaud, Editors. Kluwer Academic Publishers. p. 217-238.

[6] Cassell, J. and Thórisson, K.R. (1999) *The power of a nod and a glance: envelope vs. emotional feedback in animated conversational agents.* International Journal of Applied Artificial Intelligence, **13**(4-5): p.519-538.

[7] Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., and Espesser, R. (1996) *About the relationship between eyebrow movements and F0 variations*. in *International Conference on Speech and Language Processing*. Philadelphia, PA. p.2175-2178.

[8] Cootes, T.F., Edwards, G.J., and Taylor, C.J. (2001) *Active Appearance Models.* IEEE Transactions on Pattern Analysis and Machine Intelligence, **23**(6): p.681-685.

[9] Eisert, P. and Girod, B. (1998) *Analyzing facial expressions for virtual conferencing.* IEEE Computer Graphics & Applications: Special Issue: Computer Animation for Virtual Humans, **18**(5): p.70-78.

[10] Ekman, P. and Friesen, W. (1978) *Facial Action Coding System (FACS): A technique for the measurement of facial action*. Palo Alto, California.: Consulting Psychologists Press.

[11] Elisei, F., Bailly, G., Gibert, G., and Brun, R. (2005) *Capturing data and realistic 3D models for cued speech analysis and audiovisual synthesis*. in *Auditory-Visual Speech Processing Workshop*. Vancouver, Canada

[12] Guenter, B., Grimm, C., Wood, D., Malvar, H., and Pighin, F. (1998) *Making faces*. in *SIGGRAPH*. Orlando - USA. p.55-67.

[13] Krahmer, E., Ruttkay, Z., Swerts, M., and Wesselink, W. (2002) *Pitch, eyebrows and the perception of focus*. in *Speech Prosody*. Aix en Province, France. p.443-446.

[14] Lewis, J.P. (1995) *Fast Template Matching.* Vision Interface: p.120-123.

[15] Odisio, M. and Bailly, G. (2004) *Tracking talking faces with shape and appearance models.* Speech Communication, **44**(1-4): p.63-82.

[16] Pighin, F.H., Szeliski, R., and Salesin, D. (1999) *Resynthesizing facial animation through 3D model-based tracking.* International Conference on Computer Vision, **1**: p.143-150.

[17] Revéret, L., Bailly, G., and Badin, P. (2000) *MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation*. in *International Conference on Speech and Language Processing*. Beijing - China. p.755-758.

A Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing Phenomena

Jens Allwood
University of Göteborg
jens.allwood@ling.gu.se
Kristiina Jokinen
University of Helsinki
kristiina.jokinen@helsinki.fi
Loredana Cerrato
TMH/CTT, KTH, Sweden
loce@speech.kth.se
Costanza Navarretta, Patrizia Paggio
CST, University of Copenhagen
{Patrizia,Costanza}@cst.dk

ABSTRACT
This paper deals with the MUMIN multimodal annotation scheme, which is dedicated to the study of hand gestures and facial displays in interpersonal communication, with focus on the role played by multimodal expressions for feedback, turn management and sequencing. The scheme has been tested on the analysis of multimodal behaviour in short video clips in Swedish, Finnish and Danish. These preliminary results show that the categories defined are reliable, and points at a few necessary revisions.

Author Keywords
Multimodal annotation, feedback, hand and facial gestures

INTRODUCTION
The creation of a multimodal corpus often reflects the requirements of a specific application and thus constitutes an attempt at modelling either input or output multimodal behaviour. On the contrary the MUMIN coding scheme [4], developed in the Nordic Network on Multimodal Interfaces MUMIN (www.cst.dk/mumin), is intended as a general instrument for the study of hand gestures and facial displays in interpersonal communication, focusing on the role played by multimodal expressions for feedback, turn management and sequencing. It builds on previous studies of feedback strategies in conversations [9, 1], and on work where vocal feedback has been categorised in behavioural or functional terms [2,3,7]. In what follows, we briefly describe the annotation categories starting with the functional ones, and then deal with coding procedure, materials and results from three case studies. We conclude with a few reflections on the potential applications of the scheme.

ANNOTATION CATEGORIES
The main focus of the coding scheme is the annotation of the feedback, turn-management and sequencing functions of multimodal expressions, with important consequences for the annotation process and results. First of all, the annotator is expected to *select* hand gestures and facial displays to be annotated *only* if they play an observable communicative function. Moreover, the attributes concerning the shape or dynamics of the observed phenomena are not detailed, because they only seek to capture features that are significant when studying interpersonal communication. However, the annotation of gesture shape and dynamics can be extended for specific purposes, for example to construct computer applications, without changing the functional level of the annotation.

The first kind of annotation considered is modality-specific, and concerns the expression types, the second concerns multimodal communication. For each hand gesture and facial display taken into consideration, a relation with the corresponding speech expression (if any) is also annotated. However, the scheme does not provide tags for the annotation of verbal expressions since the focus is on the facial displays and hand gestures which can be synchronized with spoken language.

Feedback
The production of feedback is a pervasive phenomenon in human communication. Participants in a conversation give feedback to show that they are willing and able to continue the interaction and that they are listening, paying attention, understanding or not understanding, agreeing or disagreeing with the message being conveyed. They elicit feedback to know how the interlocutor is reacting in terms of attention, understanding and agreement. While exchanging feedback, both speaker and listener can show emotions and attitudes. Both feedback giving and eliciting are annotated by means of the same three sets of attributes: *Basic, Acceptance*, and *Attitudinal emotions/attitudes*.

Function attribute	Function values
Basic	CPU, CP
Acceptance	Accept, Non-accept
Additional Emotion/ Attitude	Happy, Sad, Surprised, Disgusted, Angry, Frightened, Other

Table 1. Feedback Annotation Features

Basic features define hand gestures or facial displays in terms of whether they express or elicit i. continuation/contact and perception (CP), where the

dialogue participants acknowledge contact and perception of each other; ii. continuation/contact, perception and understanding (CPU), where the interlocutors also show explicit signs of understanding or not understanding of the message. The two categories capture what [9] call *acknowledgement*. *Acceptance* indicates that the interlocutor has not only perceived and understood the message, but also shows or elicits signs of either agreeing with its content or rejecting it. Basic and Acceptance can be compared with process-related and content-related in [13]. Finally, feedback annotation relies on a list of *emotions* and *attitudes* that can co-occur with one of the basic feedback features and with an acceptance feature. The list includes the six basic emotions [111,5] plus an "other" value.

Turn management

The turn management system regulates the interaction flow and minimises overlapping speech and pauses. It is coded by the three general features *Turn gain*, *Turn end* and *Turn hold*. In addition, a turn gain is either a *Turn take* if the speaker takes a turn that wasn't offered, possibly by interrupting, or a *Turn accept* if the speaker accepts a turn that is being offered. Similarly, turn end can be achieved in different ways: the speaker can release the turn under pressure (*Turn yield*), offer the turn to the interlocutor (*Turn offer*), or signal completion of the turn and end of the conversation at the same time (*Turn complete*).

Sequencing

Sequencing concerns the organisation of a dialogue in meaningful sequences, corresponding to what in other frameworks has been described as sub-dialogues, i.e. a sequence of speech acts which may extend over several turns. In other words, sequencing is orthogonal to the turn system. *Opening sequence* indicates that a new speech act sequence is starting. *Continue sequence* indicates that the current speech act sequence is going on, for example when a gesture is associated with enumerative phrases such as "the first… the second… the third…". *Closing sequence* indicates that the current speech act sequence is closed, which may be shown by a head turn or another gesture while uttering a phrase like "that's it, that's all".

MULTIMODAL EXPRESSIONS

Under normal circumstances, in face-to-face communication feedback, turn management and sequencing all involve use of multimodal expressions, and are not mutually exclusive. For instance, turn management is partly done by feedback. A turn can be accepted by giving feedback and released by eliciting information from the other party. Within each feature, however, only one value is allowed, since the focus of annotation is on the explicit communicative function of the phenomenon under analysis. For example, a head nod which has been coded as CPU (continuation/contact, perception and understanding) cannot be assigned accept and non-accept values at the same time.

An example of a multifunctional facial display coded with ANVIL [12] is shown in the frame in Figure 1: the speaker frowns and takes the turn while agreeing with the interlocutor by uttering: "ja, det synes jeg" (Yes, I think so). By means of the same multimodal expression (facial display combined with speech utterance) he also elicits feedback from the interlocutor and encourages her to continue the current sequence.

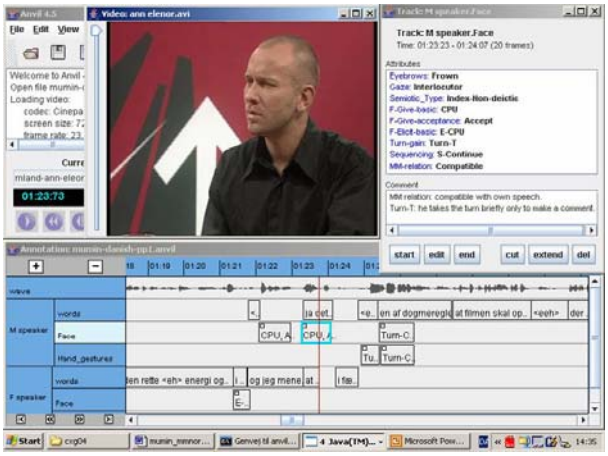


Figure 1: A multifunctional facial display: turn management and feedback

The components of a multimodal expression can have different time spans. For instance, a cross-modal relation can be defined between a speech segment and a slightly subsequent gesture. To define a multimodal relation, we make a basic distinction between two signs being *dependent* on or *independent* from each other. If they are dependent, they are either *compatible* or *incompatible*. For two signs to be compatible, they must either complement or reinforce each other, while incompatibility arises if they express different contents, as e.g. in ironic contexts.

FACIAL DISPLAYS AND HAND GESTURES

Facial displays and hand gestures are annotated with respect to the shape and dynamics of the movement. Although the categories proposed here, as already noted, are not very detailed, they should be specific enough to be able to distinguish and characterise non-verbal expressions that play a role in feedback, turn management and sequencing. They are concerned with the movement dimension of facial displays and hand gestures, and should be understood as dynamic features that refer to a movement as a whole or a protracted state. Internal gesture segmentation is not considered since it

doesn't seem relevant for the analysis of communicative functions we are pursuing.

The term *facial display* [6] refers to timed changes in eyebrow position, expressions of the mouth, movement of the head and of the eyes. The coding scheme includes features describing *General face* expressions such as *Smile* or *Scowl*, features of *Eyebrow movements,* such as *Frown* or *Raise*, features referring to *Eye movement*, features for *Gaze direction*, for movements of the *Mouth* and position of the *Lips*. Finally, a number of features refer to *Head* movements. The total number of different features is 36.

The annotation of the shape and trajectory of hand gesture is a strong simplification of the scheme used at the McNeill Lab [10]. The features, 7 in total, concern the two dimensions of *Handedness* and *Trajectory*, so that we distinguish between single-handed and double-handed gestures, and among a number of different simple trajectories analogous to what is done for gaze movement.

Finally, semiotic categories have also been defined common to both facial displays and hand gestures building on Pierce's semiotic types. They are *Indexical Deictic*, *Indexical*, *Non-deictic, Iconic* and *Symbolic*.

CODING PROCEDURE, TOOLS AND MATERIAL

The coding procedure was iteratively defined in several MUMIN workshops, and annotations have been carried out by means of the several coding tools, e.g. ANVIL [12]. The annotated material consists of a) one minute clip from an interview of a Danish actress for Danish television; b) one minute interview of the Finnish finance minister for Finnish television provided by the courtesy of the Centre of Scientific Computing; c) one minute clip from the Swedish film "Show me love".

The Danish case study

Two independent annotators with limited experience annotated gestures in the Danish clip using ANVIL. They started by annotating the non-verbal expressions of one of the interlocutors together to familiarise themselves with the coding scheme. Then they did the annotation task for the other dialogue participant independently in order to evaluate the reliability of the coding scheme.

In order to align the two annotations, it was decided that two segments referred to the same gesture if they covered the same time span, plus or minus ¼ of a second at the onset or end of the gesture. The first coder annotated 37 facial displays, and the second one 33. Of these, 29 were common to both coders. The agreement in recognition of facial gestures is thus 0.83. Concerning hand gestures, the first coder annotated 6, the second 4. Of these only two were in common. Therefore, only hand gestures have been considered for the κ-score evaluation.

The κ-scores obtained on the features concerning gesture shape and semiotic type are all in the range .83-.96 with the exception of those concerning *Gaze* (.54) and *Head* (0.2). This low agreement is partly due to the fact that one coder privileged head position over gaze (head up, no gaze), while the other in such cases ignored head movements and annotated gaze. There are also inconsistencies: in some cases the tag is *Gaze side* with the comment "away from the interlocutor", in others *Gaze other* with the comment "away from the interlocutor". Thus, the interaction of head movement and gaze needs a more careful treatment in the coding manual.

In the coding of communicative functions, on the other hand (Table 2), the annotators achieved satisfactory κ-scores with the exception of *sequencing*, particularly the feature *Continue sequence*. The issue needs further investigation.

	P(A)	P(E)	Kappa
F-Give Basic	.79	.33	.68
F-Give acceptance	.86	.25	.81
F-Give Emotion	.86	.08	.84
F-Elicit basic	.93	.33	.9
F-Elicit acceptance	1	.25	1
F-elicit emotion	.93	.08	.92
Turn-gain	.89	.33	.83
Turn-end	.93	.33	.89
Turn-hold	.96	.05	.92
Sequencing	.69	.25	.59
MM-relation	.82	.25	.76

Table 2: κ-scores for classification of communicative function features

While they show a good reliability for most of the categories used, the κ-scores don't tell us anything about the coverage of the scheme. The material in the Danish case study is quite limited, so it is not surprising that many of the categories are not used. However, it is worth noting that one of the basic feedback features, *F-elicit-acceptance*, never appears (thus the κ-score concerns the default value "none"). The other case studies show that this is an idiosyncratic characteristic of this dialogue rather than evidence of empirical inadequacy of the feature. Concerning lack of necessary categories, on the other hand, it is obvious already from this limited study that body posture, which is not included in the scheme, is important for feedback: both coders noted in their comments that a relevant movement of the torso should have been annotated.

The Swedish and Finnish case studies

The Swedish video clip consists of a one-minute emotional conversation between two actors who interpret father and daughter. They are mostly filmed in close ups of their faces, so that the hands are rarely in the picture, making it impossible to annotate hand gestures. The actor that speaks is not always in focus, so in two cases in which the actors utters a feedback expression, the face cannot be observed.

Only one expert annotator coded the film scene, so the reliability of the coding scheme was evaluated only by means of an inter-variance test, which checks whether the same coder varies their judgments over time. The coder annotated the material once and after about six months repeated the coding. A total of 12 facial displays related to feedback were coded both times, with complete intercoder agreement. The coded facial displays related to turn management functions were 12 the first time and 13 the second time, which means that the percentage of turn management identification was 95%.

Since the video-clip is extracted from a film, all the conversational moves are pre-defined and therefore only few turn-gain and turn-hold facial displays occur, moreover no sequencing facial displays or gestures were identified, probably due to the fact that the flow of discourse is pre-defined not leaving space to a spontaneous organisation of the discourse structure.

Given the emotional scene, it is not surprising that most of the feedback phenomena annotated have been labelled as *F-Give-emotion/attitude* (7, against 2 for *F-Elicit-acceptance*, and 1 for *F-Give-acceptance*, *F-Elicit-basic* and *F-Elicit-emotion/attitude*). The fact that *F-Elicit-acceptance* was used points to the fact that the category is useful, and that its absence from the Danish data is due to the different communicative situation. On the other hand, in the Swedish clip there are no examples of *F-Give basic*, which in spontaneous conversation has been found to be one of the most frequent feedback categories [8].

The distribution of turn management features was 10 for *Turn-end*, and 1 for *Turn-gain* and *Turn-hold*.

The Finnish 1-minute clip is similar to the Danish in that it is also an interview edited for broadcasting. The most important contribution of this study – still in the process of being analysed – again points to the fact that a broader selection of gestures are needed to cover the analysis of communicative functions. In particular, tilting of the head was recurrently used by the interviewee to elicit feedback from the interviewer.

CONCLUSIONS

The MUMIN coding scheme constitutes an attempt at defining a scheme for the annotation of feedback, turn management and sequencing multimodal behaviour in human communication. The preliminary results of the reliability test run in the Danish study case confirm the general reliability of the categories defined for the purpose of coding feedback and turn taking functions, although gaze, head and sequencing features seemed problematic in some cases, and not enough detailed in others (Finnish results). Body posture, which is not part of this version of the coding scheme, is a needed extension. Future revisions and extensions to the current version of the scheme will seek to accommodate these problems. We are now gathering additional experience by applying the coding scheme in graduate courses on multimodal communication.

The availability of such a scheme is an important step towards creating annotated multimodal resources for the study of multimodal communicative phenomena in different situations and different cultural settings, and for investigating many different aspects of human communication. Examples of issues that can be investigated empirically by looking at annotated data are to what extent gestural feedback co-occurs with verbal expressions; in what way different non-vocal feedback gestures can be combined; whether specific gestures are typically associated with a specific function; how multimodal feedback, turn management and sequencing strategies are expressed in different cultural settings.

REFERENCES

1. Allwood, J., Nivre, J. and Ahlsén, E. On the Semantics and Pragmatics of Linguistic Feedback. *Journal of Semantics* 9 (1992), pp. 1–26.
2. Allwood, J. Dialog Coding – Function and Grammar. Gothenburg Papers. *Theoretical Linguistics*, 85. Gothenburg University, 2001.
3. Allwood, J. and Cerrato, L. A study of gestural feedback expressions. In Paggio et al (eds.) *First Nordic Symposium on Multimodal Communication*, 2003.
4. Allwood, J., Cerrato, L., Dybkær, L., Jokinen, K., Navarretta, C. and Paggio, P. *The MUMIN multimodal coding scheme*. Technical report available at www.cst.dk/mumin/stockholmws.html, 2004.
5. Beskow J., Cerrato L., Granström B., House D., Nord-strand M., Svanfeldt G. The Swedish PF-Star Multimoda Corpora. *LREC Workshop on Models of Human Behaviour*, 2004.
6. Cassell, J. Nudge Nudge Wink Wink: Elements of Face-to-Face Conversation for Embodied

Conversational Agents, in Cassell, J. et al. (eds.), *Embodied Conversational Agents* (2000), pp. 1–27. Cambridge, MA: MIT.

7. Cerrato, L. A coding scheme for the annotation of feedback phenomena in conversational speech. *LREC Workshop on Models of Human Behaviour*, 2004.

8. Cerrato, L. Some characteristics of feedback expressions in Swedish, TMH.OPSR Vol.43 *Fonetik* (2002), pp. 101-104.

9. Clark H. and Schaefer E. Contributing to Discourse. *Cognitive Science* 13 (1989), pp. 259–94.

10. Duncan, S. *McNeill Lab Coding Methods*. Available from http://mcneilllab.uchicago.edu/topics/proc.html (last accessed 26/4/2004).

11. Ekman P. Basic emotions. In T. Dagleish and T. Power (eds.) *The Handbook of Cognition and Emotion* NY J. Wiley, 1999, pp. 45–60.

12. Kipp, M. Anvil – A Generic Annotation Tool for Multimodal Dialogue. In Eurospeech 2001. pp. 1367–1370.

13. Thórisson, Kristinn R. Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action. In B. Granström et al (eds.) *Multimodality in Language Speech Systems*. Kluwer Academic: Dordrecht, the Netherlands, pp. 173–207, 2002.

A Framework for Analyzing Embodied Communicative Feedback in Multimodal Corpora

Jens Allwood 1 & 4
jens@ling.gu.se

Karl Grammer 2 & 4
karl.grammer@univie.ac.at

Stefan Kopp 3 & 4
skopp@techfak.uni-bielefeld.de

Elisabeth Ahlsén 1 & 4
eliza@ling.gu.se

1) Department of Linguistics
Göteborg University, Box 200
SE-40530 Göteborg, Sweden

2) Ludwig Boltzmann
Institute for Urban Ethology, Althanstrasse
14, 1090 Vienna, Austria

3) Artificial Intelligence Group
Bielefeld University, P.O. 100131,
D-33501 Bielefeld, Germany

4) ZiF – Center for Interdisciplinary
Research
P.O. 100131, D-33501 Bielefeld, Germany

ABSTRACT
Communicative feedback refers to unobtrusive (usually short) vocal or bodily expressions whereby a recipient of information can inform a contributor of information about whether he/she is able and willing to communicate, perceive the information, and understand the information. This paper provides a theory for embodied communicative feedback, describing the different dimensions and features involved. It also provides a corpus analysis part, describing a first data coding and analysis method geared to find the features postulated by the theory.

Author keywords
Communicative embodied feedback, contact, perception, understanding, emotions, multimodal, embodied communication

INTRODUCTION
The purpose of this paper is to present a theoretical model of communicative feedback, which is to be used in a VR agent capable of multimodal communication. Another purpose is to briefly present the coding categories which are being used to obtain data guiding the agent's behavior. Below, we first present the theory.

The function/purpose of communication is to share information. This usually takes place by two or more communicators taking turns in contributing new information. In order to be successful, this process requires a feedback system to make sure the contributed information is really shared. Using the cybernetic notion of feedback of Wiener (1948) as a point of departure, we may define a notion of communicative feedback in terms of four functions that directly arise from basic requirements of human communication: Communicative feedback refers to unobtrusive (usually short) vocal or bodily expressions whereby a recipient of information can inform a contributor of information about whether he/she is able and willing to (i) communicate (have contact), (ii) perceive the information (perception), and (iii) understand the information (understanding). In addition, (iv) feedback information can be given about emotions and attitudes triggered by the information, a special case here being an evaluation of the main evocative function of the current and most recent contributions (cf. Allwood, Nivre & Ahlsén 1992 and Allwood 2000, where the theory is described more in detail).

The central role of feedback in communication is underpinned already by the fact that simple feedback words like *yes*, *no* and *m* are among the most frequent in spoken language. A proper analysis of their semantic/pragmatic content, however, is fairly complex and involves several different dimensions. One striking feature is that these words involve a high degree of context dependence with regard to the features of the preceding communicative act, notably the type of speech act (mood), its factual polarity, information status and evocative function (cf. Allwood, Nivre & Ahlsén 1992). Moreover, when studying natural face-to-face interaction it becomes apparent that the human feedback system comprises much more than words. Interlocutors incessantly coordinate and exchange feedback information by nonverbal means like posture, facial expression or prosody. In this paper, we extend the theoretical account developed earlier to cover *embodied communicative feedback* and provide a framework for analyzing it in multimodal corpora.

DIMENSIONS OF COMMUNICATIVE FEEDBACK

Communicative feedback can be characterized with respect to several different dimensions. Some of the most relevant in this context are the following:

(i) Degrees of control (in production of and reaction to feedback)
(ii) Degrees of awareness (in production of and reaction to feedback)
(iii) Types of expression or modality used in feedback (e.g. audible speech, visible body movements)
(iv) Types of function/content of the feedback expressions
(v) Types of reception preceding giving of feedback
(vi) Types of appraisal and evaluation occurring in listener to select feedback
(vii) Types of communicative intentionality associated with feedback by producer
(viii) Degrees of continuity in feedback signal
(ix) Semiotic information carrying relations of feedback expressions

These dimensions and others (cf. Allwood 2000) play a role in all normal human communication. Below, we will describe their role for embodied communicative feedback. Table 1 shows how different types of embodied feedback behavior can be differentiated according to these dimensions. The table is discussed and explained in the 8 following sections (cf. also Allwood 2000, for a theoretical discussion).

Degrees of awareness and control and embodiment

Human communication involves multiple levels of organization involving physical, biological, psychological and socio-cultural properties. As a basis, we assume that there are at least two (human) biological organisms in a physical environment causally influencing each other, through manipulation of their shared physical environment. Such causal influence might to some extent be innately given, so that there are probably aspects of communication that function independently of awareness and intentional control of the sender. Other types of causal influence are learned and then automatized so that they are normally functioning automatically, but potentially amenable to awareness and control. Still other forms of influence are correlated with awareness and/or intentional control, on a scale ranging from a very low to a very high degree of awareness/control. In this way, communication may involve

1) innately given causal influence
2) potentially aware and intentionally controllable causal influence
3) actually aware and intentionally controlled causal influence.

Human communication is thus "embodied" in two senses, (i) since it always relies on and exploits of physical causation, (ii) because its physical actualization occurs through processes in a biological body. The feedback system as an aspect human communication shares these general characteristics. The theory has a perspective on communication and feedback, which implies processes occurring on different levels of organization or put differently as can be seen in table 1 as implying processes that occur with different levels of awareness and control (intentionality). In addition to this, the theory also involves positing several qualitatively different parallel concurrent processes.

Perceptual modality of feedback expression

Like other kinds of human communication, the feedback system involves two primary types of expression, (i) visible body movements and (ii) audible vocal sounds. Both of these means of expression can occur on the different levels of awareness and control discussed above. That is, there is feedback which is mostly aware and intentionally controllable, like the words *yes, no, m* or the head gestures for affirmation and negation/ rejection. There is also feedback that is only potentially controllable, like smiles or emotional prosody. Finally there is feedback behavior which one is neither aware of nor able to control, but that is effective in establishing coordination between interlocutors. For example, speakers tend to coordinate the amount and energy of their body movements without being aware of it.

Types of function/content of the expressions

Communicative feedback concerns expressive behaviors that serve to give or elicit information, enabling the communicators to share information more successfully. Every expression, considered as a behavioral feedback unit, has thus two functional sides. On the one hand it can evoke reactions from the interlocutor, on the other hand it can respond to the evocative aspects of a previous contribution. Giving feedback is mainly responsive, while eliciting feedback is mainly evocative. Each feedback behavior may thereby serve different responsive functions. For example, vocal verbal signals (like *m* or *ye*s) inform the interlocutor that contact is established (C) that what has been contributed so far has been perceived (P) and (usually also) understood (U). Additionally, the word *yes* often also expresses acceptance or agreement with the main evocative function of the preceding contribution (A). Thus, four basic responsive feedback functions (C, P, U and A) can be attached to the word *yes.* In addition to these functions, further emotional, attitudinal information (E) may be expressed concurrent to the word *yes*. For example, the word may be articulated with enthusiastic prosody and a friendly smile, which would give the interlocutor further information about the recipient's emotional state. Similarly, the willingness to continue (facilitating

communication) might be expressed by posture mirroring.

Types of reception

As explained above, feedback behavior is a more or less aware and controlled expression of reactions and responses based on appraisal and evaluation of information contributed by another communicator. We think of these reactions and responses as produced in two main stages: First, an unconscious appraisal is tied to the occurrence of perception, emotions and other primary bodily reactions. If perception and emotion is connected to further processing involving meaningful connections to memory, then understanding, empathy and other cognitive attitudes, like surprise or hope, might occur. Secondly, this stage can lead to more aware appraisal, or evaluation concerning the evocative functions (C, P, U) of the preceding contribution and especially its main evocative function (A), which can be accepted, rejected or possibly met with some form of intermediary reaction, (often expressed by modal words like *perhaps*, *maybe* etc). We distinguish between these two types of reception and use the term "reactive" when the behavior is more automatic and linked to earlier stages in receptive processing, and the term "response" when the behavior is more aware and linked to later stages. For example, vocal feedback words like *yes*, *no* and *m* as well as head gestures are typically responses associated with evaluation, while posture adjustment and facial gestures are more reactive and linked more directly to appraisal and perception.

Types of appraisal and evaluation

Responses and reactions with a certain feedback function occur as a result of continuous appraisal and evaluation on the part of the communicators. We suggest that the notion of "appraisal" be used for processes that are connected to low levels of awareness and control, while "evaluation" is used when higher levels are involved. The functions C, P, U all pose requirements that can be evaluated as to whether they are met or not (positive or negative). Positive feedback in this sense can be explicitly given by the words *yes* and *m* or head nod (or implicitly by making a next contribution), and negatively by words like *no* or head shakes. The attitudinal and emotional function (E) of feedback is more complex and rests upon both appraisal, i.e. processes with a lower degree of awareness and control, as well as evaluation processes. What dimensions are relevant here is not clear. One possibility is the dimensions suggested by Scherer (1999), where it is suggested that the appraisal dimensions most relevant are (i) novelty (news value of stimulus), (ii) coping (ability to cope with a stimulus), (iii) power (how powerful does the recipient feel in relation to the stimulus), (iv) normative system (how much does the stimulus complies with norms the recipient conforms to), (v) value (to what extent does the stimulus conform to values of the recipient). The effect of appraisal that runs sequentially along these dimensions is a row of emotional reactions, which may include a certain prosody or other behavioral reactions, primarily through prosody and facial display. Additionally, there will be a cognitive evaluation of whether or not the recipient is able and/or willing to comply with the main evocative function of the preceding contribution (A), e.g., can the statements made be believed, the questions answered or the requests complied with.

Types of communicative intentionality

Like any other information communicated by verbal or bodily means, feedback information concerning the basic functions (C, P, U, A, E) can be given on many levels of awareness and intentionality. Although such levels almost certainly are a matter of degree, we, in order to simplify matters somewhat, here distinguish three levels from the point of view of the sender (cf. Allwood 1976): (i) Indicated information is information that the sender is not aware of, or intending to convey. This information is mostly communicated by virtue of the recipient's seeing it as an indexical (i.e., causal) sign. (ii) Displayed information is intended by the sender to be "showed" to the recipient. The recipient does not, however, have to recognize this intention. (iii) Signaled information is intended by the sender to "show" the recipient that he is displaying and, thus, intends the recipient to recognize it as displayed. Display and signaling of information can be achieved through any of the three main semiotic types of signs (indices, icons and symbols, cf. Peirce 1955/1931). In particular, we will regard ordinary linguistic expressions (verbal symbols) as being signals by convention. Thus, a linguistic expression like *It's raining*, when used conventionally, is intended to evoke the receiver's

	Bodily coordination	**Facial expression, posture, prosody**	**Head gestures**	**Vocal verbal**
Awareness and control	Innate, automatic	Innate, potentially aware + controlled	Potentially/mostly aware + controlled	Potentially/mostly aware + controlled
Expression	Visible	Visible, audible	Visible	Audible
	C, P, E	C, P, E	C, P, U, E, A	C, P, U, E, A
Type of reception	Reactive	Reactive	Response	Response
Type of appraisal	Appraisal, evaluation	Appraisal, evaluation	Appraisal, evaluation	Appraisal, evaluation
Intentionality	Indicate	Indicate, display	Signal	Signal
Continuity	Analogue	Analogue, digital	Digital	Digital
Semiotic sign type	Index	Index, icon	Symbol	Symbol

Table 1. Types of linguistic and other communicative expressions of feedback.

(C = Contact, P = Perception, U = Understanding, E = Emotion, A = Attitude)

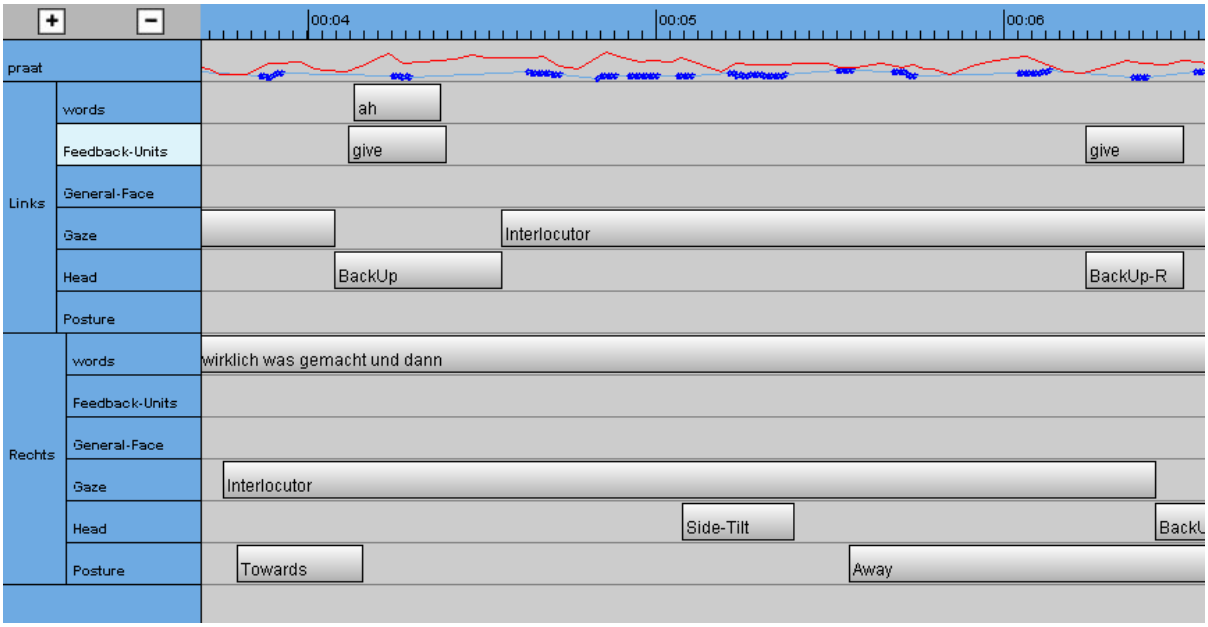


Figure 1. Snapshot of the annotation board for analyzing embodied communicative feedback.

recognition not merely that "it's raining" but that he/she is "being shown that it's raining".

Degree of continuity (i.e. analog vs. digital)

Feedback information can be expressed in analog ways, such as prosodic patterns in speech, continuous body movements and facial expressions, which evolve over stretches of interaction. It may also be more digital and discrete, such as feedback words, word repetitions or head nods and shakes. Normally, analog and digital expressions are used in combination.

Type of semiotic information carrying relation

Following Peirce's semiotic taxonomy, where indices are based on contiguity, icons on similarity and symbols on conventional, arbitrary relations between the sign and the signified, we can find different types of semiotic information expressed by feedback.

FALSIFICATION AND EMPIRICAL CONTENT

A relevant question to ask in relation to all theories is the question of how the theory could be falsified. Since the aspect of the theory that has been presented in this paper mainly consist of a taxonomy of the theoretical dimensions of the theory, falsification in this case consists in showing that the taxonomy is ill-founded, i.e. that it is not homogeneous, that the categories are not mutually exclusive, not perspicuous, not economical or not fruitful. Since the question of whether the above criteria are met or not can be meaningfully asked, we conclude that the theory has empirical content, ie can be falsified.

EMPIRICAL BASIS

To test our theoretical framework for its adequacy and usability in analyzing multimodal corpora, we have started to gather and analyze data on 30 video-recorded dyadic interactions with two subjects in standing position. The dyads were systematically varied with respect to sex and mutual acquaintance. The subjects were university students and their task was to find out as much as possible about each other within 3 minutes. Extractions of one minute from the video-recordings were transcribed and coded, according to an abbreviated version of the MUMIN coding scheme for feedback (Allwood et al. 2005). The coding schema identifies the feedback units (either verbal or non-verbal), which are coded for function type (giving, eliciting) and attitudes (continued contact, perception, understanding; acceptance of main evocative function; emotional attitudes). It further captures the following non-vocal behaviors: posture shifts, facial expressions, gaze, and head movements. In addition, intensity and pitch of the (single) audio track were computed using the PRAAT software; movement analysis was applied to measure how the interlocutors' movements vary and coordinate over time. Finally, subjects were asked to fill in a questionnaire about their socio-cognitive perception of the other (e.g. rapport). Fig. 1 shows a snapshort of the annotation board during a data coding session.

CONCLUSIONS

We have presented a theory for communicative feedback, describing the different dimensions involved. This theory is supposed to provide the basis of a framework for analyzing embodied feedback behavior in natural interactions. We have started to design a coding scheme and a data analysis method suited to capture those features that are decisive in this account (such as type of expression, relevant function, or time scale). Currently, we are investigating how the resultant multimodal corpus can be analyzed for patterns and rules as required for a predictive model of embodied feedback. Ultimately, such a model should afford its simulation and testing in a state-of-the-art embodied conversational character.

ACKNOWLEDGEMENTS

We thank the Ludwig Bolzmann Institute for Urban Ethology in Vienna for help with data collection and transcription.

REFERENCES

1. Allwood, J. (1976). Linguistic Communication as Action and Cooperation. *Gothenburg Monographs in Linguistics* 2. Göteborg University, Department of Linguistics.
2. Allwood, J. (2000). Structure of Dialog. In Taylor, M., Bouwhuis, D. & Neel, F. (eds.) *The Structure of Multimodal Dialogue* II, Amsterdam, Benjamins.pp. 3 - 24.
3. Allwood, J., Cerrato, L., Dybjær, L., Jokinen, K., Navaretta, C. & Paggio, P. (2005). The MUMIN Multimodal Coding Scheme. *NorFA Yearbook* 2005.
4. Allwood, J, NivreJ, & Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback, *Journal of Semantics*, vol. 9, no. 1, 1-26.
5. Peirce, C. S. (1931). *Collected Papers of Charles Sanders Peirce*, 1931-1958, 8 vols. Edited by Charles Hartshorne, Paul Weiss, and Arthur Burks. Cambridge, Mass., Harvard University Press.
6. Scherer, K. T- (1999). Appraisal Theory. In T. Dalgleish & M. J. Power (Eds.) *Handbook of Emoiton and Cognition* (pp.637-663). Chichester: New York.
7. Wiener, N. (1948). *Cybernetics or Control and Communication in the Animal and the Machine*. BIT Press.

A Study into Multimodal Behaviour in Error Correction

Marie-Luce Bourguet
Computer Science Dept.
Queen Mary, University of London
Mile End Road, E1 4NS, London, UK
mlb@dcs.qmul.ac.uk

ABSTRACT
Recognition based interaction techniques (e.g. speech and gesture recognition) are still error prone. Research is needed to understand users' multimodal behaviour when faced with recognition errors. In this paper we present an experimental study into users' synchronisation of speech and pen inputs in error correction. The results of the study suggest that users are likely to modify their synchronisation patterns in the belief that it can help error resolution. Further investigation is now required and multimodal corpora are needed where a range of error resolution situations have been clearly annotated.

Author Keywords
Recognition-based interaction, error-handling, multimodal behaviour, synchronisation patterns.

ACM Classification Keywords
H5.2.User Interfaces: Input devices and strategies (e.g., mouse, touch screen).

INTRODUCTION
Natural modalities of interaction, such as speech and hand gestures, rely on recognition-based technologies, which are inherently error prone. Speech recognition systems, for example, are sensitive to vocabulary size, quality of audio signals and variability of voice parameters. Signal and noise separation also remains a major challenge in speech recognition technology, as current systems are extremely sensitive to background noise and to the presence of more than one speaker. Speech recognition errors include *false rejection*, which is when the user has spoken correctly, but the system cannot recognize the spoken input and does not deliver any recognition result; and *misrecognition*, which is when the recognizer returns a result with words that are different from what the user spoke. In both cases, the possible causes of the error include: some words spoken by the user are not in the application's vocabulary; the spoken sentence does not match the application's grammar; the system is not ready to listen; there are similar sounding words in the application's vocabulary; the user pauses too long between words; the user produces disfluent speech; the user's voice is too different from stored voice models; the computer's audio is not properly configured or the microphone is not properly adjusted.

Previous research has uncovered typical user strategies to handle errors in recognition-based multimodal interfaces. Studies of speech interfaces have found that the most instinctive way for users to correct mistakes is to repeat [3]. In handwriting, a similar strategy is to overwrite a misrecognised word. Linguistic adaptation is another strategy that has been observed where users choose to rephrase their speech, in the belief that it can influence error resolution: a word may be substituted for another, or a simpler syntactic structure may be chosen [6]. In multimodal systems, it has been suggested that users are willing to repeat their input at least once, after which they will tend to switch to another modality. For example, if speech input failed repeatedly when entering data in a form, users may switch to the keyboard in order to type their entry [6]. Alternative strategies include locating a recognition error by touching a misrecognised word on a writing-sensitive screen where recognition output is displayed, then correcting the error by choosing from a list of alternative words, typing, handwriting, or editing using gestures drawn on the display [7].

In speech interfaces, one of the most natural user correction strategies consists in repeating the misrecognised input. However, although repeating might be the most obvious way to correct when the system mishears, it is often the worse for the system [7]. The main reason for this is that when repeating, users tend to adjust their way of speaking (e.g. by over-articulating) to what they believe is easier for the recogniser to interpret, which often has the opposite effect. The purpose of the experiment presented in this paper is to study if users exhibit similar strategies of modifying some aspects of their input when repeating a complex multimodal command (e.g. a command that combines speech and a pen gesture), in the belief that it can help error resolution. More precisely, we are interested in comparing users' modality synchronisation patterns in normal situations of interaction, and in situations of error correction.

Synchronisation patterns are important to devise accurate and efficient modality integration techniques. For example, synchronisation patterns will allow a multimodal system to differentiate between independent inputs ("concurrent multimodality") and synergistic inputs ("compound multimodality"), and to set appropriate timeouts between subsequent synergistic inputs. Previous work in the area of multimodal user behaviour has uncovered typical patterns of natural integration and synchronisation of input modes [1][5]. In [5] for example, integration patterns for speech and pen inputs are described, in which pen onset usually precedes speech onset in sequential constructions (when one input is completed before the onset of the other) as well as in simultaneous constructions (when there is a temporal overlap between inputs in different modalities). In [1], it is shown that 3D hand pointing gestures are usually synchronised with either the nominal or deictic expression of a phrase. It is also shown that the timing of such gestures is predictable in the [-200 ms, 400 ms] interval around the beginning of their related expressions. However, as far as we are aware, no previous work has been reported on modality integration patterns in situations of error correction. It is thus important to understand if and how synchronisation patterns may change in situations of error correction, when users repeat a complex multimodal command.

EMPIRICAL STUDY

Task

We implemented a multimodal speech and pen application in which the participants in the experiment were asked to place famous London landmarks on a map. They had to reproduce a map similar to a model provided (see Figure 1), where images of specific sizes had to be positioned in specific places. On the model, the names of the landmarks (e.g. "Tower Bridge") were written on each picture and the participants were instructed to use these names when speaking to the system. The task included: making the images appear on the empty map, resizing the images, and positioning the images as precisely as possible on the map. This task was chosen because it includes both verbal (naming the landmarks) and spatial (positioning the images) elements, for which multimodal interaction is highly appropriate.

To accomplish the task, four different interaction styles were implemented:

- *speech-gesture* commands (e.g. to make a picture appear on the map, the user draws a "P" and pronounces the name of the landmark);

- *speech-point* commands (e.g. to resize a picture, the user points at a picture and says "smaller");

- *speech- point- point* commands (e.g. to swap two images, the user says "swap" and points at two different images);

- *speech-draw* commands (e.g. to move an image, the user says "move it here" and draws a line from the image to its desired location).

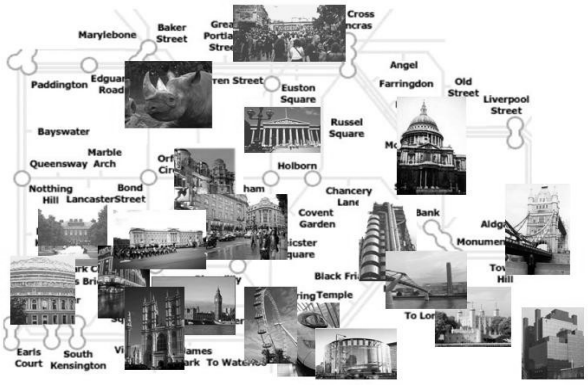


Figure 1. The London map application.

The vocabulary of the London map application is simple and does not allow linguistic adaptation (i.e. rephrasing). It comprises approximately thirty entries, including the names of the landmarks and command words or expressions such as "smaller", "move it here", etc.

Although the set of multimodal commands and vocabulary is fixed, inputs can be entered in any order (e.g. pen first, speech first or pen and speech overlapping).

The application was built using the multimodal toolkit described in [2]. Speech recognition was carried out using the ViaVoice™ engine and gesture recognition performed via the Satin toolbox [4] on a Wacom LCD tablet.

Method

Eight paid subjects, with no or little experience with recognition based systems, participated in the experiment. All of them received a brief demonstration of the application during which the experimenter was careful to show a range of different synchronisation patterns in order not to influence the execution of the multimodal commands. The participants also had the opportunity to undertake a short training session to familiarise themselves with the task and the different interaction styles.

No visual feedback was provided on the activity and performance of the speech and gesture recognisers. This means that, in case of interaction problems, users could only realise that a recognition error had occurred when evaluating the response of the system. For example, if nothing appears on the screen after a "draw image" command, users can only assume that either a speech or a gesture recognition error has occurred. Not being able to precisely determine where it went wrong was believed to be an incentive to repeat the entire command.

The experiment would only end when the entire task had been completed, i.e. when the produced map was similar to the model provided, with all the landmarks of the correct size and roughly positioned in the right place. This provided another necessary incentive to try and correct recognition errors.

Data Collection

During the experiments, automatic logs were set up to record various data: timing of every pen-down and pen-up event, speech onset and offset, and speech and gesture recognition results. The experiments were also videotaped. The log files were then compared with the video recordings in order to identify the situations of recognition error recovery. We were only interested in users' strategies to cope with errors made by the recognition systems, so ungrammatical or out of vocabulary user inputs, i.e. inputs where the participants, as opposed to the recognition systems, had made an error were discarded. Only well-formed user inputs were kept, (i.e. inputs complying with one of the four interaction styles and making use of the correct vocabulary).

Four types of recognition errors were observed: speech false rejection, speech misrecognition, gesture false rejection, and gesture misrecognition. For speech-gesture commands, combinations of speech and gesture recognition errors were also possible. To illustrate the different recognition errors, let us imagine that the user said "Tower Bridge" while drawing a "P" gesture on the screen. The possible recognition errors are:

- User's speech has not been recognised (speech false rejection): nothing appears on the screen.

- User's speech has been misrecognised (speech misrecognition). If the system recognised the name of another landmark, an unexpected image appears on the screen.

- User's gesture has not been recognised (gesture false rejection): nothing appears on the screen.

- User's gesture has been misrecognised (gesture misrecognition). If the system recognised a "delete" gesture and the gesture was drawn on an image, the image unexpectedly disappears. If the gesture was drawn on an empty space of the screen, nothing happens.

User inputs were then classified into one of the two following categories:

- *New commands*: when a command is entered in normal situation of interaction;

- *Recovery commands*: when a command is repeated, in response to a recognition error. If the user corrects an unexpected result (such as deleting an unexpected image) before repeating the initial command, the repeated command is not considered a recovery command, but a new command.

Results and Discussion

A total of 1073 multimodal commands were collected, of which 279 were entered in situations of error recovery. Figure 2 summarises the most commonly observed synchronisation patterns for the four interaction styles. In each case, the following information is shown: (1) total number of commands observed; (2) average pattern (the top line represents speech and the bottom line pen; the lines are proportional to event durations); and (3) proportions of the two most frequently observed patterns. For example, 133 speech-gesture commands were collected in a normal situation (new commands). For these commands, the average pattern is characterised by pen onset first, followed by speech. Speech onset occurs approximately in the middle of the gesture execution and finishes after the gesture has been completed. 79% of speech-gesture new commands conform to this typical pattern. 19% of speech-gesture new commands conform to a different pattern where speech onset precedes pen onset and where the gesture finishes before the end of the speech.

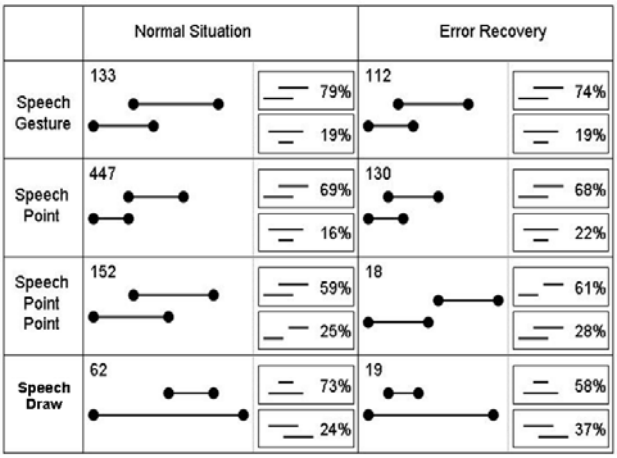


Figure 2. Synchronisation patterns.

At first glance, it can be seen that across the data (independently of the interaction style and of the command category), pen onset tends to precede speech onset. This result corroborates the main finding reported in [5] and is valid across the different interaction styles, in both normal and error recovery modes.

For the speech-gesture commands, the integration patterns in normal and recovery modes are similar. In recovery mode, the participants tend to repeat their commands in exactly the same manner as in the normal mode of interaction. This could be explained by the fact that, when a command is unsuccessful, both modalities of interaction, speech and gesture, can be held responsible for the error. Users cannot easily determine which recogniser has made

an error because the multimodal UI does not provide any feedback on the activity of the recognisers and on the recognition results they return. In these circumstances, users do not seem to be able to devise a strategy for error recovery. In order to verify this hypothesis, a further experiment should be conducted where users are provided with adequate feedback on the recognition processes.

However, in the speech-point, speech-point-point, and speech-draw cases, where speech is the only input that is subject to recognition errors, some differences can be observed between the new commands and the recovery commands.

For the speech-point and speech-draw commands, the data suggest that speech onset is shifted towards the beginning of the pen stroke in situations of error recovery. For speech-draw commands, a significant proportion of error recovery inputs (37% compared to 24% for normal inputs) shows in fact a precedence of speech over pen. In this case, it seems that users tend to deal first with the error-prone input (speech).

For the speech-point-point commands, the proportions of the two frequently observed patterns are reversed. In recovery mode, it seems that users are more likely to have completed their pen inputs before speech onset. This change of behaviour may be attributed to the complexity of the interaction style. In recovery mode, users tend to avoid multi-tasking by adopting sequential patterns of integration, where there is no temporal overlap between inputs in different modes.

CONCLUSION

Small and mobile computing devices, in the form of personal digital assistants (PDAs), mobile phones, and wearable computers, have become common, but compared with desktop computers, their screens are small or non existent, and their small keyboards are hard to use when on the move. On these platforms, typical multimodal error handling strategies such as linguistic adaptation, modality switch, and lists of alternative words, may not be available, leaving repetition as the only viable strategy. The study presented in this paper has shown that, when repeating a multimodal command, and given that the source of the error can be identified, different modality synchronisation patterns are likely to enter in users' strategies for influencing the performance of recognition-based modalities. Consequently, multimodal interfaces should not constrain users to use modalities in any fixed order. Synchronisation patterns that significantly depart from typical patterns should be interpreted with in view the possibility that the user is in error recovery mode, and modality integration techniques should be able to adapt to changing synchronisation patterns.

More empirical research is still needed to gain a thorough understanding of multimodal behaviour in error correction. This will necessitate the collection of multimodal corpora in a variety of error situations, including:

- Different error recognition types: misrecognition, false recognition, and misfire (when the recognition system returns a result in the absence of user speech).
- User mistakes: out of vocabulary input, ungrammatical input, and disfluent speech.
- And type of feedback provided to users on the different recognition processes.

ACKNOWLEDGMENTS

This research was supported by the Nuffield Foundation under grant NUF-NAL 00. Many thanks to Sarah Talbot for running the experimental study and to all the participants.

REFERENCES

1. Bourguet, M.L. and Ando, A. Synchronisation of Speech and Hand Gestures during Multimodal Human-Computer Interaction. *Ext. Abstracts CHI 1998*, ACM Press (1998), 241-242.

2. Bourguet, M.L. A Toolkit for Creating and Testing Multimodal Interface Designs. In *Companion Proc. UIST 2002*, 29-30.

3. Halverson, C., Horn, D., Karat, C. and Karat, J. The Beauty of Errors: Patterns of Error Correction in Desktop Speech Systems. In *Proc. INTERACT 99*, IOS Press (1999), 133-140.

4. Hong, J. and Landay, J. Satin: A Toolkit for Informal Ink-Based Applications. In *Proc. UIST 00*, ACM Press (2000), 63-72.

5. Oviatt, S., DeAngeli, A. and Kuhn, K. Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction. In *Proc. CHI 1997*, ACM Press (1997), 415-422.

6. Oviatt, S. Taming recognition errors with a multimodal interface. *Communications of the ACM 43*, 9 (2000), 45-51.

7. Suhm, B., Myers, B. and Waibel, A. Multimodal Error Correction for Speech User Interfaces. *ACM Transactions on Computer_Human Interaction 8*, 1 (2001), 60-98.

Making a Case for Spatial Prompting in Human-Robot Communication

Anders Green
KTH School of Computer Science
100 44 STOCKHOLM
green@csc.kth.se

Helge Hüttenrauch
KTH School of Computer Science
100 44 STOCKHOLM
hehu@csc.kth.se

ABSTRACT
In this paper we present an analysis of a set of examples of how verbal and non-verbal behavior of a service robot influence users' way of positioning themselves during interaction, using concepts from theories of non-verbal behavior. Based on the analysis we propose a design case where a robot utilizes a (naïve) *spatial prompting* strategy to influence the spatial positioning and communicative behavior of the user.

INTRODUCTION
A design requirement of a personal service robot is that it should be configured and provided with work tasks by the user in an interactive and intuitive way. These robots are intended to provide service tasks in the home, possibly offering wide range of services. Typically they are envisioned to be equipped with multimodal spoken dialogue systems, to reduce the complexity in the user interface.

In this paper we argue that theories of spatial positioning need to be considered when developing the communicative system of the robot. Furthermore we present an empirical account of the way spatial behavior of robots influence human users. We also propose the term *spatial prompting*, which refers to active strategies of the robot that are intended to influence users to position themselves in a way that is advantageous for further communicative actions.

Positioning, as it has been approached as a research challenge for human-robot interaction, is considered as providing adaptive *physical* movements of the robot. A result of this is that the *communicative dimension* of positioning typically has been ignored in systems that interactively position themselves in relation to their users. One requirement that is typically put forward is that the robot should position itself in a *socially appropriate manner* [1, 4]. The parameters that concern these approaches are typically derived from research on non-verbal behavior.

In robotics the problem of maintaining the robot localized and situated within a geometric representation of the world has been framed as the Simultaneous Localization and Mapping (SLAM) problem [13]. Recent advances in Human-Robot Interaction (HRI) have raised the interest in detecting and tracking the position of users during interaction. When the position of the user is known, the robot can plan how to position itself [1].

The research on spatial reasoning applied to robotics is well advanced but primarily focused on natural language understanding of spatial relations, providing for exchanges concerning locations of objects in the environment [3].

RESEARCH ON SPATIALITY IN COMMUNICATION
There are several research approaches for human-human that are relevant for spatial management between humans and robots. Hall [14] studied interpersonal distances and distinguished four different distances: *intimate (*0-1.5 ft), personal *(*1.5-4 ft), *social* 4-12 ft, and *public* (> 12 ft). These distances vary both with respect to the current activity and cultural factors. Another dimension that is relevant to spatiality is the concept of *territoriality,* according to Sack, i.e., "the attempt by an individual or group to affect, influence, or control people, phenomena, and relationships, by delimiting and asserting control over a geographic area" [15].

Kendon [12] also studied the spatial configuration of the participants, using the term F-formations, for instance the *L-shape* which describes the relation when two participants have a common visual focus. The shared space, the so called *o-space,* or the *transactional space* is then located in front of the participants, and it is within this area that the interaction is conducted. Clark [5] refers to this space as the *workspace*, where perceptual co-presence is established between speakers [5, 10]. In this context, research on perception and especially visual perception plays an important part for maintaining common ground between participants [10, 8]. Gill [9] has investigated the communicative effects that participants achieve by using nonverbal behavior, focusing on the functional rather than the morphological perspective of nonverbal behavior. One such function is the category *focus* which is a meta-discursive function that signals a shift in the center of attention in the discussion, e.g., a shift in body posture with the same meaning as the utterance "I am going to focus on this spot".

Another, less obvious, but nevertheless important concept is Schegloff's notion of *body torque* [15], a state of the bodily configuration when two different body segments are oriented in different directions. According to Schegloff [15]

Body torque "project change", i.e., when some part of the body is organized in an unstable way, the participants may predict that a change in posture is pending. For instance, when turning the head, this might predict a change of the general body orientation. During interaction, speakers monitor the action of others, interpreting purposeful actions that lead towards a common joint goal as compliance [10]. Human-robot interaction is situated in a physical context, where understanding and reference to actions of the human partner during interaction explicitly needs be taken into account. This makes research on *virtual* collaborative [6] environments interesting also in this context, since it is concerned with models that explicitly represent spatiality and reference.

CORPUS ANALYSIS OF SPATIAL MANAGEMENT

We have analyzed a video corpus, collected in a European project [7], containing transcribed data of about 20 user sessions, (approximately 20 minutes each) where a user talks to a robot and teaches it the names and locations of objects using a combination of gestures and speech.

By viewing the video corpus we identified and analyzed instances where the robot movements or verbal actions appear to influence the actions of the user. The examples reflect three different ways in which the robot actively influences the user to act:

- Primary verbal: by using a spoken command
- Primary non-verbal: by movements
- Multi-modal: using movement as trigger for a verbally specified (or grounded action)

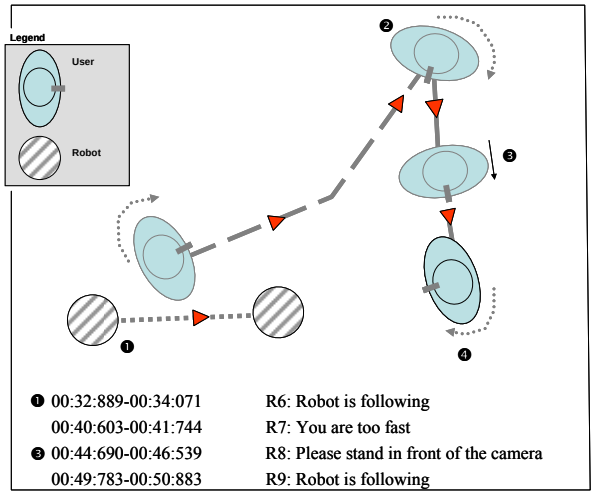


Figure 1: The user and the robot in a follow sequence. The dotted lines show the trajectory of the robot and user.

Primary verbal influence (Example 1)

The example in Figure 1 describes how the events unfold as the user has commanded the robot to follow, after acknowledging this in (❶) the robot starts moving. The robot follows and the user follow the paths depicted in Figure 1. In the second phase of the follow sequence the user has moved to a position that the robot considers too far away. Then the robot says "You are too fast", which triggers the behavior in (❷): the user turns towards what can be characterized as the robot's transactional space (according to Kendon [12]) or workspace (according to Clark [5]). Then the user starts moving slightly towards the robot workspace. When the robot speaks "Please stand in front of the camera", the user quickly moves in front of the robot (❸-❹), something which may be seen as a Give-turn Body Move (according to Gill [9]) that may be seen as a display of the users willingness to interact [11]. In other words, the user displays her attention towards the robot. The visual attention is aimed at the robot once the user has turned around (throughout phases ❷-❹).

The first example (Figure 1) illustrates how verbal action directives influence the physical actions of the user. There are instances of this type of example in each of the 20 sessions that the corpus covers.

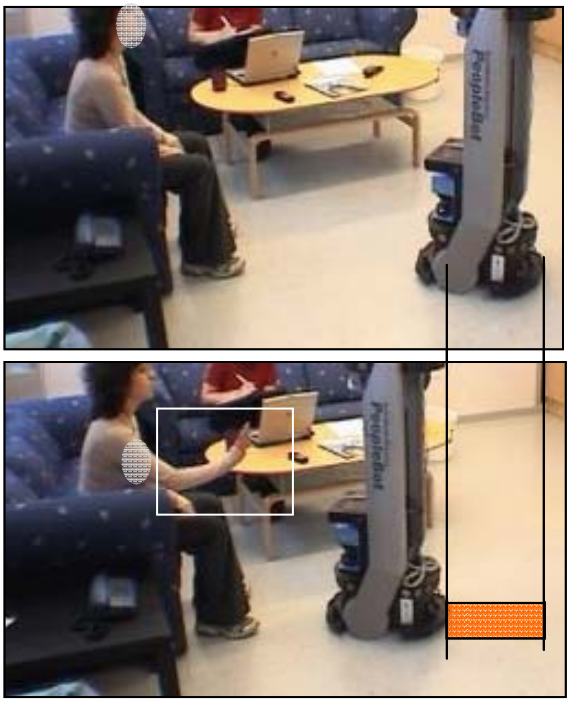


Figure 2: An example of a non-verbal action triggering user action, i.e., a stop gesture (the square in the lower image). The field in the second image, with lines pointing to the first image, illustrates the movement of the robot platform before the gesture is displayed.

Primary non-verbal influence (Example 2)

An example of how the (non-verbal) movements of the robot platform can trigger actions of the user is depicted in Figure 2. In this case the action triggered is a gesture, but it could be another action. In the preceding sequence (not shown) the user has commanded the robot to follow (by saying "Follow me"). Then the user sits down and waits as the robot is approaching as seen in Figure 2 (upper image). During the approach the user raises the arm and displays a "Stop" gesture. It appears as if the robot comes too close;

perhaps crosses the border between a *social* to an *intimate* distance, in terms of Hall [14] or triggers a behavioral reaction as the robot breaches a territorial border upheld by the user [15].

On the other hand we might interpret the raising of the hand in a "Stop" gesture as an indication to the robot that this is an advantageous position for the task at hand, i.e., the "Stop" gesture is displayed as part of a joint goal (according to [10]).

U: ok

U: now go to telephone

R: Going to telephone ❶

Figure 3: An example of how communicative actions and spatial configurations of the robot are interrelated in different modalities.

Multi-modal influence (Example 3)

In the moment that passes before the example depicted in Figure 4 the user has acknowledged that the robot has completed the task of finding an object (by establishing a common reference to the object located in front of the robot). Then the user stoops and looks into the camera of the robot, while uttering the command: "now go to the telephone". Then, when the robot confirms the request by saying "Going to telephone", the user changes into an upright position ❶ (Figure 3).

Eye contact is maintained during the whole sequence. Our analysis of this is that the user is attempting to require (visual) attention on the part of the robot. We suggest that the moving camera of the robot provides a biomimetic display that makes the user assume a transactional space located in front of the robot. In terms of Body torque, the stooping is a temporary disconfiguration of a posture, and the change back to the more neutral posture in ❶ (Figure 3) is a return to what Schegloff calls a home position [15]. In our understanding, the torque, i.e., the stooping posture and the user's attempt to require contact can be contributed to the spatial influence of the transactional space [12] and the displayed "eye" gaze of the camera.

A SPATIAL PROMPTING STRATEGY

In the analysis of the scenario we have found examples that suggest that the robot platform may influence the spatial behavior, i.e., posture, gaze direction and gesture displays (e.g., stop gesture). Typically a (multimodal) natural language user interface that is used for human-robot interaction is concerned with the aspects of spatiality that are encoded in language, such as referencing using spatial relations (e.g. "behind", "beside", "in front of X" etc) and deictic gestures [e.g. 3].

A system that encompassed a model for spatial influence could provide *spatial prompts* aimed to influence the spatial positioning of the user, for instance, to ensure an optimal configuration for further communicative behavior. An example of such a design is depicted in Figure 4. In terms of a dialogue system design, we can frame the display of Stop gesture in the second example (Figure 2) as a system goal, i.e., to reach a state in interaction where a Stop gesture has been displayed to provide an end-point in the robot's approach to a position. This requires an internal representation that considers the movement of the robot platform (and the user) together with the task state. In the follow episode in Example 2 the overall task is Follow, but

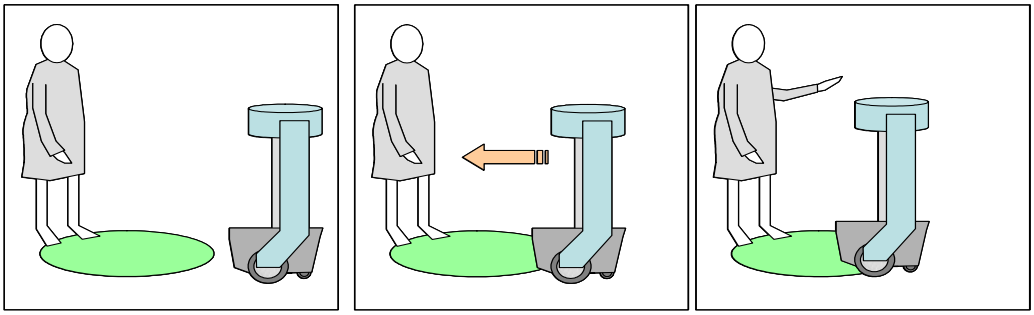


Figure 4: In the first image of the sequence, the robot is commanded to follow the user. As the user stops and the robot approaches the transactional space the robot increases its speed for a short moment. The speed increase is a spatial prompt intended to trigger a response by the user. A safety feature would stop the robot if no action is detected by the user.

as the user sits down the robot can be said to Approach the user. This could be taken into account in the system, for instance by *increasing* the speed of the robot, slightly in order to prompt the display of a gesture that shows the robot where to stop. It is obvious that a robot behavior like this raises some concerns in terms of acceptability [4], but the point of this example is merely to illustrate that spatial prompting is a possible strategy to actively influence the behavior of the users. We could also imagine an example when the robot prompts the user to tell when the robot is close enough, e.g., by saying: "say stop" while slowly approaching.

Examples like these show how we can turn empirical observations into design proposals. In terms of Body moves [9], the increase in speed in the example in Figure 4, would put Focus on the transactional space and an obligation on the user to react and specify where the robot should be positioned – before the safety feature of the robot stops it at a default distance. Then a prediction could be made, e.g., based on corpus data, so that the robot could provide a contribution that is relevant to the predicted task, e.g., "Show me an object", instead of "Stopped following" (as it is obvious to the user).

CONCLUSIONS

We have discussed a set of examples from our corpus of human-robot interactions arguing that verbal and non-verbal behavior of the robot actively influence users' spatial configuration during interaction. We also provide a scenario where a robot could utilize a spatial prompting strategy. In the future we aim to identify ways of spatially influence users by further analyze corpus data and validate the design proposal by implementing spatial prompting strategies to be tested on a robot platform.

ACKNOWLEDGEMENTS

The work described in this paper was conducted within the EU Integrated Project COGNIRON ('The Cognitive Robot Companion' - www.cogniron.org) and was funded by the European Commission Division FP6-IST Future and Emerging Technologies under Contract FP6-002020

REFERENCES

1. Alami, R., Clodic, A., Montreuil, V., Sisbot, E. A., and Chatila, R. (2005). Task planning for human-robot interaction. In Proceedings of the 2005 Joint Conference on Smart Objects and Ambient intelligence: innovative Context-Aware Services: Usages and Technologies Grenoble, France.
2. Pacchierotti, E., Christensen, H., and Jensfelt, P., (2005). Embodied social interaction in hallway settings: a user study. In IEEE Workshop on Robot and Human Interactive Communication Ro-Man2006, Nashville, USA, pp. 164-171.
3. Skubic, M., Perzanowski, D., Blisard, S., Schultz, A., Adams, W., Bugajska, M. and Brock, D. (2004). Spatial language for human-robot dialogs, IEEE Transactions on Systems, Man and Cybernetics, Part C,, vol.34, no.2pp. 154- 167.
4. Walters, M.L., Dautenhahn, K., te Boekhorst, R., Koay K. L., Kaouri C., Woods, S., Nehaniv, C., Lee, D. and Werry, I. (2005) The Influence of Subjects' Personality Traits on Personal Spatial Zones in a Human-Robot Interaction Experiment. Proc. IEEE Ro-man 2005, pp. 347-352
5. Clark, H. H. & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. Journal of Memory and Language , 50(1), 62-81.
6. Gergle, D., Kraut, R. E., and Fussell, S. R. (2004). Action as language in a shared visual space. In *Proceedings of the 2004 ACM Conference on Computer Cooperative Work* (Chicago, Illinois, USA, November 06 - 10, 2004). CSCW '04. ACM Press, New York, NY, 487-496.
7. Green, A., Hüttenrauch, H., Severinson Eklundh, K. (2004). Applying the Wizard-of-Oz Framework to Cooperative Service Discovery and Configuration In Proceedings of IEEE Ro-Man, September 20-22, 2004, Kurashiki, Japan.
8. Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. Neuroscience & Biobehavioral Reviews, 24 (6), 581–604.
9. Gill, S.P., Kawamori, M., Katagiri, Y., Shimojima, A. (2000) The Role of Body Moves in Dialogue. In RASK, April 2000, pp. 89-114.
10. Clark, H. H. (1996). Using language. Cambridge: Cambridge University Press.
11. Allwood, J., Nivre, J., and Ahlsén E. (1992). On the Semantics and Pragmatics of Linguistic Feedback, Journal of Semantics 1992 9(1):1-26.
12. Kendon, Adam. (1990). Conducting interaction – Patterns of behavior in focused encounters. Studies in interactional sociolinguistics. Cambridge, NY, USA: Press syndicate of the University of Cambridge.
13. Smith, R., Self, M, and Cheeseman P (1986) Estimating uncertain spatial relationships in robotics. UAI 1986: 435-461
14. Hall, E.T. (1966). The Hidden Dimension: Man's Use of Space in Public and Private. The Bodley Head Ltd, London, UK.
15. Sack, R.D. 1(986). Human Territoriality: Its Theory and History. Cambridge: Cambridge University Press.
16. Schegloff, E.A., (1998). Body Torque. Social Research, 65:3, 535-596.

The MIMUS Corpus

Pilar Manchón Portillo
Universidad de Sevilla
pmanchon@us.es

Carmen del Solar Valdés
Universidad de Sevilla
carsolval@alum.us.es

Gabriel de Amores Carredano
Universidad de Sevilla
jgabriel@us.es

Guillermo Pérez García
Universidad de Sevilla
gperez@us.es

ABSTRACT
This paper describes the motivation, collection and format of the MIMUS corpus. MIMUS (**M**ult**IM**odal, **U**niversity of **S**eville) is the result of multimodal WOZ experiments conducted at the University of Seville (**USEV**) as part of the TALK project. The main objective of the MIMUS corpus has been to gather information about different users and their performance, preferences and usage of a multimodal multilingual natural dialogue system in the Smart Home scenario. This corpus is focused on (although not restricted to) wheel-chair-bound users, since they are especially motivated to use this kind of technology, and they may have specific needs.

Author Keywords
Multimodal corpus, HCI, Multimodal Experiments

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
The paper is organized as follows. First, the WOZ platform will be briefly described. The next section describes the experiments and their motivation. Then, the data collected and the format and tools used will be discussed. Finally, some conclusions will be drawn.

THE USEV WOZ PLATFORM
The platform is based on Delfos, the original spoken dialogue system developed at the University of Seville (**USEV**). Since the objective of this corpus is to obtain relevant information to design, implement and configure the next multimodal version of Delfos, all the previous spoken functionality was made available as well as the new multimodal additions.

In terms of hardware, the platform consists of a PC used by the wizard, a tablet PC used by the subject, a Wifi router by means of which both PCs can communicate, and a set of real home devices which make up the Smart Home set up. In addition, software consisting of a set of wizard agents and subject agents has been developed.

The former set consists of:

1. A Wizard Helper, which is a control panel that enables the wizard to talk to the user and remotely play audio and video files.
2. A Device Manager, which enables the user to control the physical home devices and to see what the subject is clicking on, if that were the case.

The set of subject agents consists of:

1. A home Setup agent, which displays the virtual house and its devices and where the subject may click using a pen or mouse.
2. A telephone simulator, where the subject can simulate a phone call and other regular telephone options.
3. A TTS Manager, which synthesizes the wizard´s messages when appropriate.
4. A Log Manager, where all the interaction data is logged, and
5. A Video Client, used to simulate an outside camera.

Figure 1. The subject´s touchscreen display

THE USEV MULTIMODAL WOZ EXPERIMENTS

Motivation and Objectives
The MIMUS corpus is the result of a multimodal WoZ set of experiments. The original objective of these experiments was to collect data in order to extend and configure an existing spoken dialogue system (Delfos) by adding new input and output modalities. The goal was to identify and gather information regarding:

- any possible obstacles or difficulties to communicate

- any biases that prevent naturality
- a corpus of natural language in the home domain
- preferred modality in relation to task
- preferred modality in relation to task and scenario **(*)**
- preferred output modality in relation to information type
- modality preference in relation to system familiarity
- task completion time
- combination of modalities for one particular task
- inter-modality timing
- user evolution, learnability and change in attitude
- new modalities impact on interaction in other modalities
- context relevance and interpretation in multimodal environments
- pro-activity and response thresholds in multimodal environments
- relevance of scenario-specific factors/needs
- multimodal multitasking: multimodal input fusion and ambiguity resolution

(*) Different scenarios may render different results for the same task

The experiments investigate users' speech and pen multimodal integration patterns on a system application that controls the lights, a blind, a radio, a heater, an alarm, the main door, a security camera, and a telephone. The interactions between users and the human wizard were recorded from different perspectives.

Subjects

Two groups of informants, all of whom can be described as completely naïve subjects, were recruited. A primary group is formed by a number of wheel-chair bound subjects (16); a secondary group includes subjects without disabilities (7). Informants' ages range between 19 and 54 years old. At the moment, there is a total of 7 women and 16 men. All of them are native speakers of Spanish, and show varying levels of computer expertise.

Experiments

The experiments took place in a lab especially prepared to simulate a smart house, where all the devices to be controlled were at sight. Subjects were alone and undisturbed during the experiments. The set consisted of 2 complementary experiments were the subjects were interacting with an expert wizard, and 1 experiment where the naïve subjects became naïve wizards.

Instructions were provided for all tasks, which ranged from simple actions (turning a light on) to more complex simultaneous actions (making a phone call while monitoring the camera and opening the door).

Additional information on the subjects (computer expertise, etc.) was collected prior to the experiments in recorded interviews; more information about their perception of the interaction with the system and the system performance was collected after the experiments in forms.

Figure 2. Naïve subject during the experiments in the lab

User-Wizard interactions

The interaction between subject and system was recorded from different perspectives. A digital camera recorded the progression of the experiment. A web camera captured the subjects' face as they performed tasks (Figure 2). The touch-screen activity was logged.

Logging

This is the information recorded in execution time during the experiment. The logging is therefore focused on low-level information, and especially on the time at which each event occurs.

The information automatically logged can be summarized as:

- Modality
- Clicks
- Time of events
- Wizard Messages
- Wizard Message Time

The format chosen to record the information is EMMA, W3C working draft on Extensible MultiModal Annotation markup language, which distinguishes two properties for the annotation of input modality: (1) indicating the broader medium or channel (medium) and (2) indicating the specific mode of communication used on that channel (mode). The input medium is defined from the users' perspective and indicates whether they use their voice (acoustic), touch (tactile), or visual appearance/motion (visual) as input.

DATA ANNOTATION

Annotation

Given the different types of analyses to be performed on these data, different levels of annotations have been established. These levels can be summarized as follows:

- Personal information and user profile
- Experiment conditions and procedure
- Tasks and Subtasks

- Automatic Logging
- Dialogue
- Gestures

Personal Information and User Profile
This is the information related to the users, and their computer skills, disabilities, age, gender, cultural level, degree of familiarity with speech and/or graphical interfaces, nationality and language proficiency.

It also includes the information collected on pre-experimental and post-experimental surveys regarding the users´ biases towards automated interfaces and their opinions, suggestions or satisfaction level after interacting with the system.

Experiments Conditions and Procedure
This is the type of information that defines the conditions under which the experiments were conducted and the procedures followed to ensure the data reliability and coherence.

Time of the day at which the experiments were conducted, duration, general instructions given to the subjects, incidents or mistakes, if any, are the main parameters to be taken into account at this level.

Tasks and Subtasks
The description of the tasks and subtasks to be performed by the subjects is recorded at this level. In this case it is also relevant to record the exact way in which the subjects were given the information to perform each task, and when and how such information was provided. This is particularly important since some of these tasks and subtasks were especially designed to encourage or at least allow subjects to perform several tasks simultaneously. It is also important in order to determine the cognitive load imposed on the subjects.

Automatic Logging
This includes all the information logged automatically during the experiments. This is mainly low level information (time stamps, modality, icons clicked on, etc). It also includes all the information predetermined and/or introduced manually by the wizard (predetermined or spontaneous wizard messages, etc).

Dialogue
This level includes transcription and segmentation of the user´s utterances as well as the Dialogue Move and Subdialogue annotation.

In MIMUS, dialogue-level annotations follow the classification of the Natural Command Language Dialogues (NCLDs), as defined in [2]. Since it is the broader concept of NCL that encapsulates the present framework of analysis, it seems natural to also employ Dialogue Moves (DMs) in annotating dialogue turns.

Another reason for choosing the NCL approach over Traum's [8] Conversation Acts is that the former focuses on the internal aspects of dialogue, whereas the latter builds up "*to a level of common ground that is necessary for communication of beliefs, intentions, and obligations*" [8]. That is, a model built on the grounds of a NCL should be based more on what is said than what is in the minds of the participants when things are said. In other words, it should try to model external aspects of the dialogue rather than the participants' internal state.

The Dialogue Moves are therefore classified as follows:

- Command-Oriented Dialogue Moves

  askCommand: The system requests the user to specify a command or function to be performed.

  specifyCommand: A specific command or function is selected.

  informExecution: The system acknowledges the execution of the task.

- Parameter-Oriented Dialogue Moves

  askParameter: The system asks for the value of a specific parameter.

  specifyParameter: The assignment of some value to one parameter.

- Interaction-Oriented Dialogue Moves

  askConfirmation: Once a command has been completed, some situations will require an explicit and/or implicit confirmation.

  answer YN: The user replies yes/no.

  askContinuation: The system asks for the continuation of the dialogue.

  askRepeat: Any of the participants may request the other to repeat the last utterance, or even a specific parameter or command.

  askHelp: A petition for help (general, a specific command, or a specific parameter).

  answerHelp: The reply to an askHelp move.

  errorRecovery: For a situation in which the continuation of the dialogue is impossible.

  greet: The usual greeting operation.

  quit: The usual closing operation.

As to Subdialogue Annotation, and as [2] state, "*An important aspect of NLCDs is that they exhibit functional embeddings . . . [that] occur when the goal of a sub-dialogue shifts to another dialogue type.*", the following types of sub-dialogues are distinguished within a NCLD:

1. Deliberation dialogue
2. Action-oriented dialogue

3. Information-seeking dialogue
4. Negotiative dialogue

Gestures

The Gestures classification and annotation is defined according to a closed set of values for the attribute "gestureType". These have been adapted from the SmartKom Project collection of multimodal data [7]:

1. anger/irritation
2. pondering/reflecting
3. joy/gratification (being successful)
4. surprise
5. helplessness
6. neutral/anything else
7. face partly not visible

TOOLS

ANVIL is the annotation tool used for the transcription process and the encoding of the elements recorded during the experiments. The resulting data will also be translated into the TALK NXT format.

The ANVIL track "UserInput.spoken" will include the manual segmentation and transcription mentioned above. The track "UserInput.graphical" will be generated automatically from the information logged (in execution time) in the XML file "gui_in.xml". Also automatically loaded are the tracks "GUIOutput.spoken" (from the log "speech_out.xml") and "GUIOutput.graphical" (from the log "gui_out.xml").

The resulting ANVIL tracks are listed below:

1. **Track 1**: Waveform
2. **Track 2**: WizardActions
3. **Track 3**: UserInput.graphical (user´s clicks)
4. **Track 4**: UserInput.spoken (manual segmentation and transcription of user's speech)
5. **Track 5**: GUIOutput.graphical (graphical output)
6. **Track 6**: GUIOutput.spoken (endpointed TTS speech)
7. **Track 7**: DialogueMoves (hand-annotated)
8. **Track 8**: Subdialogues (hand-annotated)
9. **Track 9**: FacialExpressions (hand-annotated)

CONCLUSION

The MIMUS corpus is the result of a conscientious design work, and a rigorous and methodic predefined procedure to conduct the experiments as well as to log and annotate all relevant information. This is therefore a reliable and growing source of information for research on HCI and Multimodal Dialogue Systems as well as other related disciplines, which so far consists of 73 dialogues, by 23 different users in 32 different tasks total. It will be available for research purposes at the conclusion of the current European project.

ACKNOWLEDGMENTS

The work described in this paper has been partially funded by EU Project Talk (Contract No 507802) and the Spanish Ministry of Science and Technology under Project TIC2002-00526.

REFERENCES

1. Manchón P., Pérez G. & Amores G., WOZ experiments in Multimodal Dialogue Systems. Proceedings of the ninth workshop on the semantics and pragmatics of dialogue, 131-135. Nancy, France. June, 2005.
2. Amores, J. Gabriel, & Quesada J. Francisco. "Cooperation and Collaboration in Natural Command Language Dialogues." In Johan Bos, and Mary Ellen, and Colin Matheson (eds.), *Proceedings of the sixth workshop on the semantics and pragmatics of dialogue (EDILOG),* September 4--6, 2002.
3. Amores, J. Gabriel, & Quesada.J. Francisco "Dialogue Moves for Natural Command Languages." In *Procesamiento del Lenguaje Natural,* 27: 89-96, 2001.
4. Amores, J. Gabriel, & Quesada. J. Francisco "Dialogue Moves in Natural Command Languages." In *SIRIDUS Project*, D. 1.1, September, 2000.
5. Becker, Tilman, et al. "Proposed Methods for Multimodal Experiments." In *TALK Project*, D.6.1, November 2004.
6. Johnston, Michael. "EMMA: Extensible MultiModal Annotation markup language." W3C Working Draft, September 2005. http://www.w3.org/TR/emma
7. Steininger, Silke, Florian Schiel, and Angelika Glesner. "User-State Labeling Procedures For The Multimodal Data Collection Of SmartKom." In *SmartKom Project*, Report 28, October 2002.
8. Traum, David, and Tim Allen. "A Speech-Acts Approach to Grounding in Conversation." In *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP-92)*, pages 137-40, October 1994.