The Workshop Programme

9:00 - 9:10 Welcome

9:10 - 10:30 SESSION: HEAD MOVEMENTS / TOOLS (4 talks)

- *Feedback in Head Gestures and Speech* (Patrizia Paggio and Costanza Navarretta)
- *Repeated head movements, their function and relation to speech* (Max Boholm and Jens Allwood)
- *Using head movement to detect listener responses during multi-party dialogue* (Stuart Battersby and Patrick Healey)
- *Developing heterogeneous corpora using the Digital Replay System (DRS)* (Dawn Knight, Paul Tennent, Svenja Adolphs and Ronald Carter)

10:30 - 11:00 Coffee break

11:00 - 13:00 SESSION: INTERACTION, EMOTION & CULTURE (6 talks)

- *A Multimodal Corpus for Studying Dominance in Small Group Conversations* (Oya Aran, Hayley Hung and Daniel Gatica-Perez)
- *D64: A Corpus of Richly Recorded Conversational Interaction* (Catharine Oertel, Fred Cummins, Nick Campbell, Jens Edlund and Petra Wagner)
- *DYNEMO: A Corpus of dynamic and spontaneous emotional facial expressions* (Brigitte Meillon, Anna Tcherkassof, Nadine Mandran, Jean Michel Adam, Michel Dubois, Damien Dupré, Anne Marie Benoît, Anne Guérin-Dugué and Alice Caplier)
- *A Filipino Multimodal Emotion Database* (Jocelynn Cu, Merlin Suarez and Madelene Sta. Maria)
- *Automatic Segmentation of Spontaneous Data using Dimensional Labels from Multiple Coders* (Mihalis Nicolaou, Hatice Gunes and Maja Pantic)
- *The UTEP-ICT Cross-Cultural Multiparty Multimodal Dialog Corpus* (David Herrera, David Novick, Dusan Jan and David Traum)

13:00 - 14:30 Lunch break

14:30 - 15:50 SESSION: MOTION CAPTURE AND VISION TECHNOLOGIES (4 talks)

- *The USC CreativeIT Database: A Multimodal Database of Theatrical Improvisation* (Angeliki Metallinou, Chi-Chun Lee, Carlos Busso, Sharon Carnicke and Shrikanth Narayanan)
- *Exploiting Motion Capture for Virtual Human Animation* (Alexis Heloir, Michael Neff and Michael Kipp)
- *Linking Conversation Analysis and Motion Capturing: How to robustly track multiple participants?* (Karola Pitsch, Bernhard Brüning, Christian Schnier, Holger Dierker and Sven Wachsmuth)
- *3D Vision Technology for Capturing Multimodal Corpora: Chances and Challenges* (Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise and Luc Van Gool)

15:50 - 17:20 SESSION: POSTERS & DEMOS (includes coffee 16:00-16:30)
(see full list of posters and demos below)

17:20 - 18:40 SESSION: GESTURE ANNOTATION & ANALYSIS (4 talks)

- *Multi-level Annotations of Nonverbal Behaviors in French Spontaneous Conversation* (Ning Tan, Gaëlle Ferré, Marion Tellier, Edlira Cela, Mary-Annick Morel, Jean-Claude Martin and Philippe Blache)
- *A multimodal corpus for gesture expressivity analysis* (George Caridakis, Johannes Wagner, Amaryllis Raouzaiou, Zoran Curto, Elisabeth Andre and Kostas Karpouzis)
- *Timing Relationships between Speech and Co-Verbal Gestures in Spontaneous French* (Gaëlle Ferré)
- *The Bielefeld Speech and Gesture Alignment Corpus (SaGA)* (Andy Lücking, Kirsten Bergmann, Florian Hahn, Stefan Kopp and Hannes Rieser)

18:40 - 19:00 Final discussion

LIST OF POSTERS

- *A Multimodal Corpus Recorded in a Health Smart Home* (Anthony Fleury, Michel Vacher, François Portet, Pedro Chahuara and Norbert Noury)
- *A multi-software integration platform and support for multimedia transcripts of language* (Christophe Parisse and Aliyah Morgenstern)
- *The LDOS-PerAff-1 Corpus of Face Video Clips with Affective and Personality Metadata* (Marko Tkalčič, Jurij Tasič and Andrej Košir)
- *Corpus-based analysis of users' emotional strategies to convince virtual characters* (Magalie Ochs and Rui Prada)
- *SALEM (Statistical AnaLysis of Elan files in Matlab)* (Marc Hanheide, Manja Lohse and Angelika Dierker)
- *Collecting and Annotating Conversational Eye-Gaze Data* (Kristiina Jokinen, Seiichi Yamamoto and Masafumi Nishida)
- *SensHome: Towards A Corpus for Everyday Activities in Smart Homes* (Jochen Frey, Robert Neßelrath, Christian H. Schulz and Jan Alexandersson)
- *Look at me!: An emotion learning reinforcement tool for children with severe motor disability* (Beatriz López-Mencia, David Pardo, Néna Roa-Seiler, Álvaro Hernández-Trapote and Luis A. Hernández)
- *Capturing multimodal interaction at medical meetings in a hospital setting; Opportunities and Challenges* (Bridget Kane, Saturnino Luz and Jing Su)
- *The role of the signal redundancy in the emergence of sign language prosody* (Svetlana Dachkovsky)

LIST OF DEMOS

- *Automatic face tracking in Anvil* (Bart Jongejan)
- *Insight Interaction. A multimodal and multifocal dialogue corpus* (Geert Brône, Bert Oben and Kurt Feyaerts)
- *Capturing massively multimodal dialogues: affordable synchronization and visualization* (Jens Edlund and Jonas Beskow)
- *Generating and annotating corpora of multimedia telecommunications of pediatric cancer patients and their families and friends* (Thomas Bliesener)

Workshop Organiser(s)

Michael Kipp, DFKI, Germany
Jean-Claude Martin, LIMSI-CNRS, France
Patrizia Paggio, Univ. of Copenhagen, Denmark
Dirk Heylen, Univ. of Twente, The Netherlands

Workshop Programme Committee

Elisabeth Ahlsén, Univ. of Göteborg, Sweden
Jan Alexandersson, DFKI, Germany
Jens Allwood, Univ. of Göteborg, Sweden
Gerard Bailly, GIPSA-lab, Univ. de Grenoble, France
Philippe Blache, Univ. de Provence, France
Stéphanie Buisine, Arts et Métiers ParisTech, France
Susanne Burger, Carnegie Mellon Univ., USA
Nick Campbell, Trinity College Dublin, Ireland
Loredana Cerrato, Acapela Group, Sweden
Piero Cosi, ISTC CNR, Italy
Marco Gillies, Goldsmiths, Univ. of London, UK
Sylvie Gibet, Univ. de Bretagne Sud (UBS), France
Joakim Gustafson, KTH, Sweden
Kristiina Jokinen, Univ. of Helsinki, Finland
Kostas Karpouzis, NTUA, Greece
Daniel Loehr, MITRE, USA
Maurizio Mancini, Univ. of Genova, Italy
Costanza Navarretta, Univ. of Copenhagen, Denmark
Michael Neff, UC Davis, USA
Fabio Pianesi, Bruno Kessler Foundation, Italy
Isabella Poggi, Univ. of Rome Three, Italy
Andrei Popescu-Belis, Idiap Research Inst., Switzerland
Matthias Rehm, Aalborg Univ., Denmark
Florian Schiel, LMU Munich, Germany
Thomas Schmidt, Univ. of Hamburg, Germany
Ielka van der Sluis, Trinity College Dublin, Ireland
Peter Wittenburg, MPI for Psycholinguistics, The Netherlands

Table of Contents

Author Index

Feedback in Head Gestures and Speech

Patrizia Paggio, Costanza Navarretta

University of Copenhagen
Centre for Language Technology (CST)
Njalsgade 140, 2300-DK Copenhagen
paggio@hum.ku.dk, costanza@hum.ku.dk

Abstract

This paper addresses the issue of how linguistic feedback expressions and head gestures relate to each other. The data are a collection of eight video-recorded Danish map-task dialogues annotated with phonetic and prosodic information. We focus on the analysis of the listeners' head movements and face expressions and on how they relate to feedback expressions in their different prosodic realisations. The study shows that head gestures and prosodic features significantly improve automatic classification of dialogue act labels for these expressions.

1. Introduction

The fact that head movements and speech reinforce each other in conversation is generally acknowledged. In fact, several studies indicate that the relation between the two communication modalities is more than general physiological co-ordination.

Several authors have noted that head movements and posture shifts tend to occur in conjunction with turn shifts or syntactic clause boundaries (Dittmann and Llewellyn, 1968; Kendon, 1972; Hadar et al., 1984). Moreover, aspects of head movement seem to correlate with different communicative functions. The fact that head movements relate to feedback (or backchanneling) has been observed in several studies and for different languages (Yngve, 1970; Duncan, 1972; Maynard, 1987; Cerrato, 2007). Others (Hadar et al., 1985; McClave, 2000) mention feedback, but also agreement, turn management, aligning with the interlocutor's stressed syllables and pauses.

Given the fact that several scholars converge on considering head movements as relevant to feedback phenomena, in this paper we look at Danish multimodal data and systematically analyse the relation between *head gestures* (i.e. head movements and facial expressions) and specific feedback words and phrases in their different prosodic realisations.

Other studies have looked at gestures in annotated multimodal corpora. For example, in Jokinen and Ragni (2007) and Jokinen et al. (2008) it was found that machine learning algorithms could be trained to recognise some of the functions of head movements. And Reidsma et al. (2009) show that there is a dependence between focus of attention (a combination of head, gaze and body features) and the assignment of dialogue act labels. In this study, we investigate whether the discourse function of linguistic feedback expressions can be automatically classified based on prosodic and gestural information. We find that prosodic information improves the classification results, and that head gestures, where they occur, contribute to the semantic interpretation of feedback expressions in a significant way.

In Section (2) we describe the multimodal Danish corpus on which we have based our investigations. In Section (3), we explain how the prosody of feedback expressions is annotated, how their content is coded in terms of dialogue act labels, and how head gestures are annotated. Section (4) contains a description of the resulting data and a discussion of the results obtained in the various classification experiments. Section (5) is the conclusion.

2. The Danish map-task dialogues

The Danish map-task dialogues from the DanPASS corpus (Grønnum, 2006) are a collection of dialogues in which 11 speaker pairs cooperate on a map-task. The dialogue participants are seated in different rooms and cannot see each other. They talk through headsets, and one of them is recorded with a video camera. Each pair goes through four different sets of maps, and changes roles each time, with one subject giving instructions and the other following them. The resulting material is transcribed orthographically with an indication of stress, articulatory hesitations and pauses. In addition to this, the acoustic signals are segmented into words, syllables and prosodic phrases, and annotated with POS-tags, phonological and phonetic transcriptions, pitch and intonation contours. All the annotations were done using Praat (Boersma and Weenink, 2009).

We decided to use the DanPASS dialogues in our study for a number of reasons. First of all, we expected to find a large number of feedback expressions due to the nature of the task. Since the dialogue participants' maps differ slightly, negotiating the route is not entirely easy, and the two have to give each other confirmation and feedback quite often. A preliminary investigation of the speech annotations in the DanPASS dialogues revealed in fact a rather high frequency of feedback words and expressions. Feedback expressions make up a rather heterogeneous group. Here, we focus on what we will call *Yes* and *No* expressions, i.e. in Danish words like *ja* (yes), *jo* (yes in a negative context), *jamen* (yes but), *nej* (no), *næh* (no), cf. Table (1).

Yes and *No* feedback expressions represent about 7% of the approximately 47,000 running words in the corpus. This is a rather high proportion compared to other corpora. In the spontaneous spoken corpus of Danish BYSOC (Gregersen and Pedersen, 1991), for example, *Yes* and *No* expressions occur 72.028 times, i.e. 0.05 % of the over 1,3 million words in the corpus. In the Brown corpus of written Amer-

Expression	Count	%
Stressed *Yes*	3020	85
Unstressed *Yes*	173	5
Stressed *No*	282	8
Unstressed *No*	57	2
Total	3532	100

Table 1: *Yes* and *No* expressions in the DanPASS dialogues

ican English, the word *yes* only occurs 144 times, which makes for 0.01% of the corpus.

The *Yes* and *No* tokens are sometimes a single word and sometimes a sequence of words. However, we have only considered word sequences if the speech annotators have grouped them in the transcription based on phonetic considerations. Examples are: *ja altså* (yes actually), *ja ja* (yes yes), *ja okay* (yes okay), *jo aldrig* (yes never).

Having video-recordings of one of the participants allows for an investigation of how head movements are used in this context. Moreover, the detailed phonetic and prosodic annotations that are provided together with the corpus, also make it possible to include these dimensions into our investigation.

A disadvantage of the DanPASS material is the technical setup, especially the fact that the two subjects cannot see each other, a circumstance that makes their non-verbal behaviour likely to be quite limited. Nevertheless, participants do move their heads in ways that are clearly related to their speech behaviour. Communicative head gestures are in fact easily distinguished from movements done to follow the route on the map. Feedback behaviour, both in speech and gestures, can be observed especially in the person who is receiving the instructions, who will be called the *follower*. This is in accordance to what found in other studies, where it is especially the listener that gives feedback through head movements (Duncan, 1972; Maynard, 1987)

Therefore, we decided to focus our analysis only on the follower's part of the interaction. For time reasons, we also restricted the study to four different subject pairs and two interactions per pair. This left us in the end with eight videos for a total of about an hour of video-recorded dialogue interaction. Adding semantic annotations to the feedback expressions and gesture annotations took four annotators about two work weeks altogether including discussions and corrections.

3. Feedback annotation

3.1. Feedback expressions and their prosody

As already mentioned, all words in DanPASS are marked with whether they are stressed or not, and with articulatory hesitation at word onset or offset where appropriate. In addition, the speech annotators have also marked additional stress prominence where they found it, as well as whether stress is followed by a higher, a lower or an even posttonal pitch. See Grønnum (2006) for a more detailed explanation. We will refer to these details of prosodic annotation as tone attributes.

In the subset of the corpus considered here, 82% of the feedback expressions bear stress or tone information, and 12% are unstressed; 7% of them are marked with onset or offset hesitation, or both.

3.2. Semantic annotation of feedback expressions

All feedback expressions were annotated for this study with an agreement feature where relevant. Moreover, to distinguish among the various functions that feedback expressions have in the dialogues, we selected a subset of the categories defined in the emerging ISO 24617-2 standard for semantic annotation of language resources. Finally, the two turn management categories *TurnTake* and *TurnElicit* were also used. All the attributes are shown in Table (2).

Dialogue act	
Answer	Answer to a question, whether direct or indirect.
Accept	Acceptance of a request or an offer, i.e. when the speaker asks the follower to take a certain route on the map, and the follower accepts.
Decline	Decline of a request or an offer, i.e. when the follower denies to take a certain route on the map.
RepeatRephrase	Repetition or rephrasal of own expression or expression by the interlocutor.
Agreement	
Agree	Agreement to a certain state of affairs.
NonAgree	Non-agreement to a certain state of affairs.
Turn Management	
TurnTake	Signals willingness to take the turn.
TurnElicit	Encourages the interlocutor to take the turn.

Table 2: Semantic annotation of feedback expressions

It should be noted that the same expression may be annotated with a label for each of the three semantic dimensions. For example, a *yes* can be the *Answer* to a question like "You have a lake on your map, don't you?". In the same context, it will be coded with *Agree* since it expresses agreement to the knowledge presupposed in the question, and it may possibly also have a *TurnElicit* function if intonation and face expression indicate that more input is being asked for. In many cases, however, only one or two of the semantic dimensions will be relevant. Table (3) shows how the various types are distributed across the 453 feedback expressions in our data.

Dialogue Act		
Answer	70	15%
RepeatRephrase	57	13%
Accept	127	28%
None	199	44%
Agreement		
Agree	166	37%
NonAgree	14	3%
None	273	60%
Turn Management		
TurnTake	113	3%
TurnElicit	85	19%
None	354	78%

Table 3: Distribution of semantic categories

3.3. Gestures

All communicative head gestures occurring in the videos were found and annotated with ANVIL (Kipp, 2004), us-

ing a subset of the attributes defined in the MUMIN annotation scheme (Allwood et al., 2007; Paggio and Navarretta, 2009). The MUMIN scheme is a general framework for the study of gestures in interpersonal communication. The attributes defined in the scheme concern facial expressions, head movements, body posture and hand gestures, and they relate to the annotation of gesture shape and dynamics as well as their semiotic and functional interpretation. In this study, we do not deal with functional classification of the gestures in themselves, but rather with how gestures contribute to the semantic interpretations of linguistic expressions. Therefore, only a few of the MUMIN attributes have been used. They are shown in Table (4).

Attribute	Values
Face	Smile, Laughter, Scowl, FaceOther
HeadMovement	Nod, Jerk, Tilt, SideTurn, Shake, Waggle, Other
HeadRepetition	Single, Repeated

Table 4: Face expression and head movement annotation

The intercoder agreement for these categories was measured in a previous study (Jokinen et al., 2008), and found to be quite high (in the range 83-96%).

If a gesture was judged by the annotator to complement one or more words, a link was also established between the gesture under consideration and the relevant speech sequence. The link was then used to extract gesture information together with the relevant linguistic annotations on which to apply machine learning.

The total number of head gestures annotated is 236. Of these, however, only 56 (21%) co-occur with feedback expressions, with Nod as by far the most frequent type (20 occurrences) followed by FaceOther as the second most frequent (12). The other tokens are distributed more or less evenly with a few occurrences (1-6) per type. We have not yet analysed the many remaining gestures to see what other types of utterances they associate with since our focus here is feedback expressions.

4. Analysis of the data

The multimodal data we obtained by combining the linguistic annotations from DanPASS with the gesture annotation created in ANVIL, comprise all *Yes* and *No* expressions not accompanied by gestures, and all those that are accompanied by a face expression or a head movement, as shown in Table (5).

Expression	Count	%
Yes without gestures	407	90
No without gestures	46	10
Total without gestures	453	100
Yes with gestures	51	91
No with gestures	5	9
Total with gestures	56	100

Table 5: *Yes* and *No* datasets

In the experiments reported here we focus on the *Yes* expressions because they make up the largest material, thus providing a more reliable basis for machine learning.

The two datasets of *Yes* expressions were thus used for automatic dialogue act classification using the Weka system (Witten and Frank, 2005). The first was used to investigate the effect of prosody on the classification, and the second the combined effect of prosody and head gestures. Overall, the classifier which gave the best results is Weka's SMO, which is an implementation of a support vector classifier, so we only report results obtained with this classifier in addition to the ZeroR baseline, which alwasy selects the most frequent class. Ten-folds cross-validation was used in all experiments.

In the first group of experiments, then, we only took into consideration the datasets without gestures (407 *Yes*). We started by totally leaving out prosodic information, and running the classification based on Agreement, Turn and, of course, Dialogue Act features. In the second run we added information about stress (stressed or unstressed); in the third we added tone attributes, and in the fourth information on hesitation.

In Table (6) results are provided in terms of precision (P), recall (R) and F-measure (F). These are calculated in Weka as weighted averages of the results obtained for each class.

dataset	Algor	P	R	F
Yes	ZeroR	19.8	44.5	27.4
	SMO	49.8	47.7	40.6
+stress	SMO	48.5	46.9	40.9
+stress+tone	SMO	56.2	56.5	49.1
+stress+tone+hesitation	SMO	52.4	57.2	49.8

Table 6: Classification results with prosodic features

The results indicate that the annotation significantly improves the classification of dialogue acts with respect to the baseline in all four experiments (with an improvement of 13.2, 13.5, 21.7 and 22.4% respectively). In particular, annotation improves each time a prosodic feature is added. The largest improvement is obtained when tone information is added, while the improvement in accuracy when hesitations are introduced is not significant. In Table (7) we show the confusion matrix from the last experiment in which stress, tone and hesitation are added to the data.

a	b	c	d	classified as
0	0	20	27	a = Repeat-Rephrase
0	2	16	36	b = Answer
0	1	106	18	c = Accept
0	1	54	126	d = NONE

Table 7: Confusion matrix for data with prosodic features

From the confusion matrix we can see that the classifier is best at identifying *Accept*, and is very bad at identifying *RepeatRephrase*. This must be seen in light of the fact that the former type is very frequent in the data. As for the latter, prosodic information does not correlate with it in any systematic way. Information on what word precedes would

probably be the most valuable feature, especially in case of repetition.

The second group of experiments was conducted on the dataset where feedback expressions are accompanied by gestures (51 *Yes*). The Precision, Recall and F-measure of the ZeroR classifier on these data are 13.9, 37.3 and 20.2, respectively. However, for these experiments we used as a baseline the results obtained by the SMO algorithm on the basis of *Yes* expressions and all prosodic information, the combination that gave the best results on the larger dataset. Together with the prosodic information, Agreement, Turn and Dialogue Act attributes were included as before. We then run a number of trials adding face expressions alone first, then head movements alone, and finally both gesture types together. The results are shown in Table (8).

dataset	Algor	P	R	F
Yes+prosody	SMO	27.7	35.3	29.3
+face	SMO	35.5	43.1	37.4
+headm	SMO	38.9	39.2	38
+face+headm	SMO	49	47.1	44.9

Table 8: Classification results with prosody and head gesture features

The results indicate that adding head gesture information on top of the prosodic features improves the classification of dialogue acts in this reduced dataset. The largest improvement is achieved when both face expression and head movement attributes are taken into consideration. In Table (9) we show the confusion matrix generated when using prosodic features alone and in Table (10) we show the confusion matrix generated when head gestures are included.

a	b	c	d	classified as
0	1	1	4	a = Repeat-Rephrase
0	7	5	7	b = NONE
0	7	2	3	c = Answer
0	0	0	14	d = Accept

Table 9: Confusion matrix for data with prosodic features alone

a	b	c	d	classified as
2	0	2	2	a = Repeat-Rephrase
0	13	0	6	b = NONE
2	4	2	4	c = Answer
1	3	0	10	d = Accept

Table 10: Confusion matrix for data with prosodic features and head gestures

The comparison of the two matrices shows that the classification improves for all types of dialogue acts when gesture information is used, with the exception of *Accept* which is now classified incorrectly in 4 out of 14 occurrences.

If we look at how different gesture types are weighted by the classifier to distinguish between the classes, we can see that head movements, in particular nods and jerks, are associated with *Answer* rather than *Accept*, and with *RepeatRephrase* more strongly than either *Answer* or *Accept*. We can also see that an *Answer* is often accompanied by a smile when there is no head nod: in other words, smiles and nods are somewhat complementary. The results are interesting because they confirm, even for the very specific setup provided by map-task dialogues, the claim made in several of the studies quoted earlier, that head nods often express feedback.

5. Conclusion

In this study we have experimented with the automatic classification of positive feedback expressions into different dialogue acts in a multimodal corpus of Danish. We have conducted two sets of experiments, first looking at how prosodic features contribute to the classification, then testing whether the use of head gesture information improved the accuracy of the classifier. The results indicate that prosodic features improve the classification, and that where feedback expressions are accompanied by head gestures, gesture information is also useful. More than three quarters of the head gestures in our data co-occur with other linguistic utterances than those targeted in this study. Extending our investigation to those and including negative feedback expressions, as we plan to do, will provide us with a larger dataset and therefore presumably with even more interesting and reliable results.

6. Acknowledgements

This research has been supported by the Danish Council for Independent Research in the Humanities and by the NOMCO project (http://sskkii.gu.se/nomco/), a collaborative Nordic project with participating research groups at the universities of Gothenburg, Copenhagen and Helsinki. The project is funded by the NOS-HS NORDCORP programme. We would also like to thank Nina Grønnum for allowing us to use the DanPASS corpus, and our gesture annotators Josephine Bødker Arrild and Sara Andersen.

7. References

Allwood, Jens, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The mumin coding scheme for the annotation of feedback, turn management and sequencing. *Multimodal Corpora for Modelling Human Multimodal Behaviour. Special Issue of the International Journal of Language Resources and Evaluation*, 41(3–4):273–287.

Boersma, Paul and David Weenink, 2009. *Praat: doing phonetics by computer*. Retrieved May 1, 2009, from http://www.praat.org/.

Cerrato, Loredana. 2007. *Investigating Communicative Feedback Phenomena across Languages and Modalities*. Ph.D. thesis, Stockholm, KTH, Speech and Music Communication.

Dittmann, Allen and Lynn Llewellyn. 1968. Relationship between vocalizations and head nods as listener responses. *Journal of Personality and Social Psychology*, 9.

Duncan, Starkey. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292.

Gregersen, Frans and Inge Lise Pedersen, editors. 1991. *The Copenhagen study in urban sociolinguistics*. Reitzel.

Grønnum, Nina. 2006. DanPASS - a Danish phonetically annotated spontaneous speech corpus. In Nicoletta Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, and D. Tapias, editors, *Proceedings of the 5th LREC*, pages 1578–1583, Genoa, May.

Hadar, Uri, T. J. Steiner, E. C. Grant, and F. Clifford Rose. 1984. The timing of shifts of head postures during conservation. *Human Movement Science*, 3(3):237–245, September.

Hadar, Uri, T. J. Steiner, and F. Clifford Rose. 1985. Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4):214–228, December.

Jokinen, Kristiina and A. Ragni. 2007. Clustering experiments on the communicative prop- erties of gaze and gestures. In *Proceeding of the 3rd. Baltic Conference on Human Language Technologies*, Kaunas, October.

Jokinen, Kristiina, Costanza Navarretta, and Patrizia Paggio. 2008. Distinguishing the communicative functions of gestures. In *Proceedings of the 5th MLMI*, LNCS 5237, pages 38–49, Utrecht, September. Springer.

Kendon, Alan. 1972. Some relationships between body motion and speech. In A. Seigman and B. Pope, editors, *Studies in Dyadic Communication*, pages 177–216. Elmsford, New York: Pergamon Press.

Kipp, Michael. 2004. *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Ph.D. thesis, Saarland University, Saarbruecken, Germany, Boca Raton, Florida, dissertation.com.

Maynard, Senko. 1987. Interactional functions of a nonverbal sign: Head movement in japanese dyadic casual conversation. *Journal of Pragmatics*, 11:589–606.

McClave, Evelyn. 2000. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32:855–878.

Paggio, Patrizia and Costanza Navarretta. 2009. Integration and representation issues in the annotation of multimodal data. In *Proceedings of the NODALIDA 2009 workshop Multimodal Communication – from Human Behaviour to Computational Models*, pages 25–31. NEALT, May.

Reidsma, Dennis, Dirk Heylen, and Rieks op den Akker. 2009. On the Contextual Analysis of Agreement Scores. In Michael Kipp, Jean-Claude Martin, Patrizia Paggio, and Dirk Heylen, editors, *Multimodal Corpora From Models of Natural Interaction to Systems and Applications*, number 5509 in LNAI, pages 122–137. Springer.

Witten, Ian H. and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, second edition.

Yngve, Victor. 1970. On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society*, pages 567–578.

Repeated head movements, their function and relation to speech

Max Boholm, Jens Allwood
SCCIILL (SSKKII) Center
University of Gothenburg, Box 200, SE 40530 Gothenburg, Sweden
E-mail: jens@ling.gu.se, max.boholm@telia.com

Abstract

This paper presents a study of multimodal communication in spontaneous "getting to know each other conversations". The study focuses on repeated head movements (head-nods and head-shakes) and the speech co-occurring with them. The main function of such repeated head movements is found to be communicative feedback. This is also the most frequent function of the speech co-occurring with the head movements. However, there is mostly no 1-1 relation between repetition in head movement and vocal words. Repeated head movements are more often accompanied by single than repeated words. Both repeated head movements and repeated vocal words can also occur without accompaniment in the other modality. In such cases, the most frequent function for the head movements is still communicative feedback. However, the most frequent function of repeated words without accompaniment in the other modality is own communication management. Frequent functions of repeated head movements, besides feedback, are emphasis, self-reflection, citation, self-reinforcement and own communication management. Other findings in the study are that affirmative repeated head nods mostly start with an upward movement and involve two repetitions.

1. Introduction

Research based on naturalistic multimodal face-to-face communication is important both for a deeper understanding of human-human communication and human communication with virtual agents. This is the underlying motivation of the NOMCO project – a cooperative project between Sweden, Denmark and Finland. In the project we are investigating human-human communication and creating an annotated corpus that will be available for other researchers over the Internet. The project is focused on studies of feedback, turn management, own communication management (OCM) and information structure. In this paper we present a study of repeated head movements (nods, jerks, shakes and tilts) and their relation to speech and attempt to answer the following questions:

(i) Which head movements occur in repeated form?
(ii) What is uttered (if anything) simultaneously with a repeated head movement?
(iii) What is the function of repeated head movements and of multimodal contributions involving both repeated head movements and speech?

A specification of the second question concerns whether repetition in head movement accompanies repetition in speech, or vice versa. Figure 1 below shows the possible combinations that exist in multimodal contributions regarding repetition in speech and/or repetition of gesture:

		Speech +	Speech -
Repetition of head movement	+	**A**	**B**
	-	**C**	

Figure 1. Possible relations between repeated head movements and repetition in speech.

Since head movements (mainly head nods and head shakes) frequently are used to express communicative feedback (Allwood and Cerrato 2003), we will in this paper focus our interest on the relationship between head movements as communicative feedback and the vocal counterpart of repeated head movement feedback, namely repeated feedback words.

2. Method, data and procedure

The data consists of four video recordings of strangers meeting for the first time who during approximately seven minutes get to know each other. There are two participants in each recording. In two of the recordings both participants are female, in two of them one participant is female and the other one male. The interactions are in Swedish.

The three recordings have been coded for

(i) Repeated head movements
(ii) What is uttered (if anything) by the head mover simultaneously with the movement
(iii) Repeated words, which do not occur with repeated head movements

The head movements have been classified according to overall movement type (nod, shake or other), number of repetitions and (initial) direction (up or down for head nods and left or right for head shakes). Head nods and head jerks are not separated as different main types (both are classified as head nods). They can, however, be differentiated on the basis of the initial direction of the repeated head nods, i.e. up (sometimes called *jerk*) and down (*nod* in a more restricted sense than in this article).

The identification of a repeated head movement is based on the definition of a single use of a head movement, as indicated below (cf. Allwood et al 2005).

- Nod: up, then down, *or* down, then up
- Shake: left, then right, *or* right then left
- Tilt: leaning head to one side

A repeated head movement occurs whenever the movement continues over and above the single head movements described above. Since the last repeated movement in a sequence of repeated movements can have the same direction as the first movement in the sequence, it is possible that the number of movements in one direction exceeds the number of movements in the other direction. For example a head nod can be produced as up-down-up-down-up. To account for this, we have counted half units of repeated head movements. The case of an up-down-up-down-up nod has been coded as a "2+½ up-nod".

Slow motion playback has been used to identify repeated head movements. In addition, a functional analysis has been done of the movements.

To test inter-annotator reliability a random selection of 15 coded head nods (137 in total) and five coded head shakes (total: 21) was made. Based on the sample for coded head nods there is 93% agreement on the direction of head nods (κ=0.86, according to Cohen's kappa; Cohen 1960) and 80% agreement on number of repetitions. There is less agreement on the coding of the head shakes. Three of the five selected head shakes are agreed upon (both direction and number). Concerning the functional coding there is complete agreement on the function, or functions, in 85% of the 20 randomly selected cases. The 15% of non-complete agreement consist in cases of agreement on one function (feedback: CPU), but where one of the annotators has suggested an additional function to the one agreed upon, not suggested by the other (e.g. agreement).

3. Results

3.1 Overview

There are a total number of 162 repeated head movements in the four recordings: 137 repeated head nods, 21 repeated head shakes, three repeated tilts, and one repeated circling of the head. Most of the repeated head nods start in an upward direction (cf. the notion of jerk). The number of repetitions ranges from one up to 8+½ repetitions of the movement. Tables 1 and 2 provide information on type of head movement, initial direction, no. of repetitions and the frequency of repeated head nods and head shakes.

No. of repetitions	**Right**	**Left**	*Σ*
2	4	2	*6*
2+½	4	1	*5*
3	4	0	*4*
4	1	1	*2*
4+½	1	0	*1*
5	2	0	*2*
6	1	0	*1*
Σ	*17*	*4*	*21*

Table 1. Initial direction, number of repetitions and frequency of head shakes

No. of repetitions	**Up**	**Down**	*Σ*
1+½	2	3	*5*
2	26	16	*42*
2+½	8	2	*10*
3	11	10	*21*
3+½	3	3	*6*
4	9	9	*18*
4+½	2	3	*5*
5	10	3	*13*
5+½	1	1	*2*
6	6	1	*7*
6+½	2	0	*2*
7	2	2	*4*
8	1	0	*1*
8+½	1	0	*1*
Σ	*84*	*53*	*137*

Table 2. Initial direction, number of repetitions and frequency of head nods

A majority of the repeated head movements (121, i.e. 74%) are produced simultaneously with speech. The same person produces speech and head movement making them part of a complex multimodal contribution. The movements not produced simultaneously with speech belong to two types: either they are produced in an overlap with another speaker or they are part of the head mover's turn, who however does not speak at the time of producing the head movement.

In most cases the speech produced simultaneously with head movements has a feedback function (for the Swedish system of communicative feedback see Allwood 1988, Allwood and Ahlsén 1999). In the schema below, we show the coded repeated head movements and their co-occurrence with speech.

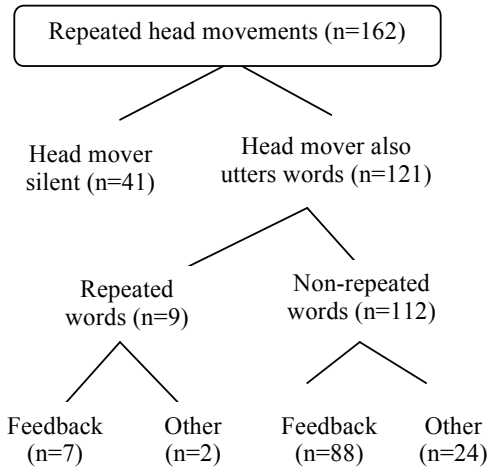


Figure 2: Repeated head movements and their co-occurrence with speech

In two instances the simultaneous speech was incomprehensible. No functional analysis was done in these two cases.

In addition to the repeated speech co-occurring with repeated head movements there is also repeated speech that does not co-occur with repeated head movements (19 instances). These vocal verbal repetitions include repetition of feedback words and phrases (8 instances), but mainly repetitions of other words with the function of own communication management for choice (i.e. word finding) (9 instances as well as a a repeated greeting *hej hej* ('hi hi').

3.2 Feedback

Communicative feedback is given by (unobtrusive) vocal and bodily expressions to inform an interlocutor about ability and willingness to (i) continue the interaction (C), to (ii) perceive (P), and (iii) understand (U) what is communicated, as well as about (iv) other attitudinal and emotional reactions (A) (see Allwood 1988, Allwood and Ahlsén 1999, Allwood et al 2007).

Thirty-three (33) of the 41 of the repeated head movements produced under silence have feedback function, expressing contact/continuation, perception and understanding. In some cases the function seems also to be stronger by also expressing acceptance or agreement with what is said by the interlocutor.

When simultaneous with speech, repeated head nods and headshakes most often are simultaneous with vocal (verbal) feedback expressions, making up multimodal feedback (95 of the 162 (59%) repeated head movements are simultaneous with feedback words). However, in only seven cases, a repeated nod or shake occurs simultaneously with repeated feedback words. Table 3 shows the vocal feedback (FB) expressions co-occurring with repeated head movements.

Non-repeated FB expressions	
m ('m')	31
{j}a ('yes')	9
okej ('ok')	9
ja ('yes')	7
{j}a precis ('yes exactly')	3
jo ('yes (rejection of negation)')	3
jaha ('yes, I see')	2
mhm ('mhm')	2
nä ('no')	2
{j}a / de{t} blir de{t} ('yes it will be like that')	1
{j}a / de{t} e0 ju de{t} ('yes it is you know')	1
{j}a de{t} e0 en bra början ('yes that's a good start')	1
{j}a de{t} e0 första boken ('yes it's the first book')	1
{j}a precis nej ('yes exactly no')	1
amen precis {j}a ('yes exactly yes')	1
ja okej ('yes ok')	1
ja precis {j}a ('yes exactly yes')	1
japp ('yup')	1
m // ja ('m // yes')	1
m okej ('m ok')	1
nä / de{t} e0 de{t} ('no it is')	1
nä ja{g} hö{r} de{t} knappt alls ('no I hardly hear it at all')	1
näe ('no')	1
nämen precis ('no but exactly')	1
okej / {j}a ('ok / yes')	1
okej {j}a ('ok yes')	1
okej ja ('ok yes')	1
okej jag fattar ('ok I understand')	1
okej men ('ok but')	1
Repeated FB expressions	
m / m ('m / m')	3
jo precis precis ('yes exactly exactly')	1
nä det stämmer det stämmer det stämmer ('no that's correct that's correct that's correct')	1
nä nä de{t} / de{t} ja{g} tänkte ('no no that / that I thought')	1
okej akej ('ok ak')	1

Table 3. Vocal feedback expressions simultaneous with repeated head movements

Comment: slash "/" is used for short pauses and brackets, "{" and "}", surrounds non-pronounced part of standard orthographical representation of a word, e.g. the words *det* ('it') and *ja* ('yes') are pronounced without final and initial consonants respectively (cf. Nivre 1999).

All repeated headshakes involved in feedback co-occur with *nä* ('no') or related variants (*näe*, *nämen*), with the exception of one headshake that co-occurs with *jaha* ('yes, ok'), where the verbal message expresses acceptance of what previous speaker has said but the head shake expresses lack of understanding of what has been accepted. Head nods occur with more positive feedback words like (*m*, *mhm*, *okej, {j}a* ('yes') etc.). However, three repeated head nods co-occur with nä ('no'). These occurrences of *nä* all follow negative claims by the interlocutor where the *nä* expresses acceptance (by agreement on negative polarity) and the head nod also expresses acceptance but without the negative agreement. All instances of bodily and multimodal feedback (including those containing *nä*) discussed here therefore express CPU and in some cases also acceptance.

Based on the data presented above, we may now consider the following two questions:

- What does repeated head movement add to the function expressed by the vocal verbal feedback?
- Does the head movement have an independent meaning from the vocal component?

Considering the first question one answer is that the head movement strengthens the function expressed by the vocal expression. The complex multimodal expression of head movement and words is more robust and perceptually salient in interaction than words or gesture alone. However, answering the second question, one might consider the possibility that the gesture not only intensifies the functions of the vocal component. For example the repeated head movement sometimes also indicates self-reflection, over and above understanding, a kind of activation perhaps assisting the head mover in understanding what is communicated.

Another observation is that all uses of repeated head movements that co-occur with the word *okej* ('ok') are

instances of a repeated upward nod with two or three reduplications. The function of *okej* (alone and in combination with the upward repeated head nods) is that of expressing the notification (being made aware of) of a fact. Further investigation might deepen the understanding of short upward head nods as a bodily means of expressing the same or a similar function to that which is expressed by vocal *okej*.

In addition to multimodal feedback there also are occurrences of repeated feedback words, not occurring with repeated head movements. In the observed sample, the following eight instances of repeated vocal feedback expressions are not accompanied by repeated head movement: *ja ja ja* ('yes yes yes'), *jo jo* ('yes yes' (rejection of negation)), *nä nä* ('no no'), *nä nä nä* ('no no no'), *precis / precis* ('exactly / exactly'; two instances), *{j}a ja{g} vet {j}a ja{g} vet* ('yes I know yes I know'), and *ja de{t} e0 klart de{t} e0 klart* ('yes that's clear that's clear').

In comparison with the Gothenburg Spoken Language Corpus (GSLC), the relative frequency of repeated feedback words in the four interactions under investigation is slightly lower. GSLC contains 1.4 million words and almost 3 600 instances of repeated feedback words, not interrupted by pauses (approximately 1:400 ratio). (The comparison relies on frequency of repeated feedback words without intervening pauses to reduce the complexity of making the comparison). The four recordings, analyzed in this paper comprise slightly more than 6 300 words, but only five instances of repeated feedback words not interrupted by pauses, i.e. *nä nä* (2), *nä nä nä*, *ja ja ja* and *jo jo* (approximately 1:1 200 ratio). Hence, repeated feedback words are approximately three times more common in GSLC. A possible explanation for the difference might be the lack of familiarity and slight awkwardness often present in a first encounter. Perhaps repetition is facilitated by familiarity.

3.3 Other functions of repeated head movements

Even though most of the repeated head movements have a feedback function (alone or in combination with speech), all do not. In total there are 32 instances of repeated head movement that do not have a feedback-function. Other functions identified are:

- *Emphasising words and phrases uttered*. Head nods and headshakes simultaneous with speech are used to emphasize parts (words) of what is said (cf. "batons" of Ekman and Friesen 1969: 68; also see Bull and Connelly 1985, McClave 2000). This is also the function of one of the repeated tilts and the repeated circling of the head identified in the empirical data. All except two of the head shakes that do not co-occur with the feedback-word *nä*, instead co-occur with negations such as *inte* ('not') and *aldrig* ('never'), i.e. are used to emphasise the negation in what is uttered (10 instances).
- *Self-affirmation*. Another function of repeated head nods apart from feedback is that of expressing self-affirmation. For example, repeated head nods are used after statements as an expression of self-reflection, acceptance and reinforcement of what the speaker him/herself has said (12 instances of repeated head nods).
- *Citation and imitation*. In one case of a repeated head nod, and one case of a repeated tilt, the function is that of marking a switch from direct to indirect discourse, i.e. the introduction of quotes (see McClave 2000). Alternatively to *introduce* quotes, the function of these head movements can be that they are used as *part of* the quote, to imitate not only what others have said, but also what they have gestured.
- *Part of own communication operations of choice* (word finding). In the two cases of reduplication of both speech and gesture that are not feedback (see schema above), the function identified is "own communication management" of choice, i.e. activation, planning and hesitation about what to say (cf. Allwood et al 2006, McClave 2000). In total, two repeated head shakes and three repeated nods are involved in own communication management.

We may also consider the question of what the function is of repetition, whether it be of head movement or of vocal words. There are at least three possible answers to this question: (i) there is a common function of repetition for both modalities, (ii) there is one common function for head movements and a different common function for vocal words or (iii) there are several functions of repetition both for vocal words and head movements.

We believe it is possible to give an analysis of the function of repetition that provides an answer of the first type by claiming that repetition always has the function of activation.

If we relate this claim to the specific functions of repeated head gestures, we have found in the data, we suggest the following analysis. Repetition always means increased activation of the expression and/or content of the gesture or vocal word that is repeated. All functions we have found associated with repeated forms also occur with the single vocal words and gestures corresponding to the repeated forms, but in less activated form

This increased activation can be used for emphasis (intensification) of the element that is repeated e.g. feedback or other types of content. A special case of emphasis is self-reinforcement, i.e. reinforcement of something one is communicating. Another special case is hyperbole and irony, where the intensification is exaggerated or unmotivated.

The increased activation, however, can also be used for improved activation of content or expression which has not been completely successfully activated (e. g. the choice function in OCM (word finding)) and, as a special

case of this, self-reflection, where the same expression and content are repeatedly activated to activate further thoughts. Repetition with this function can also have the function of turn keeping or turn holding. One keeps the turn by indicating or displaying by repetition that one has more to say.

Two further special functions noted above are citation through imitation of someone else's repetition and switching from direct to indirect discourse. However, we do not believe that these two functions are specific to repetition since repetition like any other communicative feature can be imitated and that repetition to indicate a breach or a switch in the flow of discourse could also be achieved by other noticeable communicative means. The fact that repetition can be used in this function could perhaps also (besides its salience) be seen as an extension of the OCM choice function.

In sum, we, thus, suggest that the main function of repetition is activation of expression (vocal or gestural) and content of an element of communication. This activation has two main uses:

(i) Increased activation if the element is not sufficiently activated, e.g. word finding, self-reflection.
(ii) Emphasis (intensity) and self-reinforcement

4. Conclusions

Repeated head movements mainly function as feedback. They may be part of multimodal feedback expressions or serve as bodily expressions of feedback without speech. There is no strong tendency for repetition in head movement to co-occur with repetition in speech or vice versa In addition to strengthening the information provided by the feedback words, making the signaling of feedback functions more robust and perceptually salient, repeated head movements can have the functions of emphasising the words and phrases uttered, self-affirmation and self reflection as well as assisting in choice related own communication management. In general, repetition always means increased activation of the expression and/or content of the gesture or vocal word that is repeated.

5. Acknowledgements

This study was funded by the Swedish Research Council (VR), project Embodied Feedback and by the Nordic council (Nordcorp). We are grateful to Elisabeth Ahlsén for comments and discussion.

6 References

Allwood, J. (1988) Om det svenska systemet för språklig återkoppling. In: P Linell, V. Adelswärd & P. A. Pettersson (ed.) *Svenskans beskrivning 16*, vol. 1. Linköping: Tema kommunikation, Linköpings universitet.

Allwood, J. and Ahlsén, E. (1999) Learning how to manage communication, with special reference to the acquisition of linguistic feedback. *Journal of Pragmatics*, 31: 1353-1389.

Allwood, J. & Cerrrato, L. (2003) A study of gestural feedback expressions, *First Nordic Symposium on Multimodal Communication,* Paggio P. Jokinen K. Jönsson A. (eds), Copenhagen, 23-24 September 2003, ISSN 1600-339X, pp.7-22

Allwood, J., Cerrato, L., Dybkjaer, L., Jokinen, K., Navaretta, C. and Pagio, P. (2005) The MUMIN multimodal coding scheme. In *NorFA Year book 2005*.

Allwood, J., Ahlsén, E., Lund, J. and Sundqvist, J. (2005) Multimodality in Own Communication Management. In: *Proceedings from the Second Nordic Conference on Multimodal Communication*, Göteborg, 2005.

Allwood, J., Kopp, S., Grammer, K., Ahlsén, E., Oberzaucher, E. & Koppensteiner, M. (2008). The analysis of embodied communicative feedback in multimodal corpora: a prerequisite for behavior simulation. *Language Resources and Evaluation*, 41(3-4): 255-272

Bull, P. and Connelly, G. (1985) Body movement and emphasis in speech. *Journal of Nonverbal Behavior*, 9(3): 169-187.

Cohen, J. (1960) A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20(1): 37-46.

Ekman, P. and Friesen, W. V. (1969) The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1: 49-98.

McClave, E. Z. (2000) Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32: 855-878.

Nivre, J. (1999) *Transcription Standard Version 6.2*. Department of Linguistics, University of Gothenburg.

Using head movement to detect listener responses during multi-party dialogue

Stuart A. Battersby & Patrick G.T. Healey

Queen Mary University Of London,
Interaction, Media & Communication Research Group,
School of Electronic Engineering & Computer Science,
London, E1 4NS

stuart@dcs.qmul.ac.uk, ph@dcs.qmul.ac.uk

Abstract

Multi-party interactions present some unique problems for the analysis of non-verbal behaviour. The potential complexity of non-verbal cues increases as the number of participants increases. Our analytical approach involves a hybrid of human annotations from video data and machine analysis using motion capture data. The literature often cites eye gaze and head orientation as the critical cues for managing dialogue. In multi-party dialogue the relative importance of these cues changes; participants can only monitor the eye gaze of one person at a time. In previous work we identified the relative contributions of changes in speaker head orientation and gesture orientation on the listeners using data from a corpus of 3-person interactions. We detail here our methodology used in analysing motion capture data and combining it with human annotations, and discuss two approximations of recipient responses designed to highlight interactional patterns: head reorientations and head nods.

1. Introduction

As face to face dialogue is an open ended activity, it presents challenges for the analysis of non-verbal behaviour. One way to overcome this is to impose constrains upon people. A common form of this is to ask participants to retell a story, typically a scene from a cartoon, to camera. Whilst such an approach provides a narrative in a controlled environment in which the relationship between gesture and speech can be examined, it masks the *interactive* nature of dialogue (see Goodwin (1979) for evidence that the verbal production of an utterance is dynamic and changes depending upon the actions of the addressee(s)). A demonstration of interactive non-verbal behaviour can be found from Furuyama (2000) who shows that, during an origami tuition task, interlocutors are able to produce gestures that are a product of their collaboration. Also Morency et al. (2008) show that interlocutors' movements such as nodding are influenced by the local interactional context and Bavelas et al. (1992) discuss the notion of interactive gestures which are said to be a class of gesture which refer to the participants rather than topic of discussion. These gestures also help to manage the dialogue.

Multi-party dialogue provides a more complex environment in which interlocutors must coordinate their dialogue activity than that offered during dyadic dialogue. The literature often cites eye gaze and head orientation as critical cues to activities such as turn taking (see e.g. Argyle (1975)). In a study of listener gaze patterns, Gullberg (2003) found that they fixate the speaker's face 96% of the time, with only 0.5% of the time spent looking at their gestures. The problem of concern here is that, whilst these cues may be representative of dyadic dialogue, they do not necessarily generalise to multi-party. Interlocutors can only monitor the eye-gaze of one person at a time. Loomis et al. (2008) demonstrate that in situations such as small group conversations, we are only able to reliably judge another's eye gaze within a 4° rotation. Head orientation can be effectively judged up to rotations of 90°. The eye gaze issue is exaggerated because, when there are more than two people in the dialogue, who is looking at whom matters more in multi-party dialogue (for example, it can help to determine the intended addressee of the word *you* (Frampton et al., 2009; Purver et al., 2009)). Utterances and actions can have different interpretations depending upon both speaker and listener orientations.

Battersby and Healey (submitted) compare the significance of changes in speaker head orientation and gesture orientation during 3-way conversations by measuring the responses of both recipients. Two measures of responses are used: head reorientations - e.g does the recipient turn their head from the speaker to the third person in the interaction - and head nods. Contributions of hand movements and head movements are compared and, in contrast to previous literature, it is shown that changes in speaker hand orientation elicit more and faster responses from recipients than changes in head orientation.

Motivated by the need to analyse this data we made use of an analytical approach which uses both human annotation of video and machine analysis of motion capture data (a similar hybrid approach has been employed in Jokinen (2010)). Human annotations were used for identifying the speaker's target events whilst machine analysis was used for detecting recipient responses. The critical factor in designing the machine analysis was to create measures that allow us approximate changes of orientation and head nods, and hence draw *comparisons* between people or conditions, rather than precisely detecting events.

In this paper we focus on the methodology used in Battersby and Healey (submitted) starting with a description of the corpus data, then a discussion of the details of the techniques and finally we present brief results and consider potential extensions and modifications to our approach.

2. Details of the corpus

The corpus used for the study was collected in the Augmented Human Interaction (AHI) lab at Queen Mary. This lab houses an optical motion capture system and video recording equipment. 33 participants took part in the study in groups of 3 meaning the corpus consists of 11 triads. Six tuition tasks were developed that consisted of either a short Java program or a description of a system of government. The material was all text based with no graphical representations.

Each group performed three rounds of tuition. On the first round one member of the triad was assigned the role of 'learner' and the remaining two members were assigned the roles of 'instructors'. On the subsequent rounds these roles were rotated so that each person was a learner once. On each round the instructors were given printed tuition material which they were asked to collaboratively teach to the learner. They were allowed to familiarise themselves with the material prior to tuition and then returned it to the experimenter. During this time the learner was removed to another room. The learner and the instructors then sat on stools in the AHI lab and the tuition commenced. There was no time limit and were no restrictions other than they were not allowed to use pen and paper. To motivate the participants to teach and learn, a post-completion test was used.

3. Methodology

3.1. Coding video data

Each round of tuition is recorded by three video cameras one above and one to either side of the participants (see Figure 1). These cameras are software synchronised via networked computers in order to start and stop together. The videos are imported into ELAN where they are coded for the target events. A code was made for every visible change in *speaker* head or gesture orientation relative to the recipients. Each of these codes was then sub-coded by the following classification:

- Head Moves: Here the head orientation changes but the gesture remains stationary. For example, the speaker may be gesturing towards the primary recipient and glance (by turning their head) towards the other, secondary, recipient.
- Hand Moves: Here the gesture moves, but the head orientation remains stationary. For example, the speaker could be gesturing with a palm up gesture towards the primary recipient and whilst continuing to look at them, turn their gesture so that it is oriented to the secondary recipient.
- Both Move: Here both the gesture and the head shift orientation. For example, the speaker could be pointing towards the primary recipient then turn their point along with their head orientation towards the secondary participant.

All coding was performed by the 1st author. A random sample of 25 events including target events and control events was coded by a second rater. The inter-rater reliability was good with Kappa = 0.78, $p < 0.001$.

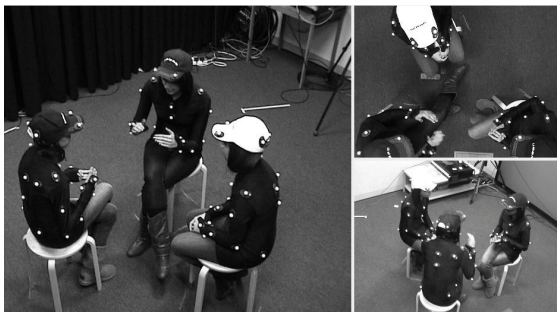


Figure 1: Three participants during a round of tuition

3.2. Machine Analysis

3.2.1. Overview of the motion capture system

The motion capture system used is an optical based system supplied by Vicon. For this study the system was set to record data at 60 frames per second. Each participant wears an upper body motion capture suit and a baseball cap. 27 reflective markers are attached to these which are then tracked by an array of 12 infra-red cameras. The software supplied with the system uses the data from each camera to reconstruct the 3D scene, providing the coordinates of each maker on each frame of data.

Once the data has been labelled it can be exported as a tab delimited file and used in other software. We have written analysis software using Python that will read the data files for the whole corpus and perform the quantitative analysis.

3.2.2. Extracting Head Orientation

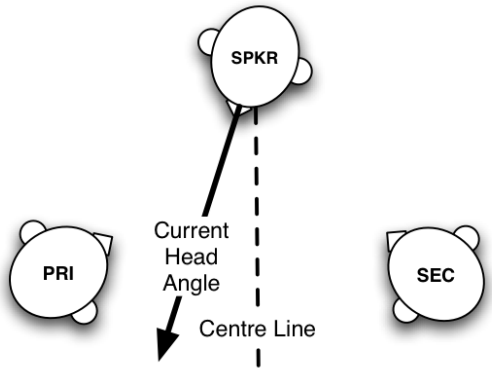


Figure 2: Head orientation detection

We extract head orientation data for each participant by using the markers on the head to create a vector which extends forwards from their forehead. Whilst it is impossible to determine the exact focus of attention for a person, we approximate attention by comparing the current head orientation to a centre line down the interaction space. For any one person, this center line is a vector that originates at the rear centre of the head and extends midway between the other two people based upon their current position (see Figure 2). This line forms a bisection of the interaction space,

indexed by the the rear of the head. We update this line on every frame of data and use it to determine which person is being oriented to by whom. If any head orientations are within 2 ° of the centre line we exclude this data.

In addition to using the head orientation as an indicator of focus of attention, we also use it to detect turning of the head, or reorientations. We look for times in which a participant's head angle crosses over the centre line and class this as an instance of a reorientation.

3.2.3. Extracting Head Nods

As with measuring head angle, it is impossible to identify exactly a nod. However, given a set of parameters we are able create a measure that can be used for comparison.

The data used for the analysis come from a front head marker in the vertical axis. This coordinate data can be interpreted as a signal. As this signal is global to the whole 3D scene, along with nodding it will contain a variety of movement including prosodic body sways, gross shifts in posture, head movement, unintentional body movements, and more. We apply signal processing techniques in order to filter out some unwanted information.

We first shift the zero position of the signal to be the mean position of the head marker (see Figure 3). This gives

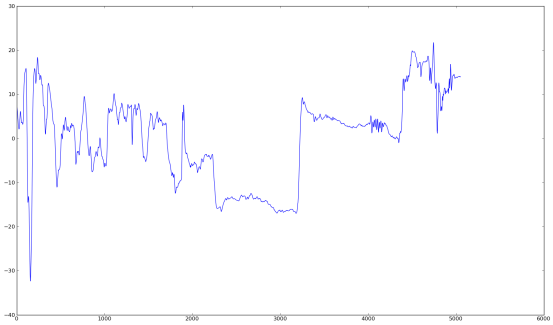


Figure 3: Zeroed to mean head marker signal showing all movement. Y axis: amplitude (-40 - 30), X axis: frame number (time)

values that show movement above and below this mean position. For the next step we exclude any low frequency movements that could account for body sway etc by removing frequencies below 2Hz. We then remove high frequency movements that could be caused by body shakes, camera error etc by removing frequencies above 8Hz. This filtering gives us a range of frequencies that could plausibly be the frequency of a nod (see Figure 4).

We then detect all the peaks and troughs in the signal (which could represent the top and bottom of a nod) that have a movement greater than 1.5cm and produce a new signal of these values. We remove from this signal any movement greater than 5cm from the mean position as this is more likely to be the result of a posture shift than a head nod. Taking the resulting signal we invert the troughs so that we have only a positive signal that represents the motion that we have narrowed down (see Figure 5). By applying a finite smoothing filter to this signal (using a window of

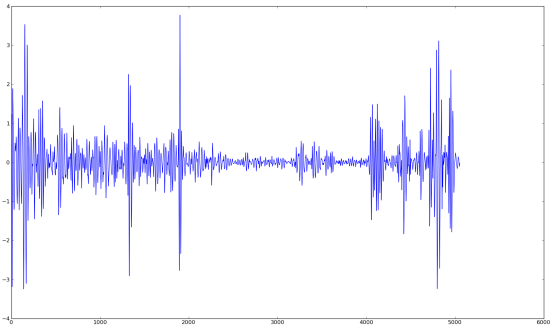


Figure 4: Filtered head marker data. Y axis: amplitude (-4 - 4), X axis: frame number (time)

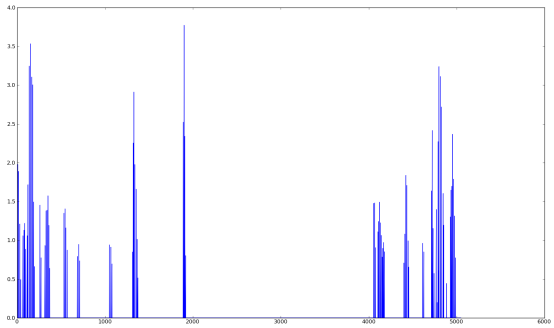


Figure 5: Peaks and inverted troughs. Y axis: amplitude (0 - 4), X axis: frame number (time)

circa 0.5 seconds) we create a signal which now represents areas of motion (see Figure 6). We define a cut off point of 0.1 above which we code for the presence of a 'nod'.

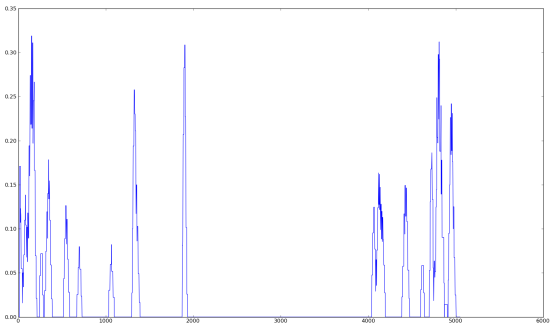


Figure 6: Smoothed final signal which represents periods of nodding. Y axis: amplitude (0 - 0.35), X axis: frame number (time)

3.3. Detecting responses

Our main focus was to compare response rates for particular classes of events. For this we need to judge whether a response has occurred or not. The ELAN transcription files contain the time information of each target event that has been annotated. These are imported into the software and combined with the times for head reorientations and head

nods. For each target event we judge the primary recipient as the recipient that the speaker is oriented to at the start of the event, the remaining person is the secondary recipient. This is judged using the head orientation data. When a target event occurs, a 5 second bracket following the event is used to identify any change in orientation and any nod from either of the recipients. The result is the frequency and delay times of response for each class of target event.

In order to compare the target events to a baseline of head orientation and head nodding throughout the dialogues, we select random times during a speaker's utterance (ensuring that these do no coincide with an actual target event). This collection of baseline events are put through the same response detection process. Both groups of events are compared for statistical significance.

4. Results

Full details of the results are reported in Battersby and Healey (submitted). Over 2 hours and 54 minutes of dialogue, 287 target events were identified. Recipient head orientation was shown to be significant (primary recipients orient to the speaker 68% of the time whereas secondary recipients' orientation is split 50-50 between the speaker and the primary recipient ($\chi^2 = 16.9, p < 0.001$)), and the target events were consequential for the interaction (they elicited a 48.6% head reorientation response rate compared to the 41.3% baseline rate ($\chi^2 = 5.75, p < 0.05$) . We found that changes of orientation of the speaker's hand reliably produced faster and more frequent recipient responses (significant head reorientation response rate of 63%) than movements of the head (for which there was no reliable difference from the baseline).

In addition to these results we can also report on the effect of recipient entry orientation and dialogue role on the response rate. For both the primary recipient and the secondary recipient, we compare their response rate for head reorientations and nodding, depending on whether they are looking at the speaker or not.

When measuring head reorientations, we observe that the primary recipient will respond 56.4% of the time when looking at the speaker (and hence turn to the secondary recipient), but 73.5% of the time when looking at the secondary recipient (and hence turn back to the speaker). This difference is significant ($\chi^2 = 8.65, p = 0.003$). This difference is not observed for the secondary recipient, they are equally like to respond depending upon their entry orientation.

When looking at head nods these patterns are not observed, there is no significant difference between response rates for entry orientations for either participants.

5. Discussion

Given that we see interactionally significant patterns which are statistically reliable emerging, it is clear that our simple and effective approach is indexing aspects of the participants' behaviour that are meaningful for the interaction. To a certain extent it does not matter whether the analysis really identified 'nods', or whether the orientations were accurately identifying focus of attention. It matters that it indexes movements that are clearly visible to the interlocutors, and have a systematic and marked effect on the interaction.

We may be able to refine our technique further. In our analysis of head nodding we make use of one single marker's movement in the vertical plane. Using the head as a complete segment with rotation data may remove the need for some of the signal processing which excludes unwanted body movements. We also make some logical assumptions as to the frequencies that should be identified in the head motion signals. A measure which allows us to tease apart the frequencies of movement within the signals as the interaction unfolds would lead us to more informed decisions when selecting parameters. A technique which is being use on human interaction data with interesting results is cross-wavelet analysis (see Issartel et al. (2006) for details of this).

We also need to consider an analysis of hand movement as we have shown that it is a highly significant cue for management within dialogue. This movement is less constrained than the movement of the head, and is more susceptible to missing data due to occlusion in the tracking environment so any approaches to automated analysis must be very robust. Our current work tries to integrate this hand movement data by using the speed of the hands to create discrete events (i.e. times of motion) and also for regression analysis.

As we are examining interaction data, rather than narratives it is necessary to understand how people move together. As Morency et al. (2008) suggest, a listener's movements are influenced by the actions of the other participants of the dialogue. A speculation is that people may coordinate their motion (e.g. jointly coordinate their head movements). One approach to this possibility would be to examine cross correlations of head movement and speech between the individuals in the group (see Lavelle et al. (2009) for similar work on interpersonal coordination involving a patient with a diagnosis of schizophrenia).

6. Conclusion

We have made initial steps towards an analysis of multiparty non verbal behaviour. By using real face-to-face dialogues to measure listener responses we have respected the interactive nature of dialogue. The technique that we have employed gives us comparable measures so that we are able to see patterns that exist naturally within interaction. These patterns now require experimental analysis to determine when they occur and when they do not. A promising approach for this is to use virtual avatars as an experimental platform.

The findings are significant, not only for understanding the processes of communication, but also for systems which make use of computer vision techniques (such as automated meeting analysis). Techniques are already available to gain access to the movement data of a participant without the need for motion capture equipment. Our findings can guide these techniques so that they are applied to interactionally significant phenomena.

7. References

Michael Argyle. 1975. *Bodily Communication*. Methuen & Co. Ltd, Bristol.

Stuart A Battersby and Patrick G T Healey. submitted. Head and Hand Movements in the Orchestration of Dialogue. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Janet Beavin Bavelas, Nicole Chovil, Douglas A Lawrie, and Allan Wade. 1992. Interactive Gestures. *Discourse Processes*, 15(4):469–489.

Matthew Frampton, Raquel Fernández, Patrick Ehlen, Mario Christoudias, Trevor Darrell, and Stanley Peters. 2009. Who is "You"? Combining Linguistic and Gaze Features to Resolve Second-Person References in Dialogue. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 273–281, Athens, Greece, March. Association for Computational Linguistics.

Nobuhiro Furuyama. 2000. Gestural Interaction between the instructor and learner in origami instruction. In *Language and Gesture*. Cambridge University Press.

Charles Goodwin. 1979. The interactive construction of a sentence in natural conversation. In G Psathas, editor, *Everyday language: Studies in ethnomethodology*, pages 97–121. Irvington Publishers.

Marianne Gullberg. 2003. Eye movements and gesture in human face-to-face interaction. In J Hyönä, R Radach, and H Deubel, editors, *The mind's eye: Cognitive and applied aspects of eye movements*, pages 685–703. Oxford: Elsevier.

Johann Issartel, Ludovic Marin, Thomas Bardainne, Philippe Gaillot, and Marielle Cadopi. 2006. A Practical Guide to TimeFrequency Analysis in the Study of Human Motor Behavior: The Contribution of Wavelet Transform. *Journal of Motor Behavior*, 38(2):139–159.

Kristiina Jokinen, 2010. *Gestures and Synchronous Communication Management*. Springer.

Mary Lavelle, Rosemarie McCabe, Patrick G T Healey, Christopher Frauenberger, and Fabirzio Smeraldi. 2009. Interpersonal Coordination in Schizophrenia: The first 3D analysis of social interaction in schizophrenia. In *Proceedings of the International Society for the Psychological Treatments of Schizophrenias and Other Psychoses conference 2009*, Copenhagen.

Jack M Loomis, Jonathan W Kelly, Matthias Pusch, Jeremy N Bailenson, and Andrew C Beall. 2008. Psychophysics of perceiving eye and head direction with peripheral vision: Implications for the dynamics of eye gaze behaviour. *Perception*, 37:1443–1457.

Louis-philippe Morency, Iwan De Kok, and Jonathan Gratch. 2008. Context-based Recognition during Human Interactions: Automatic Feature Selection and Encoding Dictionary. In *ICMI 08*, Chania, Crete.

Matthew Purver, Raquel Fernández, Matthew Frampton, and Stanley Peters. 2009. Cascaded Lexicalised Classifiers for Second-Person Reference Resolution. In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL 2009 Conference)*, pages 306–309, London, UK, September. Association for Computational Linguistics.

Developing heterogeneous corpora using the Digital Replay System (DRS)

Dawn Knight, Paul Tennent, Svenja Adolphs and Ronald Carter
The University of Nottingham
School of English Studies, The University of Nottingham, Nottingham, NG7 2RD
E-mail: Dawn.Knight@nottingham.ac.uk, pxt@cs.nott.ac.uk, Svenja.Adolphs@nottingham.ac.uk, Ronald.Carter@nottingham.ac.uk

Abstract

This paper reports on the latest developments made as part of the ESRC funded Understanding Digital Records for eSocial Science Project (DReSS) at the University of Nottingham. Specifically, it reports on some of the issues and challenges that are currently being faced in compilation and use of heterogeneous multi-modal corpora comprised of *heterogeneous* datasets; discussing some of the optimum ways in which these datasets are recorded, processed, stored and accessed/interrogated by the linguist. The paper profiles the Digital Replay System (DRS) software which is being developed to support these processes.

1. Introduction

Multi-modal corpora are comprised of video, audio and textual records of interaction (and associated metadata information) extracted from recordings of naturally occurring conversational episodes which are *streamed* in an easy-to-use interface; the multi-modal corpus workbench. A *mode* of data, in this sense, is crudely defined as the physical format in which a particular linguistic or extra-linguistic phenomenon is presented and observed; thus, in this current article *multi-modality* is defined as the culmination of these integrated and aligned data streams.

Heterogeneous digital corpora are defined as emergent multi-modal datasets which may be comprised of a variety of different records of everyday communication from a range of different discursive environments; from face-to-face to digital mediums. These may include a variety of different forms of media 'types' including interaction in virtual environments (instant messaging, entries on personal notice boards etc), GPS data, face-to-face situated discourse, phone calls, video calls and so on.

The ESRC-funded Digital Records for eSocial Science Project (DReSS, based at the University of Nottingham, see Knight et al., 2006, 2009; Knight and Tennent, 2008) project seeks to allow for the collection and collation of a wider range of heterogeneous datasets for linguistic research, with the aim of facilitating for the investigation of the interface between multiple modes of communication in everyday life. This paper examines some of the key issues and challenges faced when attempting to achieve this aim.

2. The Nottingham eLanguage Corpus

Specific forms of data that are currently being compiled for the 'Nottingham eLanguage Corpus' (NeLC), as part of DReSS, are broadly categorised into the following 'types':

- Text-based eLanguage data: Text-based communicative data sent and received through digitally based mediums.
- Language 'in the wild': Capturing 'a day in the life of your language'- a corpus of language reception (not production).

The 'text-based eLanguage data' strand includes common digitally based language resources which have become integral to modern life in the digital age, that is, any linguistic stimulus that is digitally based. For the specific purpose of the NeLC, this includes SMS/MMS messages, email activity, blogging entries and status updates on social networking sites ('SNSs'; boyd and Ellison, 2007). As far as possible, the following details are recorded in each of these cases (to be stored as forms of metadata in the corpus):

- Date and time sent/ received
- Identity of sender/receiver (age, occupation etc.)
- Location when sent/received
- Activity in location
- Content of message

The integration of this additional information starts to allow us to analyse aspects of situational language variation (Tagg, 2009), in order to make some informed observations about, for example, the sort of language used in certain digital mediums at particular times of the day and in specific locations. In other words, this (meta)data enables then analysis of features associated with the context-specificity of words used, and topics discussed, through specific digital mediums, in given communicative environs.

Building on this, the data collected for the 'language in the wild' strand of DReSS focuses more on attempting to capture the linguistic experience of specific language individuals on a *day-to-day* basis; incorporating not only text-based records, but also video, audio, field notes and so on. It is perceived that the systematic analysis of such datasets, as digital corpora, will enable a more detailed investigation of the interface between a variety of different communicative modes from an individual's perspective, tracking a specific person's (inter)actions over time (i.e. across an hour, day or even week).

3. Corpus software requirements

Given that NeLC is comprised of multiple forms of varying media types, there are lots of issues to be addressed regarding the optimum ways in which these are recorded, processed, stored and accessed/interrogated by the linguist. The methods employed at each of these stages differ from each media type because they are naturally stored in a variety of file formats, and are typically visualised and represented in different ways.

It has long been understood that 'time and fiscal constraints, as well as the traditions of different research communities make it impossible to adopt a single standard for all corpora' (Strassel and Cole, 2006: 2, a fact also explored by Lapadat and Lindsay, 1999). This is true not only for the methodological or practical procedures used to collect and compile the data, but also, to some extent, the question of how to represent corpora for future interrogation.

Consequently, current mono-modal, as with *developing* multi-modal corpora, are somewhat bespoke insofar as they are commonly designed and constructed in 'light of the investigators goals' (Cameron, 2001: 29, also see Knight et al., 2009; Lapadat and Lindsay, 1999; O'Connell and Kowal, 1999: 112; Reppen and Simpson, 2002: 93 and Roberts, 2006), that is, in order to meet a given research need and/or to allow users to focus on specific features of spoken or written language. Therefore, procedures and standards used even in multi-modal corpus collection and compilation need to be refined and somewhat redefined in order to accommodate more varied, heterogeneous datasets elicited from multi-modal and ubiquitous resources.

Of principle concern to datasets of a heterogeneous nature, as was the case for early forms of multi-modal corpora comprised of text, audio and video data; is how these multiple forms of media are synchronised and aligned. It is important, for example (in the case of the 'day in the life' data), for the user to have the ability to examine specific words spoken, gestures or actions (i.e. what they are doing) of participants at a particular time or in a given geographical locations. The natural advantage, then, with utilising video and audio files, is that they are already time-based so theoretically the synchronicity of these media files can be carried out with relative ease. However, the synchonisity of different media types poses a bigger challenge (discussed later in this paper).

To some extent this natural time-based alignment of data is also true for text-based digital SMS messages and e-mails, as these are commonly logged as being sent or received at a particular time and on a particular date. In addition to this, text-based 'time-stamps' can be added into transcriptions of speech to allow the time alignment of these additional records. Time-stamps are currently administered manually, either at a turn-by-turn basis, or if a more finite level of alignment is required (perhaps for detailed phonetic transcriptions to be undertaken when analysing the data), they can be administered on a word-by-word basis. However it is particularly difficult to do the latter manually, since each single word needs to be assigned a time code in turn. This means that it is unlikely that large quantities of data can be processed easily or with speed, in such a way. Given this, the former of these approaches is perhaps more appropriate to use, especially when dealing with large sets of dyadic or multi-party talk. In the case of single speaker monologues, the insertion of time-stamps are determined by the discretion of the analyst, so are perhaps placed at lengthy pauses and/or at natural breaks in the speech.

Once collected and synchronised, it is vital to consider how to represent the data, and how to enable the future interrogation and analysis of the corpora by the user. At the onset of the DReSS project, the following utilities were defined as being critical to integrate into a ubuiqtous corpus tool:

- The ability to search data and metadata in a principled and specific way (encoded and/or transcribed text-based data), within and/or across the three global domains of data; devices/ data type(s), time and/or location and participants/ given contributions.
- Tools that allow for the frequency profiling of events/ elements within and across domains (providing raw counts, basic statistical analysis tools, and methods of graphing such).
- Variability in the provisions for transcription, and the ability for representing simultaneous speech and speaker overlaps (for example).
- New methods for drilling into the data, through mining specific relationships within and between domain(s). This may be comparable to current social networking software, mind maps or more topologically based methods.
- Graphing tools for mapping the incidence of words or events, for example, over time and for comparing sub-corpora and domain specific characteristics.

4. DRS

The Digital Replay System (DRS[1], see French et al., 2006), software which is being built as part of the DReSS research project, seeks to provide users with these utilities (and a range of others, as discussed below). Digital Replay System (DRS) is an open source, cross platform suite of multi-purpose qualitative analysis tools, designed to support exploration of heterogeneous data sets.

Developed with a strong focus on corpus linguistics (see Knight et al., 2006), it provides several tools with a direct application for multi modal corpus examination. At its core are three functions: corpus management, corpus filtering, and corpus visualisation. A screenshot of the DRS software can be seen in figure 1.

[1] The Digital Replay System software is available to download for free. For downloads links and further information on the system, including publications, user guides and demonstrations please see the following: http://www.mrl.nott.ac.uk/research/projects/dress/software/DRS/Home.html

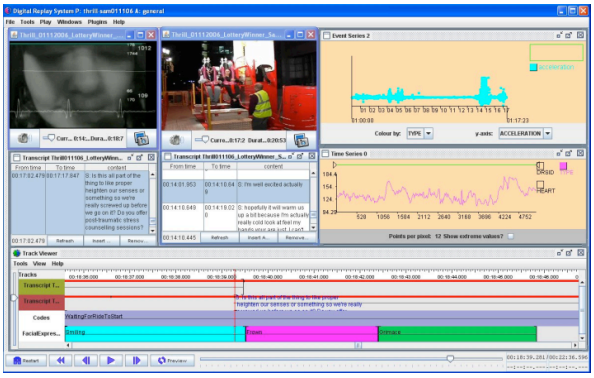


Figure 1: The Digital Replay System (DRS) interface. In this screenshot we see two synchronised videos, the transcripts of those videos (also synchronised), two graphs showing physical (acceleration) and biometric (heart rate) data and finally a track-viewer which supports the coding system, as well as showing representations of the transcripts and videos.

DRS is designed to integrate and synchronise several types of media including video, audio, images, transcriptions, databases, system logs and location data. It is in the combination and annotation of these diverse media that users are able to explore stored multimedia and multimodal corpora.

5. Corpus Management

DRS provides a method of storing and sorting several corpora simultaneously. It features a project-based system, where all the files associated with a particular corpus may be stored in a "project." Subsets of that media which have a temporal connection appropriate for synchronisation can be stored as time sensitive "analyses." Once a set of media is stored in a project, these data can then be annotated filtered and visualised elsewhere in the system.

DRS offers a client-server system, whereby projects can be uploaded to a centralised workgroup server and access can be given to a group of users to edit or view that project. An example project might contain several videos and audio clips, complete with their associated transcripts; annotations and meta-data; a coding scheme and the set of codes associated with each video/audio clip; Location data captured by GPS; coded working documents and one or more databases of additional data. All these data are stored in an external database, allowing for scalability over fairly large corpora.

DRS provides the utility for performing some basic searches of the corpora across all specific words, phrases, extralinguistic and metadata information, across particular domains or sub-categories of search domains (i.e. across devices and/or data type(s), time and/or location and participants/ contributions). Plans exist to ensure that DRS enables the user to have the flexibility to define these atoms, enabling them to be added and omitted, with ties being blocked and unblocked as required. This is to ensure the maximum usability of this tool. The 'global domains' of the DReSS corpora are depicted in figure 2.

6. Corpus filtering

The typical approach to the detailed analysis of corpus data primarily involves profiling the rate-of-use of given search terms and phrases. This basic statistical information is of use for conducting preliminary frequency profiling analyses. DRS is equipped with the basic provisions for conducting these searches, providing a more enhanced, multi-modal version of most contemporary concordancers.

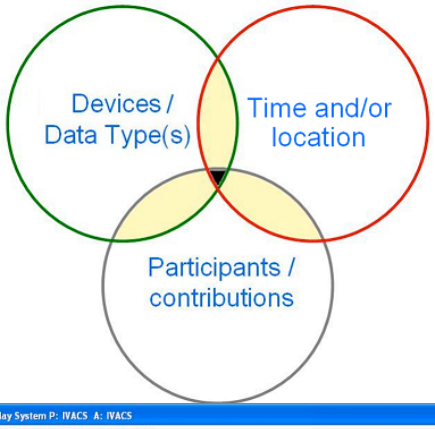


Figure 2: The global domains of digital corpora.

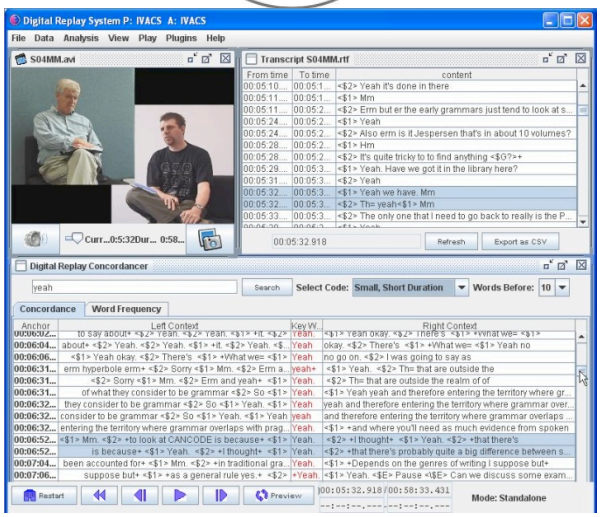


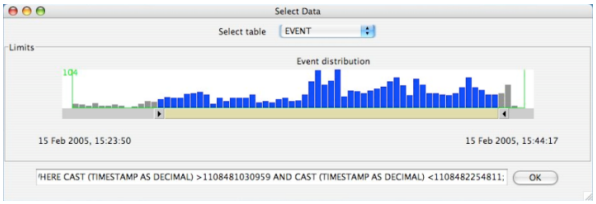
Figure 3 –The DRS conocordancer in use. At the top left we have a video of a conversation, at the top right we see a transcript of that conversation and at the bottom we have concordance lines containing the word "yeah." When the concordance lines are selected, the video and transcript jump to the correct time providing additional context to the concordance.

DRS has an integrated *text workbench*, allowing various methods of filtering corpora of text. The first of these is a concordance tool. In its most basic form, this simply searches through the utterances looking for instances of a particular word of phrase and providing the preceding and proceeding context. However, because of the

heterogeneous nature of the data storable within DRS, this context can be more than just the surrounding words. A multimodal concordance line can provide an instant view of the part of a video/audio clip where a given utterance was transcribed from; a location on a map (since geographical, location based data comprises an important part of the 'A day in the life of your language' component of the DReSS II data); any data associated with that time or person. This much richer form of context allows for a much richer filtering. Figure 3 shows the DRS concordancer.

As an example we might perform a concordance on the word "meeting," but then filter the output only to show instances where we know that person was at work. The concordance tool can also be used on non-verbal data such as codes: so as an example if the corpus contains a coding track describing gestures, it is possible to search for those gestures and see the verbal (and non-verbal) context in which those gestures were used, and of course the same filtering tools are available to us.

In practice DRS uses a graphical interface to construct Standard Query Language (SQL) statements to filter data stored within a database. This might take the form of filtering text by a particular string or logical combination of strings, or it may be more esoteric such as selecting a given location on a map. Numerical filters may also be used if the associated data or metadata contains a numeric value. As an example we might be searching for concordances of the word cold when some temperature sensor shows a temperature below zero degrees. The possibilities for sophisticated filtering are limited only by



the richness of the stored data and the time spent annotating and coding that data. Figure 3 shows an example of this interface.

Figure 3 – An example of the Graphical SQL builder in use. In this case we are filtering the data by time. The graph shows the frequency of events in a given time period, and a double ended slider allows us to easily limit that time period.

DRS provides a word collocation tool, which employs the same precursory filtering system as the concordance tool. Once filtered, tables can be displayed showing the frequency counts of the words surrounding it. As with the concordance tool, a rich view of context can further enhance the usefulness of this tool, as this allows a user to explore the way in which certain words may be interchangeable but only under certain circumstances.

DRS also offers a simple word/code frequency tool, which is similarly subject to the same filtering system as the concordances and collocation tools.

7. Corpus visualisation

One key feature of DRS is the ability to create visualisations of your data. A suite of general purpose synchronised visualisation tools is available to achieve this. Based on earlier work in Morrison et al. (2006), DRS offers a series of graphical representations of your data. These consist of Event Series, time series, histograms, bar charts, pie charts, scatter plots, correlation charts and map-displays. The system is also extensible allowing additional bespoke visualisations to be created for specific purposes. Each of these graphing components can be configured to take any appropriate dimension of a given table of data and apply that to any available visualization dimension. Figures 4 and 5 show examples of these visualisations together with explanations of how they can be used in a coordinated fashion.

A system of brushing and linking (Becker and Cleveland, 1987) has been implemented to allow complex selection and filtration by using combinations of visualisations; that is a selection made in one graph is instantly reflected on each of the others, so for example an event series may show speakers on the Y-Axis, Time on the X-Axis and plot instances of a particular word. If we combine this with a spatial distribution of words, we can then select an individual speaker on the event series and see just that speaker's use of the words by location on the map, or in reverse we can select a particular location on the map and see on the event series which utteraces occurred in that place. A similar example from a corpus of SMS messages can be seen in figures 4 and 5.

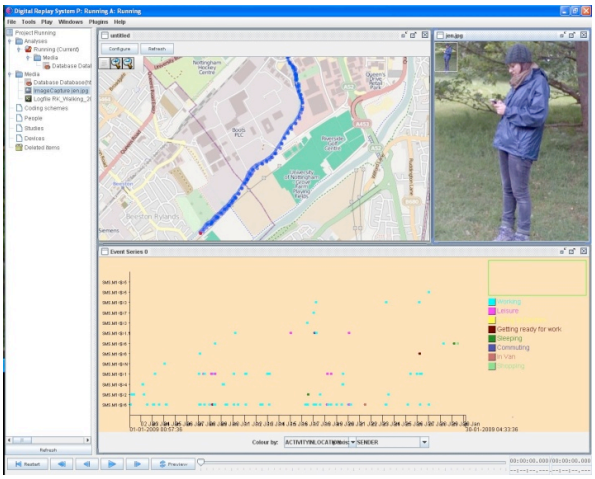
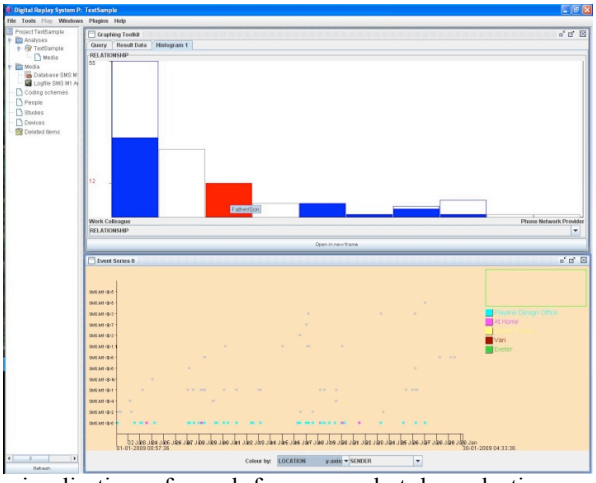


Figure 4: The DRS geographical mapping utility. In this image we see a GPS trace of a user with a red dot indicating when an SMS message was sent. We have an image of the message being sent, which is correctly embedded in the timeline to provide additional context, and we have an overview of the database of messages sent over time shown on an event series.

In other words, this map would be fully interactive insofar as users should be able to select a specific part of the map in order to view the accompanying video and/or transcript (where present) and to investigate, for example, specific patterns of language use in given contexts. This will allow for a micro-level analysis of the corpora, allowing users to mine/ search the corpora according to a specific word, phrase, tag or code. While this functionality is not available with the current release of

DRS, there are possible plans to integrate it in the future.

Similarly the histogram gives us a good general



visualisation of word frequency, but by selecting an individual speaker or place, we can see what percentage of that frequency chart applies to that person, as seen in figure 5.

Figure 5: Coordinated Data Visualisation. In this image we see an event series showing an overview of SMS messages sent with sender on the y-axis, time on the x-axis and coloured by location. The histogram shows the number of messages sent to each receiver. On the event series, a single user has been selected (with all unselected events greyed out) and that selection of events is reflected in the histogram, showing us what percentage of the total texts to each receiver was sent by the selected user.

As a further example, if we don't have actual location information, but rather a more general coding of location, for example "home," "work," "gym," "pub" etc. We can display the percentages of these on a pie chart, then by selecting a given word, we can see at a glance the frequency with which it occurs in those locations.

DRS also provides a scatterplot clustering tool, which allows us to lay data out topographically based on a number of different dimensions. Figure 6 shows an example usage of this tool to reconstruct location data from other available recorded data.

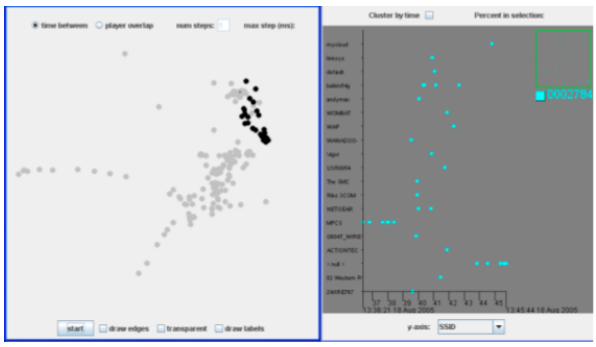


Figure 6 – Reconstructing location data topographically using other available data. In this screenshot we see on the right an event series showing the names of wireless networks a user was connected to, something that can easily be captured both inside and outside on commodity devices such as smart phones. On the left we see a topographical representation of the person's location generated using the time between accesses.

8. Import and export

As the ability to handle legacy data, as well as data in many different formats is critical, DRS provides a sophisticated import tool called the *logfile workbench* which allows users to import data in a number of standard formats such as CSV, TSV etc. However, DRS is not limited to importing and exporting in these standards. A system of extensible parsers is available to allow import and export of data in more specific formats, such as particular XML schemas. In this case, a little knowledge of programming in Java is required, but the actual work is fairly minimal, especially as template parsers are provided. Similarly, for output DRS provides the facility to export any table created by the filtration system to a CSV or TSV which may then be imported into excel or SPSS should more sophisticated statistical analysis be required. At present, the only non-standard third party format directly supported for import/export is Transana (www.transana.org).

9. Positioning with related tools

There are a fair number of tools which provide some of the features available within DRS, both commercial and academic. Examples include ELAN (elan.co.uk), Anvil (anvil-software.de), Transana (transana.org), Studiocode (studiocodegroup.com), Wordsmith (lexically.net), INTERACT (mangold-international.com) and Observer (noldus.com). What sets DRS apart from these others is its ability to handle so called "born-digital" data, that is, system recorded data such as GPS, SMS, sensor data etc. This offers the freedom to create widely heterogeneous data sets and provide incredibly rich contextual information, presented as a usable qualitative resource, as well as offering a flexible system for the inclusion of meta-data. The key factor is coordination. Each tool with in DRS is synchronised with all its counterparts as described in section 8. It is in the coordination of these visualizations that we are able to effectively filter our data in ways previously unfeasible.

10. Summary and conclusion

This paper has discussed some of the issues that arise from the design and representation of heterogeneous datasets for future ubiquitous corpora (for linguistic analysis). By integrating these different datasets within DRS, the linguist is arguably provided with a range of additional measurements of different aspects of text and context; providing a functional platform for generating useful insights into the extent to which everyday language and communicative choices are determined by different spatial, temporal and social contexts.

11. Acknowledgements

The research on which this paper is based is funded by the UK Economic and Social Research Council (ESRC), grant number RES-149-25-1067.

12. References

Becker, R. A. and Cleveland, W. S. (1987). Brushing Scatterplots. *Technometrics, 29*. pp.127-142.

boyd, d. m., and Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13(1).*

Cameron, D. (2001). *Working with spoken discourse.* London: Sage.

French, A., Greenhalgh, C., Crabtree, A., Wright, W., Brundell, B., Hampshire, A. and Rodden, T. (2006). Software Replay Tools for Time-based Social Science Data. In *Proceedings of the 2nd annual international e-Social Science conference, Manchester*.

Knight, D. and Tennent, P. (2008). Introducing DRS (The Digital Replay System): A tool for the future of Corpus Linguistic research and analysis. Proceedings of the 6th Language Resources and Evaluation Conference, Palais des Congrés Mansour Eddahbi, Marrakech, Morocco, 28-30th May 2008.

Knight, D., Bayoumi, S., Mills, S., Crabtree, A., Adolphs, S., Pridmore, T. and Carter, R.A. (2006). Beyond the Text: Construction and Analysis of Multi-Modal Linguistic Corpora. *Proceedings of the 2nd International Conference on e-Social Science, Manchester, 28 - 30 June 2006.*

Knight, D., Evans, D., Carter, R. and Adolphs, S. (2009). Redrafting corpus development methodologies: Blueprints for 3rd generation "multimodal, multimedia'' corpora. Corpora (4), 1: 1-32.

Lapadat, J.C. and Lindsay, A. C. (1999). Transcription in research and practice: from standardisation of technique to interpretative positioning. *Qualitative Inquiry 5*(1), pp.64-86.

Morrison, A, Tennent, P and Chalmers, M. (2006). Coordinated Visualisation Of Video And System Log Data. In *Proceedings of 4th International Conference On Coordinated and Multiple Views in Exploratory Visualization (Cmv2006)*, London, UK.

O'Connell, D.C. and Kowal, S. (1999). Transcription and the issue of standardisation. *Journal of Psycholinguistic research 28*(2), pp.103-120.

Reppen, R. and Simpson, R. (2002). Corpus linguistics. In Schmitt, N. (Ed.) *An Introduction to Applied Linguistics*. London: Arnold. pp.92-111.

Roberts, C. (2006). Part one: issues in transcribing spoken discourse. *Qualitative research methods and transcription* [online academic course]. Kings College London. Available at: http://www.kcl.ac.uk/schools/sspp/ education/research/projects/dataqual.html [Accessed 18 February 2010].

Strassel, S. and Cole, A.W. (2006). Corpus development and publication. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC) 2006* [online]. Available at: http://papers.ldc.upenn.edu/LREC 2006/CorpusDevelopmentAndPublication.pdf [Accessed 18 February 2010].

Tagg, C. (2009). A Corpus Linguistics Study of SMS Text Messaging. PhD Thesis. The University of Birmingham, Birmingham, UK.

A Multimodal Corpus for Studying Dominance in Small Group Conversations

Oya Aran[1], Hayley Hung[1], and Daniel Gatica-Perez[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland
{oya.aran, hhung, gatica}@idiap.ch

Abstract

We present a new multimodal corpus with dominance annotations on small group conversations. We used five-minute non-overlapping slices from a subset of meetings selected from the popular Augmented Multi-party Interaction (AMI) corpus. The total length of the annotated corpus corresponds to 10 hours of meeting data. Each meeting is observed and assessed by three annotators according to their level of perceived dominance. We analyzed the annotations with respect to dominance, status, gender and behavior. The results of the analysis reflect the findings in the social psychology literature on dominance. The described dataset provides an appropriate testbed for automatic dominance analysis.

1. Introduction

Social interaction is a fundamental aspect of human life and is also a key research area in psychology and cognitive science. Although social psychologists have been researching the dimensions of social interaction for decades, the interest on the automatic analysis of social interaction, particularly small group conversations, is quite recent. It is an emerging field of research in several communities such as human computer interaction, machine learning, speech processing, and computer vision (Gatica-Perez, 2009; Pentland, 2005; Vinciarelli et al., 2009) and there is a crucial need for developing dedicated techniques and collecting necessary resources.

The social cues produced and exchanged during an interaction include verbal and nonverbal elements. In parallel to the verbal elements (the spoken words), the nonverbal information is conveyed as wordless messages through aural cues (voice quality, speaking style, intonation) and also through visual cues (gestures, body posture, facial expression, and gaze) (Knapp and Hall, 2009). These cues can be used to predict human behavior, personality, and social relations, in a very wide range of situations. It has been shown that, in many social situations, humans can correctly interpret nonverbal cues and can predict behavioral outcomes with high accuracy, when exposed to short segments or "thin slices" of expressive behavior (Ambady and Rosenthal, 1992). The length of these thin slices can change from a few seconds to several minutes depending on different situations.

Dominance is one of the fundamental dimensions of social interaction. It is signaled via both verbal and nonverbal cues. The nonverbal cues include vocalic ones such as speaking time (Schmid-Mast, 2002), loudness, pitch, vocal control (Dunbar and Burgoon, 2005b), turns, and interruptions (Smith-Lovin and Brody, 1989); and kinesic ones such as gesturing, posture, facial expressions, and gaze (Dovidio and Ellyson, 1982; Dunbar and Burgoon, 2005a). Dominant people are in general more active both vocally and kinesically, with an impression of relaxation and confidence (Hall et al., 2005; Burgoon and Dunbar, 2006). It has been shown that, they also have a higher visual dominance ratio (looking-while speaking to looking-while-listening ratio), i.e. they look at others more while speaking and less while listening (Dovidio and Ellyson, 1982).

There are a number of works in the literature that investigate techniques for the automatic estimation of dominance in small group conversations through nonverbal cues. (Rienks and Heylen, 2005) addressed the classification of participants dominance level (high, normal, low) and used a supervised approach based on Support Vector Machines with manually annotated audio nonverbal features. In (Jayagopi et al., 2009) a large number of automatically extracted nonverbal audio and visual activity cues were used to estimate the most dominant and least dominant participant. The difference in estimating the two dimensions of social verticality, dominance and status, is addressed in (Jayagopi et al., 2008). In (Hung et al., 2008), the authors investigated the use of visual attention cues for estimating dominance. A recent survey on the topic can be found in (Gatica-Perez, 2009). These initial works investigate the different features and models for the estimation of dominance. However for further advancement, there is a clear need for a large database, that can be used as a benchmark across different studies.

In this paper we present a new annotated multimodal dataset that can be used to assess dominance on small group conversations. The novelty of this dataset comes from the dominance annotations as the AMI meeting corpus is well known. In Section 2, we briefly describe the AMI meeting corpus. Section 3 details the dominance annotations and the experimental protocol. In Section 4, we present the resulting datasets and the estimation tasks. The detailed analysis of the annotations is given in Section 5.

2. Meeting Corpus

We use a subset of the publicly available Augmented Multiparty Interaction (AMI) corpus for this study (Carletta et al., 2005). The AMI meeting corpus includes two types of meetings: scenario meetings and non-scenario meetings. In scenario meetings, participants are given the task of designing a remote control over a series of sessions with roles assigned for each participant. One of the participants is the

project manager who has the overall responsibility. These meetings are generally based on presentations followed by discussions. The participants are not always seated. It is common that one of the participants is presenting in front of the whiteboard or slide screen. In non-scenario meetings participants were free to choose their own topic beforehand. Participants are generally seated in these meetings. Each meeting has four participants.

Meetings in the AMI corpus were carried out in a multisensor meeting room as shown in Figure 1. The room contains a table for four participants, a slide screen, and a white board. The audio is recorded via several microphones: a circular microphone array placed on the table, another one with four microphones placed in the ceiling, headset and lapel microphones. The video is recorded via seven cameras: Three cameras mounted on the sides and back of the room capture mid range and global views, respectively; four cameras mounted on the table capture individual visual activity only. Example screen shots from the corpus, from each of the cameras are shown in Figure 2.

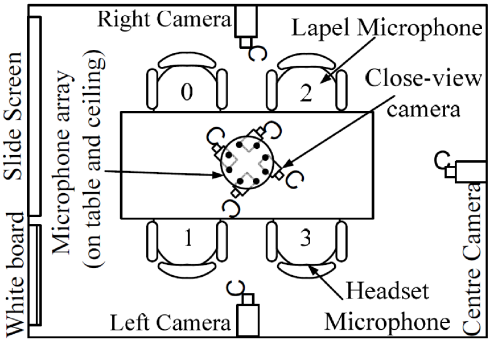


Figure 1: AMI meeting room setup.

Figure 2: AMI screen shots from seven available cameras. The top row shows the views from the right, center, and left cameras. The bottom row shows the views from the close up cameras.

3. Dominance Annotations

We collected a set of annotations on a subset of the meetings selected from the AMI corpus. We follow the "thin slice" approach and use five-minute meeting segments. Previous publications on dominance estimation on AMI corpus use a dataset that corresponds to 4.5 hours of recordings (Jayagopi et al., 2008; Jayagopi et al., 2009). We have enlarged the previously annotated data with a new set of annotations. With this new set, we double the size of our annotated dataset, which corresponds to more than 10 hours of recordings.

3.1. Annotation Questionnaire

The questionnaire asks the annotators about their perceived dominance of the meeting participants. There are two sections in the questionnaire: In the first section the annotators were asked questions on the participants' relative dominance; and in the second section, the questions are focused on evaluating each participant independently:

Dominance ranking: Each participant is ranked from 1 to 4, with 1 representing the most dominant person, and 4 representing the least dominant person in the meeting.

Dominance weight: 10 points are distributed among the participants reflecting annotator's impression of their relative dominance displayed during the meeting. More units signified higher dominance.

Confidence: To identify segments where the rankings were difficult to allocate, annotators were asked about their confidence in their rankings on a seven-point scale.

Participant characteristics: Annotators were requested to evaluate five specific characteristics of each participant independently: dominance (dominant/submissive), status (high/low), aggressiveness (aggressive/meek), dynamism (dynamic/passive), and talkativeness (talkative/silent), also on a seven-point scale. These questions were selected from social psychology work (Dunbar and Burgoon, 2005a).

3.2. Annotator Agreement

For each meeting segment, three annotators ranked the participants according to their level of perceived dominance. We then assessed the agreement between the three annotators for each meeting. If all annotators ranked the same participant as the highest (resp. lowest), we assume there is a full agreement on the most (resp. least) dominant person. If at least two annotators ranked the same participant as the highest (resp. lowest), we assume there is a majority agreement on the most (resp. least) dominant person. Following this procedure we obtained two annotated meeting datasets:

Meeting Set 1 (M1) The initial set of annotations is done on a total of 58 five-minute meeting segments with 21 independent annotators. The meetings were selected from the scenario meetings in AMI corpus. This set was previously used in several publications on automatic dominance estimation (Jayagopi et al., 2008; Hung et al., 2008; Jayagopi et al., 2009).

Meeting Set 2 (M2) We collected a new set of annotations with a completely new set of annotators. 21 annotators annotated a total of 67 five-minute meetings. The meetings were selected both from the scenario and non-scenario AMI meetings. Special care was taken to select segments where all participants were seated during the whole meeting.

Figure 3 shows the agreement statistics in M1 and M2 sets. The bars show the percentage of each type of agreement; three annotators agree (red/bottom), two annotators agree (green/middle), and no agreement (blue/top). The actual

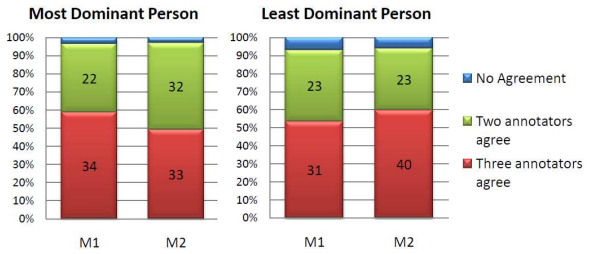


Figure 3: Distribution of agreement types in M1 and M2: Three annotators agree (red/bottom), two annotators agree (green/middle), and no agreement (blue/top)

number of meetings for each type of agreement is shown in the middle of the bars. On different meetings and with different sets of annotators, we observe similar agreement statistics: Full agreement is observed on around 50% of the meetings; whereas on almost all meetings, except a few, we observe majority agreement.

4. Experimental Protocol

4.1. Dominance Estimation Task

Following the recent work in (Jayagopi et al., 2009), we define two dominance estimation tasks:

1. Estimating the Most Dominant (MD) person: Among the participants of the meeting, we aim to estimate the most dominant person.

2. Estimating the Least Dominant (LD) Person: Among the participants of the meeting, we aim to estimate the least dominant person.

4.2. Datasets

The number of full and majority agreement meetings for MD and LD tasks for M1 and M2 sets, and also jointly, are summarized in Table 1. On the joint data, we define four datasets based on the dominance estimation tasks and annotator agreements. For each dataset, we also report the average annotator confidence (Conf - 1 being the highest, 7 being the lowest) and the average dominance weight of the agreed person (Weight - 10 being the highest, 1 being the lowest; all adding up to 10), as reported by the annotators:

FMD: **Full** agreement set, **most dominant** person estimation task (Conf: 1.85 - Weight: 4.57),

FLD: **Full** agreement set, **least dominant** person estimation task (Conf: 2.28 - Weight: 1.03)

MMD: **Majority** agreement set, **most dominant** person estimation task (Conf: 2.03 - Weight: 4.18),

MLD: **Majority** agreement set, **least dominant** person estimation task (Conf: 2.59 - Weight: 1.17)

The self-reported confidences show that the annotators gave higher confidence when annotating the most dominant person with respect to the least dominant one, indicating the latter can be a more difficult task. Furthermore, the full agreement datasets have higher self-reported confidence than the majority agreement datasets. The average relative weights assigned by the annotators also show the consistency of the dominance rankings.

	M1 (58)		M2 (67)		M1+M2 (125)	
	Full	Maj	Full	Maj	Full	Maj
MD	34	56	33	65	67	121
LD	31	54	40	63	71	117

Table 1: Number of meetings for tasks MD and LD with full and majority agreement in M1, M2, and jointly. The total number of meetings in each dataset is in brackets.

5. Analysis of Annotations

5.1. Dominance and Status

Dominance and status are two aspects of the vertical dimension of human social interactions. Although related, these two concepts are different: dominance is a personality trait, which can be defined as the ability to control others; on the other hand, status is an achieved quality and does not directly relate to the ability to control (Hall et al., 2005).
In order to investigate this fact, we analyzed the relationship between the project manager, which is the highest status in the AMI meetings, and the dominance annotations. Figure 4 shows the project manager distribution among most/least dominant participants in full and majority agreement datasets (FMD, MMD, FLD, and MLD). It can be seen that only ~50% of the most dominant participants are also a project manager; whereas the number of least dominant participants who are also the project manager is extremely low. This shows the relation and also the difference between the concepts of dominance and status, as stated by social psychology: (1) high status is not a direct indicator of high dominance, (2) high status people are not totally submissive either.

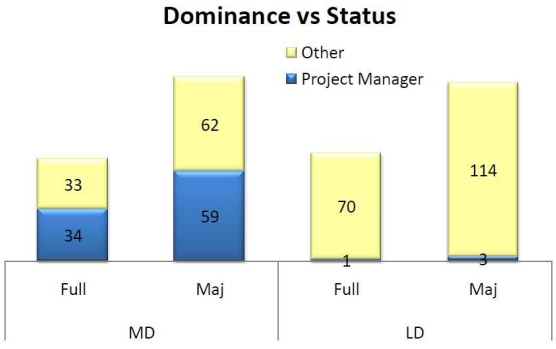


Figure 4: Distribution of project manager among most/least dominant participants in full and majority agreement datasets. Blue/bottom part shows the number of most/least dominant participants who are also the project manager. Yellow/top part shows the number of most/least dominant participants who have other roles than project manager.

5.2. Dominance and Gender

We also investigated our corpus to see the relationships between gender and dominance and gender and status.

Among the total meeting participants, the percentage of females is around 30% (156 females, 344 males). Among the project managers, it is around 50% (56 females, 69 males). We further investigated the distribution of gender for most dominant and least dominant participants. Figure 5 shows the number of males and females for most/least dominant participants in full and majority agreement datasets (FMD, MMD, FLD, and MLD), and also for the project manager (PM). It can be seen that for each case, the percentages of females and males are balanced (Percentages of females in FMD:52%, MMD:55%, FLD:56%, MLD:50%, and PM:45%).

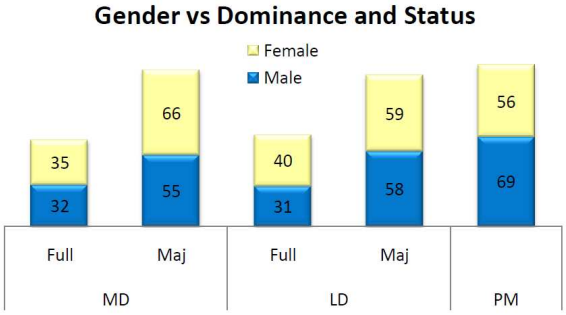


Figure 5: Gender distribution for most/least dominant participants infull and majority agreement datasets, and for project manager. Blue/bottom part shows the number of males and yellow/top part shows the number of females in each case.

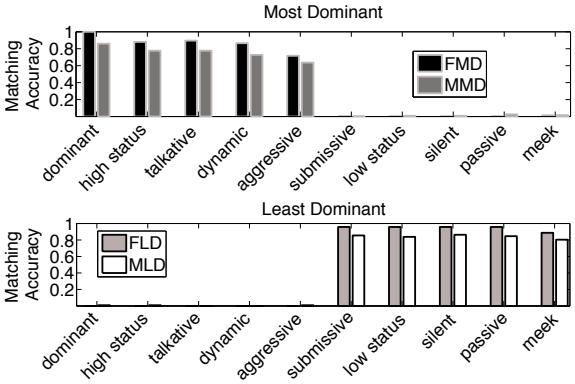


Figure 6: Person matching accuracy of the estimations based on participant evaluations with respect to most/least dominant participants for full an majority agreement datasets.

5.3. Participant Evaluations

We analyzed the participant evaluations in two aspects. The first analysis is based on comparing the participant selected with respect to the highest or lowest evaluation score in the questionnaire (e.g. talkative/silent, aggressive/meek) against the participant selected through the dominance rankings. This analysis aims to identify which of the participant characteristics are more related to perceptions of dominance. The second analysis is based on the correlations of the evaluation scores between the project manager and the most/least dominant person, which shows the relationship between these two concepts in more detail.

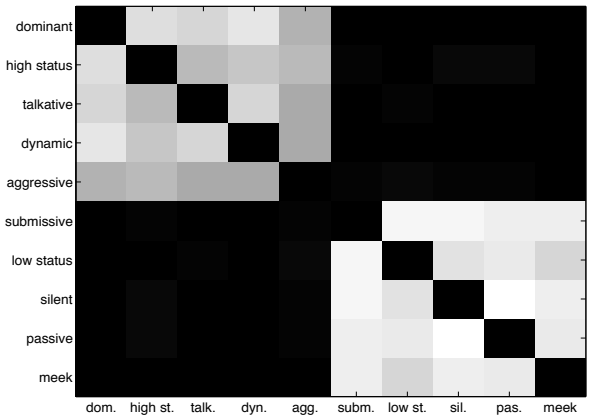


Figure 7: Person matching accuracy of the estimations based on participant evaluations across behaviors. The matrix is symmetric and in grayscale (black:0 and white:1).

For the first analysis, we computed the average of the evaluation scores for each participant by the three annotators for each of the five questions. For each question, based on the evaluation score, we can define two behavior types, one being the extreme opposite of the other. Then we select the most representative participant for each behavior, by choosing the participant with the highest or lowest average evaluation score of the related question. Taking the highest or the lowest value solely depends on which part of the seven-point scale that behavior is placed within the questionnaire. For example, in the question that asks for the dynamism of the participants, if the evaluation score is close to one, it indicates that the person is very dynamic, on the other hand if it is close to seven, it indicates passiveness.

For each behavior and for each dominance task, we count how often the person selected by each behavior was also labeled as the most or least dominant person. Figure 6 shows the person matching accuracies of the estimations based on participant evaluations with respect to the most and least dominant participants. We see that people highly scored as dominant, high status, talkative, dynamic and aggressive are more likely to be selected as most dominant, whereas people scored as submissive, low status, silent, passive and meek are more likely to be selected as the least dominant. Furthermore, Figure 7 shows a pairwise analysis across annotated behaviors. For each pair, we count how often the person selected by one behavior matches the person selected by the other behavior and calculated the person matching accuracy. In general, parallel behaviors highly match each other and low profile behaviors (e.g. silent/passive) have higher accuracies than high profile ones (e.g. dominant/dynamic). In addition, contrasting behaviors do not match at all, with accuracies very close to zero.

For the second analysis, we computed the Pearson correlation of the scores given in the five questions for the project manager and the most/least dominant person (Figure 8). In

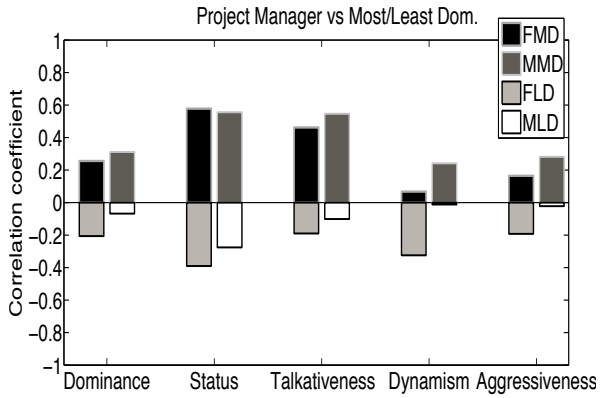


Figure 8: Pearson correlation of the scores of project manager and most/least dominant person for FMD, FLD, MMD, and MLD sets.

general, there is a positive correlation with the most dominant person and a negative correlation with the least dominant person. The highest correlation is observed for status and talkativeness.

6. Conclusion

We have described a multimodal corpus for analyzing dominance in meetings. To our knowledge, this is the first publicly available dataset that combines multiple annotations, rich sensors and multiple annotated tasks. The analysis of the annotations indicate that the annotators are quite consistent within themselves and with each other. Furthermore, the analysis results are consistent with the social psychology findings on dominance.

We believe the presented corpus provides a good testbed and benchmark for the automatic dominance analysis in small group conversations. The AMI meeting corpus is rich in terms of sensors and allows extraction of multimodal nonverbal cues for each participant in the meeting. On the other hand, the new dominance annotations and the identified datasets provide the perceived dominance of the meeting participants, as agreed by multiple annotators. This dataset allows researchers to study the links between multimodal nonverbal cues and dominance perception as well as to assess the performance of the computational models that can be used to estimate dominant and submissive behavior.

The database is available in the following address:

www.idiap.ch/scientific-research/resources/dome/

Acknowledgments: This work was supported by EU FP7 Marie Curie Intra-European Fellowship project Automatic Analysis of Group Conversations via Visual Cues in Non-Verbal Communication (NOVICOM), EU project Augmented Multiparty Interaction with Distant Access (AMIDA), and Swiss National Center for Competence in Research on Interactive Multimodal Information Management (IM2) project. We thank Dinesh B. Jayagopi (Idiap) and the annotators at Bogazici University and at Idiap for their help.

7. References

N. Ambady and R. Rosenthal. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111:256–274.

J. K. Burgoon and N. E. Dunbar. 2006. Nonverbal expressions of dominance and power in human relationships. In V. Manusov and M. Patterson, editors, *The Sage Handbook of Nonverbal Communication*. Sage.

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. 2005. The AMI meeting corpus: A pre-announcement. In *Workshop on Machine Learning for Multimodal Interaction (MLMI'05) Edinburgh, U.K.*, July.

J. F. Dovidio and S. L. Ellyson. 1982. Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening. *Social Psychology Quarterly*, 45(2):106–113, June.

N. E. Dunbar and J. K. Burgoon. 2005a. Measuring nonverbal dominance. In V. Manusov, editor, *The sourcebook of nonverbal measures: Going beyond words*. Erlbaum.

N. E. Dunbar and J. K. Burgoon. 2005b. Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships*, 22(2):207–233.

D. Gatica-Perez. 2009. Automatic nonverbal analysis of social interaction in small groups: a review. *Image and Vision Computing, Special Issue on Human Behavior*, 27(12):1775–1787, December.

J. A. Hall, E. J. Coats, and L. Smith LeBeau. 2005. Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin*, 1313(6):898–924.

H. Hung, D. B. Jayagopi, S. Ba, and and D. Gatica-Perez J.-M. Odobez. 2008. Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *International Conference on Multimodal Interfaces (ICMI'08), Chania, Greece*, October.

D. B. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez. 2008. Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues. In *International Conference on Multimodal Interfaces (ICMI'08), Chania, Greece*, October.

D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. 2009. Modeling dominance in group conversations from nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing, Special Issue on Multimodal Processing for Speech-based Interactions,*, 17(3):501–513, March.

M. L. Knapp and J. A. Hall. 2009. *Nonverbal Communication in Human Interaction*. Wadsworth Publishing, 7 edition.

A. Pentland. 2005. Socially aware computation and communication. *IEEE Computer*, 38(3):33–40.

R.J. Rienks and D.K.J Heylen. 2005. Dominance detection in meetings using easily obtainable features. In *Workshop on Machine Learning for Multimodal Interaction (MLMI'05) Edinburgh, U.K.*, July.

M. Schmid-Mast. 2002. Dominance as expressed and inferred through speaking time: A meta-analysis. *Human Communication Research*, 28(3):420–450, July.

L. Smith-Lovin and C. Brody. 1989. Interruptions in group discussions: The effects of gender and group composition. *American Sociological Review*, 54(3):424–435, June.

A. Vinciarelli, M. Pantic, and H. Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, November.

D64: A Corpus of Richly Recorded Conversational Interaction

C. Oertel, F. Cummins, N. Campbell, J. Edlund, P. Wagner

U. Bielefeld, University College Dublin, Trinity College Dublin, KTH, U. Bielefeld
Corresponding author: nick@tcd.ie

Abstract

Rich non-intrusive recording of a naturalistic conversation was conducted in a domestic setting. Four (sometimes five) participants engaged in lively conversation over two 4-hour sessions on two successive days. Conversation was not directed, and ranged widely over topics both trivial and technical. The entire conversation, on both days, was richly recorded using 7 video cameras, 10 audio microphones, and the registration of 3-D head, torso and arm motion using an Optitrack system. To add liveliness to the conversation, several bottles of wine were consumed during the final two hours of recording. The resulting corpus will be of immediate interest to all researchers interested in studying naturalistic, ethologically situated, conversational interaction.

1. Introduction

The D64 Multimodal Conversational Corpus has been collected to facilitate the quasi-ethological observation of conversational behavior. Conversational interaction in person is a fully embodied activity (Cassell et al., 1999). The role of posture, eye gaze, torso movement, head rotation, hand and arm gestures all contribute to the dynamic establishment, maintenance, and dissolution of domains of joint attention (Baldwin, 1995). Little is currently known about the structure of such transient collaborative domains, and how they might be indexed. However it is clear that the felicitous participation in any natural human-human conversation demands attention to a host of subtle movement cues that permit the ephemeral coupling among participants that constitutes conversational ebb-and-flow.

There is widespread agreement that the empirical investigation of conversational interaction demands multimodal data (Massaro and Beskow, 2002). This is important, both in furthering our understanding of naturally occurring human-human interaction, and in the development of systems that are required to interact in a human-like fashion with human speakers (Edlund et al., 2008). Along with audio recordings, it is now commonplace to include video recordings of at least the faces of conversation participants (van Son et al., 2008). Speech is, however, thoroughly embodied, and unfettered conversational behavior includes appropriate manual gesturing, torso positioning, head direction, gaze behavior, blinking, etc. Furthermore, conversation is often carried out in a dynamic context, with free movement of the participants, change over time in the set of conversational participants, and with an openness that is entirely lacking from most careful studio recordings.

The D64 Multimodal Conversational Corpus has been designed to collect data that transcends many of these limitations. It has been designed to be as close to an ethological observation of conversational behavior as technological demands permit (see also Douglas-Cowie et al., 2007). We first outline the recording setup, the planned model of distribution, and finally, some of our initial aspirations in the analysis of the rich data that results.

Figure 1: The apartment room in which all recording was conducted.

2. Recording

The recording setup chosen for the data collection described is built on the following premises:

[1] The setup ought to be as naturalistic as possible, whereby ”naturalistic” is taken to mean a recording situation that is radically different from a typical laboratory recording, carried out in a recording booth or anechoic chamber with speakers sitting or standing in carefully controlled positions. Instead, a naturalistic recording situation approximates a conversational situation speakers may experience in their daily lives. A scale for different degrees of naturalistic settings is sketched in Fig. 2. The motivation for this decision was to remove as many behavioural artefacts as possible resulting from placing the speakers in laboratory conditions. As laboratory settings are conventionally employed in the hope of removing as many confounding variables as possible, our decision deliberately allows for all kinds of unexpected effects that might influence our data collection.

[2] Unlike most corpus recordings (e.g. map tasks, tourist information scenarios etc.), the chosen setup was not task oriented. No agenda or set of topics was provided. The motivation behind this was to allow the speakers to focus on language use for the purpose of social interaction. In

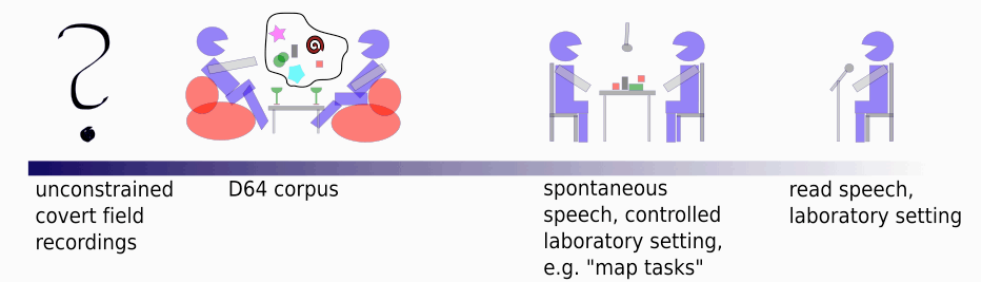


Figure 2: Spectrum of observation scenarios ranging from highly controlled to truly ethological.

task oriented dialogue, the goal of linguistic exchange is the collaborative achieve of a particular goal set by the task, e.g. to receive a particular kind of information or make an appointment. Certainly, social interaction does play an important role in task-oriented dialogue as well, but it is expected to do so to a lesser degree.

[3] Since the speakers knew that they would be recorded and filmed, our setup does not control for the observers paradox (Labov, 1997). However, it has at least the following desirable properties:

- The conversation is interpersonal, with an active and involved other (NOT just a "listener"!);
- It is both social and spontaneous;
- Participants were free to move around, or even leave;
- Speech is unprompted and unscripted;
- Recordings were made over a long period (8 hours over 2 days) thus helping to avoid stereotypical role playing;
- Subjects shared many common interests, and subjective impressions of the interaction were that it was unforced.

Figure 1 shows the domestic apartment room in which all recording was conducted. A mid-sized room with conventional furniture, with a sofa and some comfortable chairs arranged around a low coffee table was employed. Recordings were made over two days, each session being approximately 4 hours long, although the length of the corpus that will ultimately be made available has yet to be precisely determined. The first session was split into two two-hour sessions with an intervening lunch break, while the recording on the second day was continuous over 4 hours. Five participants (the first four authors and a friend) took part on Day 1, and just the 4 authors on Day 2. In order to liven up proceedings somewhat, several bottles of wine were consumed during the latter half of recording on Day 2. Participants were free to stand up, use the adjoining kitchen, change places, etc throughout. In the same spirit, no attempt was made to constrain the topic of conversation, and subject matter varied widely from technological detail (inevitable under the circumstances) to pop culture and politics.

Seven video cameras were employed. There was at least one camera trained on each participant (or one on the sofa as a whole, accommodating two participants). There were also two 360-degree cameras that captured the entire conversational field at a lower resolution. Audio was captured using both mainly wireless microphones (both head-mounted and lapel), along with a variety of strategically placed wide-field microphones. In addition, reflective markers (3 on the head, 1 on each elbow and shoulder and one on the sternum) were monitored by an array of 6 Optitrack cameras.

3. Post-processing

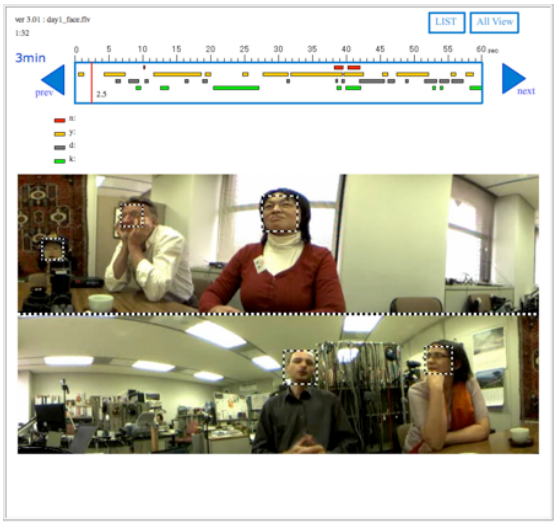


Figure 3: Sample flash interface. Speakers spoken contributions are color coded at the top. Several alternative displays are possible, this one being especially popular. For details, see Campbell and Tabata (2010).

Post-processing of the large amount of data is ongoing, including fragmentation into manageable chunks, cross-channel synchronization, and initial annotation. The data will ultimately be released in clearly indexed chunks of approximately 15 minutes duration, with a transparent indexing by both speaker and source. The entire corpus will be released under the Creative Commons Attribution-Noncommercial-Share Alike License. The raw data is al-

ready available for collaborative annotation, and the aspiration of the project team is that researchers availing of the data will contribute their annotations and other relevant information back to a central repository where others can make use of it. Interested parties may obtain the raw data from the URL www.speech-data.jp/nick/mmx/d64.html, and the project team request that they be informed of any annotation conducted.

As well as the video, audio, and motion-capture data files, the data will be presented in a custom-built flash interface that will allow the user to view, browse, search and export arbitrary subsets of the D64 corpus (Campbell, 2009). The graphical layout will make it particularly easy to search utterance sequences based on dialogue structure and speech overlaps. Each utterance will be accessible by mouse-based interaction. Moving the mouse over a bar will reveal its text, and clicking on the bar will play the speech recording associated with that specific utterance. Each speaker's data will be shown using a different colour to aid identification. Figure 3 shows the kind of flash interface envisaged, as applied to another set of conversational recordings.

Two modes of audio output will be offered for dialogue speech, since it is sometimes preferable to hear a stereo recording, which provides simultaneous natural access to two speakers' overlapping segments, while sometimes it is preferable to hear a monaural recoding, where overlapping speech does not intrude. Separate speech files can be employed in each case. Rapid and more detailed search facilities will ultimately be included. A Join-Play interactive-editing feature will allow the user to simply append the latest utterance segment (video and audio, or audio alone) to a list of related segments to build up a novel data sequence with the speech files and associated text files zipped in a form ready to burn to DVD for wider distribution.

4. The Ebb and Flow of Joint Interaction

Quasi-ethological conversational data of the sort provided by the D64 corpus have not been widely available. With rich capture of visual, audio, and movement data in a naturalistic setting, opportunities arise for the annotation and observation of both quantitative and qualitative aspects of the conversation, in a manner not otherwise possible.

It is our contention that domains of joint interaction arising in a naturalistic conversation are different in an important sense from any joint properties of the participants considered separately. This is graphically illustrated in the well-known experimental work of Murray and Trevarthen (1986) who had mothers and babies interact in real time over a video link. Infants (2 months old) were happy to interact with their mothers through this live link. However, if the infants were shown a prior recording of their mother in interaction with them at an earlier point in time, they objected and immediately withdrew from the exchange. The infants were exquisitely sensitive to the real-time push-and-pull of social interaction, and were not fooled for a moment by a recording that was incapable of responding to their own infantile selves. This work clearly illustrates that there is a meaningful coupling between the mother and infant that is not comparable to the sum of mother+infant.

The task of identifying empirical correlates of this kind of interactional fabric is a daunting challenge. As a first foray into the territory, we propose to attempt to annotate much of the D64 data using two novel quantitative scales that will need to be calibrated and assessed, to see if they may be of use in documenting the ebb and flow of joint interaction. Both variables we will use will initially be based on subjective assessment by trained observers. They will provide subjective ratings of the overall conversational *arousal* and the pairwise *social distance* between participants.

Arousal This variable is hypothesized to index the *joint* arousal of the entire group. Thus, when whole-hearted laughter breaks out all around, for example, we would note a relatively high degree of arousal, while boredom, or indeed silence, would be at the other end of the scale. These examples hide a deal of complexity, however. For example, nervous laughter may reflect a stagnation of the conversation, and thus receive a low arousal rating, and, conversely, a highly pregnant pause may be associated with high joint arousal. For this variable to index a coherent property of the group dynamic as a whole, it is necessary that there be a single conversation, rather than several, relatively independent, conversations. Natural conversation is very fluid, and there is no guarantee that arousal, as envisaged here, will be continuously documentable. Rather, an arousal rating will be provided for successive 5 second frames just in case all participants are mutually engaged.

Social Distance Social distance is a pairwise variable, which is expected to be at a relatively high level for most dyads, most of the time, but to decrease as two participants attend jointly, or engage in reciprocal interaction. We adopt a convention where a low distance value corresponds to relatively intense pairwise interaction, and a high value reflects the perception of greater distance by the annotator. An example of low rated distance would be where two people look at each other, smile at each other and address each other in conversation. The conversation does not need to be of a friendly kind. Two people having an argument would be rated low (close) just as two people confessing their love to each other. Another instance in which social distance would be indexed as relatively low would be when two people display the hallmarks of joint attention, in that they have the same point of focus, look at the same object for a rather long period of time and have the same body posture or move their heads in the same moments. In contrast, a scene in which two participants look in different directions and seem to be interested in different events would be rated as relatively high with respect to social distance.

Both of these proposed scales are highly speculative. It is not yet known whether a sufficiently high-degree of inter-rater reliability will be obtainable, even after considerable refinement of the criteria employed by annotators. Initially, annotators are being asked to base ratings on a combination of such observable characteristics as posture, torso-facing, eye gaze, head rotation, simultaneity of movement, etc. Importantly, annotators are required to use observable characteristics of the scene, and not linguistic interpretation, in their ratings. Ratings are on a scale from 1 to 10, and we freely acknowledge that there will be a period of calibration required in order to arrive at rating guidelines, based on ob-

servables, that lead to a relatively consistent evaluation of the character and dynamic of joint interaction. To bootstrap the process, a selection of extracts from the recordings will be made available on a website, and will be independently rated by at least 10 raters each. Feedback will be obtained about the observable features considered to be of most use, and inter-rater reliability will be assessed using Krippendorff's Alpha (Lin, 1989)

A second line of investigation we have been pursuing is harder to document in a static document, as it involves observation of simultaneous real-time movement of several participants at once. Several quite striking examples of simultaneity of movement of two participants have been observed, and can be viewed at http://tinyurl.com/yk2q34d. Much as spectators at a tennis match can be observed to display head movement locked to the to-and-fro of the tennis ball, so too listeners can be observed to be coupled to the ongoing flow of social interaction. Simultaneous onset of head nodding, whole body turning, etc are readily observable in the data, and are most clearly seen when the observer does not attend to the linguistic content of the ongoing discussion. We have found that the simple expedient of observing the data at a faster rate than normal, with the sound turned down, helps greatly in attending to the embodied participation of participants in the ongoing ebb and flow.

These two examples illustrate both the opportunity for rich observation, and the challenge in documenting conversational interaction as a rich form of human behavior extending far beyond mere linguistic content. The coordination of behavior in conversation has recently been described as *participatory sense-making* within the enactive tradition (De Jaegher and Di Paolo, 2007). In this approach, the process of interaction is seen as the basis for the creation, maintenance and transformation of domains of autonomy. The dimensions of social distance and arousal we have identified above may index the process by which interaction moves from the coordination and sense-making of distinct individuals to the joint process of participatory sense-making.

5. Discussion

Naturalistic data collection on the scale employed here has not hitherto been generally feasible. The utility of such large-scale oversampling will depend, to a great extent, on the usability of the web-based interfaces employed in the dissemination of the corpus. Conversely, with such rich data, it is not possible to anticipate with any certainty the kind of annotation, or the variables annotated, by specific research groups. While we have suggested some novel ways of potentially indexing the dynamics of conversational interaction, our plans here are highly speculative, and the variables, as yet, untested. We hope that the availability of multiple points of view, along with motion capture data, and extensive audio recording, will encourage other groups to consider new and ambitious ways of interpreting conversation in a natural setting.

Acknowledgements

This work has been supported by grants to Nick Campbell from the Visiting Professorships & Fellowships Benefaction Fund from Trinity College Dublin, and the Kaken-B Fund for Advanced Research from the Japanese Ministry of Information, Science & Technology, and also Science Foundation Ireland, Stokes Professorship Award 07/SK/I1218. Jens Edlund is supported by The Swedish Research Council KFI – Grant for large databases (VR 2006-7482). Catharine Oertel is supported by the German BMBF female professors programme (Professorinnenprogramm) awarded to Petra Wagner. Finally, thank you to Nike Stam for her generous participation.

6. References

D.A. Baldwin. 1995. Understanding the link between joint attention and language. *Joint Attention: Its Origins and Role in Development*, pages 131–158.

N. Campbell and A. Tabeta. 2010. A software toolkit for viewing annotated multimodal data interactively over the web. In *Proc. LREC*.

N. Campbell. 2009. Tools & Resources for Visualising Conversational-Speech Interaction. *Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*, page 176.

J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. 1999. Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: the CHI is the limit*, pages 520–527. ACM New York, NY, USA.

H. De Jaegher and E. Di Paolo. 2007. Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences*, 6(4):485–507.

E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J. Martin, L. Devillers, S. Abrilian, A. Batliner, et al. 2007. The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. *Lecture Notes in Computer Science*, 4738:488.

J. Edlund, J. Gustafson, M. Heldner, and A. Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication*, 50(8-9):630–645.

W. Labov. 1997. Some further steps in narrative analysis. *Journal of Narrative & Life History*, 7(1-4):395–415.

LI Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255.

D.W. Massaro and J. Beskow. 2002. Multimodal speech perception: A paradigm for speech science. *Multimodality in Language and Speech Systems*, pages 45–71.

L. Murray and C. Trevarthen. 1986. The infant's role in mother-infant communications. *Journal of Child Language*, 13(1):15–29.

R. van Son, W. Wesseling, E. Sanders, and H. van den Heuvel. 2008. The IFADV corpus: A free dialog video corpus. In *International Conference on Language Resources and Evaluation*.

DYNEMO: A Corpus of dynamic and spontaneous emotional facial expressions

Brigitte MEILLON*, Anna TCHERKASSOF°°, Nadine MANDRAN*, Jean-Michel ADAM*, Michel DUBOIS°°, Damien DUPRE°°, Anne-Marie BENOIT, Anne GUÉRIN-DUGUÉ°, Alice CAPLIER°**

*LIG, Université de Grenoble – CNRS, BP 53 – 38041 Grenoble Cedex 9 - FRANCE

**IEP, UPMF, BP 47 – 38040 Grenoble Cedex 9 - FRANCE

°GIPSA-LAB, UJF, BP 53 – 38041 Grenoble Cedex 9 - FRANCE

°°LIP, UPMF, BP 47 - 38040 Grenoble Cedex 9 - FRANCE

E-mail: brigitte.meillon@imag.fr, anna.tcherkassof@upmf-grenoble.fr, nadine.mandran@imag.fr, jean-michel.adam@upmf-grenoble.fr, michel.dubois@upmf-grenoble.fr, Damien.Dupre@bvra.etu.upmf-grenoble.fr, anne-marie.benoit@iep-grenoble.fr, anne.guerin@gipsa-lab.grenoble-inp.fr, alice.caplier@gipsa-lab.grenoble-inp.fr

Abstract

DynEmo is a publicly available database of significant size containing dynamic and authentic EFE of annoyance, astonishment, boredom, cheerfulness, disgust, fright, curiosity, moved, pride, and shame. All EFEs' affective states are identified both by the expresser and by observers, with all methodological, contextual, etc., elements at the disposal of the scientific community. This database was elaborated by a multidisciplinary team. This multimodal corpus meets psychological, technical and ethical criteria. 358 EFE videos (1 to 15 min. long) of ordinary people (aged from 25 to 65, half women and half men) recorded in natural (but experimental) conditions are associated with 2 types of data: first, the affective state of the expresser (self-reported once the emotional inducing task completed), and second, the timeline of observers' assessments regarding the emotions displayed all along the recording. This timeline allows easy emotion segmentations for any searcher interested in human non verbal behavior analysis.

1. Scientific background

The issue of emotional facial expressions' (EFE) recognition has been extensively investigated. So far, evidence is based for the most part on methods using a static and unnatural material, namely, still photographs of posed facial expressions of emotion (usually, Ekman's stimuli). This kind of methodology raises questions about its ecological validity (Tcherkassof, Bollon, Dubois, Pansu, & Adam, 2007). Indeed, EFE are highly dynamic social signals. The emotion message they display is reflected in facial action patterns (Bould & Morris, 2008). Yet, they have been typically studied as static displays. This is why, even though the central role of the dynamics of facial expressions is endorsed, little is still known about the temporal course of facial expressions. Furthermore, studied EFE exhibit emotions simulated or posed by actors. Yet, the lack of spontaneity and naturalness of this material constitutes a serious objection raised against such studies (Kanade, Cohn, & Tian, 2000). Thus, as it is claimed by many researchers, there is a strong need for systematic research on spontaneous and dynamic facial expressions of emotion, especially in the perspective of an ecological approach of human interactions (Back, Jordan, & Thomas, 2009). This need is even more tenacious that research on this issue is likely to provide new and important evidence which will improve our knowledge on emotions and non verbal behavior.

In order to tackle the question of the recognition of dynamic and spontaneous EFE, one must deal with three issues. First, collect a sizeable quantity of spontaneous stimuli. Second, gather on-line judgments (i.e. dynamic assessments) of the emotions displayed by the stimuli. Third, give an account of these judgments. Recently, several databases had been recorded, such as *Humaine* (Douglas-Cowie, Cox, Martin, Devillers, Cowie, Sneddon, McRorie, Pelachaud, Peters, Lowry, Batliner, Hönig, to appear), *Cohn-Kanade AU-Coded Facial Expression Database* (Cohn, Zlochower, Lien & Kanade, 1999) and the *CMU-PIE Database* (Sim, Baker, & Bsat, 2003), *The MMI face database* (Maat, Sondak,Valstar, Pantic & Gaia, 2004), *The RPI ISL facial expression database* (Tong, Liao & Ji, 2007), *the FABO* (Gunes & Piccardi 2006), *The Karolinska Directed Emotional Faces* (Goeleven, De Raedt, Leyman & Verschuere, 2008) and *The University of Texas Database* (O'Toole, Harms, Snow, Hurst, Pappas, Ayyad & Abdi, 2005), among others. However, to date, no database containing dynamic and authentic EFE in which displayed affective state is identified (both by the sender and by observers) is available (cf. Zeng, Pantic, Roisman, & Huan, 2009, for a review). Indeed, working out such a quality database is a resource-intensive task…

2. Objectives

With this in mind, Dynemo ANR (Agence Nationale pour la Recherche) project (2007 – 2008) aimed at building a publicly available significant corpus of dynamic and spontaneous facial expressions. As most of EFE databases do not identify the displayed emotional state, we intended to provide emotionally characterized data. Thus, we aimed at working out an EFE corpus combined with emotional self-report (of expressers) and on-line assessments (of observers) data.

3. Methodology

Authentic EFE are hard to collect. Thus, the first challenge of the present study was to record natural emotional expressions in realistic conditions. In doing so, we intended to work out different tasks that elicit various emotional states strong enough to be facially displayed, provided the experimental context does not explicitly

evoke the issue of emotion (for the sender not to display voluntary mimics). Second, since emotion is manifested in the course of time, it seemed relevant to collect judgments of observers along this same temporal dimension. Hence, this study also aimed at handling the issue of dynamic judgments of facial expressions of emotion, that is real-time emotional recognition. As far as we know, no existing device allows on-line emotional judgment recording. Another purpose of the present study thus consisted in working out such a device. Finally, collecting dynamic assessments necessitated to elaborate a method accounting for these temporal data.

This project has been leaded by a multidisciplinary team of researchers in psychology, computer scientist, statistician, experimentation instrumentation specialist, and jurist, working together very closely during all the different steps of the project, in order to produce a final corpus, meeting psychological, technical and ethical criteria.

4. Building emotion-inducing material

The first step was to choose a set of emotions and build appropriate tasks to be displayed on a computer screen and supposed to induce the concerned emotional states. This material consisted in specific slide shows of video, pictures, texts, without any action required from the subject, and some more active exercises, some of which materials based on "Wizard of Oz" technology.

This multimedia material has been validated during several weeks of pilot experimentation sessions. Some iterative modifications were performed on the materials depending on the results of these tests, and some emotions have been definitely discarded, because too difficult to induce.

At the end of this first step, the following emotion-inducing tasks have been retained: cheerfulness, disgust, moved, curiosity, boredom, fright, pride, annoyance, astonishment, shame. A neutral task has been added for control. The tasks last from one to 15 minutes according to the eliciting emotion. They are the following:

Cheerfulness: the participant sorts 4 funny TV advertisements according to his/her preferences.

Disgust: the participant is displayed a video report of an undergoing surgery of a firefly larva laid under a woman's scalp.

To be moved: the participant is displayed a slideshow about human misery, based on the contrast between healthy and smiling children and "soldier" children, followed by land mine wounded children, then very poor children, ending by starving children. Operatic music accompanies the slideshow from the beginning to the pictures of poor children. The end of the slideshow remains silent, making it morally more burdensome.

Curiosity: the participant is displayed a slideshow of optical illusions and their explanation followed by a TV commercial for Audi cars, illustrating these optical chimeras.

Boredom: on a white background screen, 8x6 black spots turn red one after the other, from top left to bottom right. The pace between every turning red is each time 40ms slowed down (2s between the two first turning red spots; 20,4s between the two lasting spots).

Fright: the participant is displayed a German TV commercial for an energetic drink in which a car rolls regularly along a winding road, the lens following it (with a soft and low volume music). The car disappears behind a few trees, but the lens continues its move. At the moment and at the place the car is expected to reappear, a zombie face appears suddenly with a shrill scream. Music is replaced by a loud heartbeat.

Pride: the participant is ultimately congratulated for having correctly estimated, as compared (supposedly) to other participants, the capacities of each of 20 recipients displayed without any helpful cue regarding their size.

Annoyance: the participant verbalizes in a loud voice his/her answer regarding the capacity of each of 20 recipients (displayed without any helpful cue regarding their size) to a (fake) defective automatic speech recognition system. The inefficiency of the latter compels the participant to repeat again and again his/her answers.

Astonishment: the participant is displayed an inducing attention biases video for which experimental psychology shows that 75% of viewers discover (most of the time with surprise) they had not noticed a gorilla crossing the scene they were previously displayed while focusing on a counting task.

Shame: the participant is set with a heartbeat chest strap and the experimenter (a woman for male participants; a man for female participants) remains in the room during all the experiment. The participant is get to believe that he/she is physiologically aroused by pictures of naked and/or taking provocative pose persons (women for male participants; men for female participants) thanks to the heartbeat broadcasted from the loudspeaker. The fake heartbeat accelerates when the picture is supposed to be exciting and decreases when it is not (pictures of persons in current situations like riding a bicycle, talking to another person in the office, etc.).

Neutral: a six seconds animated (spots changing color) screen.

5. Emotional induction experimentation

The first experimentation took place in the experimentation platform of Multicom (LIG laboratory), dedicated for behavioral capture of users in situation of interaction. Two rooms were necessary to design this experimentation: the experimentation room, where the encoder (subject expressing the EFE) performed his emotion-inducing task (one task per encoder) and the control room (cf. Fig. 1), where the technical experimenter conducted the experimentation and launched recordings.

Figure 1: The experimenter in the control room

5.1 Technical point of view

In the experimentation room, the screen of a tablet PC was projected on the wall, via a video projector. The encoder was sitting on a chair, in front of a little table, facing the wall, and could use the keyboard and the mouse of the tablet PC, if required by the task. A set of speakers could broadcast the sound of the multimedia material displayed to the encoder, and another speaker could be used by the experimenter in the control room to broadcast some comments needed for specific tasks. Two cameras were hidden in the room, so that the encoder could not guess he was filmed. A Pan Tilt Zoom camera was located under the wall screen, and filmed the whole face of the encoder. Another camera was hidden behind a book in a shelf and filmed the full length. A clip-on microphone was also fixed under the table so that it was possible to listen to the encoder from the control room.

In the control room, three video streams coming from the experimentation room were displayed on several monitors (encoder's zoomed in face, his length full posture, his screen activity), so that the experimenter could make sure the experimentation goes correctly. Depending on the tasks, he could take the control of the encoder's tablet PC and could remotely adjust the main camera, when the encoder was moving on his chair. The three video streams were gathered and mixed in real time (cf. Fig. 2). The resulting stream was recorded on a PC, via a data acquisition board. The video was manually named (e.g. S163_F_FR for the **163**rd encoder, **F**emale, **FR**ight emotion induction) by the experimenter in the control room. Another stream, with the zoomed in face only, was recorded on a DVD support. An experimentation notebook was manually filled in with some comments (e.g. the encoder was chewing a gum) and relevant information (e.g. match of manual video name and track and DVD number).

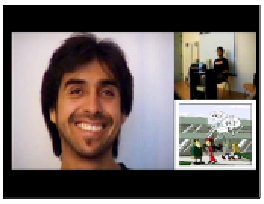


Figure 2: Encoder's zoomed in face, full length and screen activity

5.2 Experimental point of view

The encoders (358 ordinary participants: 182 women and 176 males, from 25 to 65 year old, $\mu = 48$, $\sigma = 9{,}21$) were recruited by an external private society for a study devoted to a visual ergonomic visual task (cover story). It took 8 months to complete this phase of the corpus.

During the experimentation, the encoder was left alone in the room (except for the "shame" induction task), once instructions for the concerned task have been given to him by the experimenter. In order to keep spontaneous facial expressions, and as detailed above, he was covertly videotaped by 2 hidden cameras while carrying an emotion-inducing task.

When the task was over, the encoder filled out a 51 scales questionnaire (6 points) regarding his own emotional state:

- 35 action readiness scales

Ex: *The visual task you just carried out stirs up a tendency to approach, to make contact*

- 4 dimensional scales

Ex: *The state you feel after carrying out this visual task is:*
Unpleasant …….. Pleasant

- 12 emotional labels scales

Ex. How much this visual task made you feel disgusted?
Ex. How much this visual task made you feel annoyed?

The answers of the questionnaire were recorded on line and could be linked to the video recordings, via a specific nomenclature. At the end of the questionnaire, the experimenter came into the room for a debriefing session and made sure that the subject was not psychologically destabilized by the task. Each encoder was given a 15€ coupon for his participation.

5.3 Ethical point of view

This experimentation raised several ethical problems. The first one concerns the respect of the private information regarding the encoders and their image reproduction rights. Indeed, one major stake for databases is to collect agreements of encoders in order to make the data publicly available. Another problem is the one of the hidden cameras: This snare is necessary for scientific and ecological purposes, but must also match deontological and juridical requirements.

That's the reason why we involved a jurist at the beginning of the project and spent time to solve all these problems, before starting the experimentation sessions.

Two consent papers were written and had to be signed by encoders. The first one was given them before the experimentation and asked for their agreement to use the general data issued from the test for research purposes. The second one was proposed to them after the debriefing session and concerned the use of their video to build a database dedicated for researchers only. Encoders were informed that they could go back on their decision at any moment. Few of them refused to sign and, when they did, we instantly destroyed all their materials.

The questionnaire answers were also considered as private data, from a juridical point of view, and have to stay anonymous. That's the reason why we adopted a specific nomenclature for all the files coming from the experimentation, excluding any name, so that it was impossible to associate a posteriori an encoder and his personal data from any computer with Dynemo data.

In the case an encoder goes back to his decision, the only way to locate and destroy his data is to consult the consent papers, kept in a secure place, where the coding nomenclature of his files has been written down, and remove the concerned data from the corpus.

5.4 Data validation point of view

An important validation phase followed this first experimentation. A lot of redundant information has been gathered for each encoder (coded names of the encoders, of the video files, of the questionnaires files, of the manual transcription in the notebook), so that all errors concerning a wrong file identification, or a missing information, could be solved. For example, when discrepancies among different sources were evidenced, the mixed video was visualized and the screen activity could solve the inconsistencies.

During this process, some of the recordings had to be eliminated (less than 10%) because some conditions were not satisfied (e.g. an encoder obviously drunk, a bad posture of the encoder, a bad framing…).

6. Dynamic assessment experimentation

Decoders (171 judging students) assessed the videos via *Oudjat*, a data-processing interface (Tcherkassof *et al.*, 2007). *Oudjat* allows decoders to assess on-line the emotions they perceive in the face of the encoder (cf. Fig. 3). Each video has been assessed by about 20 decoders.

A set of videos were randomly displayed to the decoders. During each video playing, the decoder marked on-line the video each time he perceived the beginning and the end of an emotion displayed by the face by pressing a key. Then, the software replayed the marked sequences and stopped after each sequence, so that the decoder assessed the beforehand marked emotional sequence by selecting one of the 12 proposed emotional terms (cf. Fig. 3).

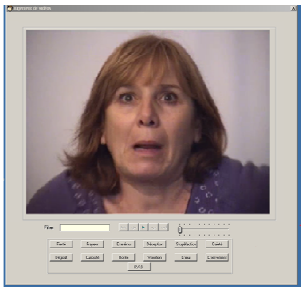


Figure 3: Rating device - Judge selects the emotional label best describing each emotional sequence

Data produced by this experimentation were recorded into text files and included information relative to the video timeline, with the decoder and the video identification, and the time-coded emotional terms for each sequence (cf. Table 1). The processing of these files has been automated in order to make the analysis easier.

Video ID	Induction	Encoder gender	N°decoder	Decoder gender	Detected emotion	Beginning Tme Code	End Time Code
DVD34_2	FR	H	45	F	FR	2567 ms	2873 ms
DVD34_2	FR	H	45	F	DE	3100 ms	3139ms
DVD34_2	FR	H	45	F	EN	3358 ms	3620 ms

Table 1: Example of data produced
First line: Decoder n°45 judging Fright (FR) from time 2567 to 2873 ms of DVD34_2

7. First results

The emotional characterization of the data was undertaken according to two modalities: the encoder's one (emotional self report) and the decoder's one (dynamic assessment).

7.1 Emotional self-report

The questionnaire provides an indicator of the emotional state of the encoder during the task. As shown in figure 4, an encoder generally reports feeling different affective states. This emotional characterization is thus very precise and allows meticulous data categorization. At the moment, only the 12 emotional labels scales have been analyzed, the other 39 emotional items being under process.

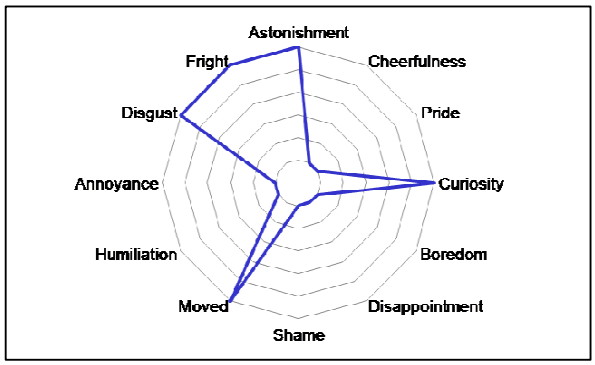


Figure 4: Encoder's self-report on the emotional labels questionnaire while carrying out a disgust-inducing task

7.2 Dynamic assessment

In order to highlight the assessment of the decoders from a dynamic point of view, we calculated each 1/10 second the in-between decoders agreement for each label, during the unfolding of the video.

It provides a timeline (cf. Fig. 5) where the emotional assessments of each decoder are superimposed (mass curves) 1/10 sec. after 1/10 sec. all along the video.

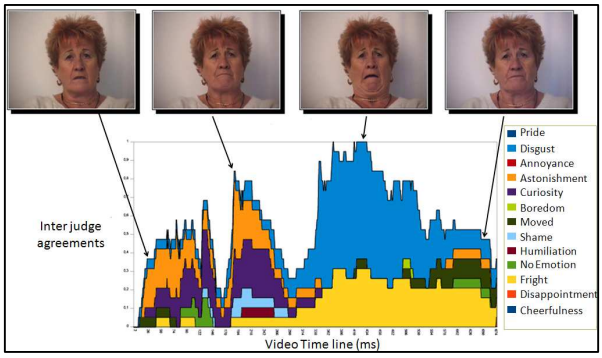


Figure 5: Frames extracted from the video of an encoder who carries out the disgust-induction task and its corresponding emotional expressive timeline underneath

At one glance, one can visually identify when the target emotion is displayed. In the Figure 5 video of a woman

confronted to a disgusting stimulus, 70% of decoders have considered that she was expressing disgust (in blue) from sec. 34 to sec. 54 essentially. It can be noticed that during the same interval, about 30% of decoders have rated her face as expressing fright instead (cf. Fig. 5, yellow mass curve). From sec. 0 to sec. 33, several labels are in competition to describe what is expressed by her face.

8. Corpus quality

Due to the great number of subjects (encoders and decoders) and the multimodality of the data produced by both experimentations (videos, questionnaire answers, encoders categorization, judgment data), each step of this corpus production has been documented in order to keep the history of the data collection and their processing chain.

For each experimentation, detailed procedures have been written down and validated, so that different project members could easily appropriate the method and be able to handle the experimentation. This point carried out the replicability of the experimental session conditions, and consequently, guaranteed the scientific value of the corpus.

Moreover, a web repository, with private access for all the project members, has been initiated to store all materials concerning the project, such as procedures, specification sheets of the software, different automated scripts, emotion-induction material, consent documents, data resulting from the dynamic assessment and questionnaires… This web site was the point of reference all along the project duration, and was updated as soon as we proceeded with our study.

All these methodological points, based on capitalization and traceability, helped us to provide a complete corpus fulfilling quality criteria.

9. Discussion

In order to investigate facial expressions of emotion, DynEmo offers a comprehensive database of videos of dynamic and spontaneous faces. Each video is associated with the expresser's emotional state and the on-line ratings of decoders who assessed all the emotions displayed during the unfolding of the expression (emotional expressive timeline). Thus, the spontaneous and dynamic expressions are characterized very precisely in real time.

Emotional expressive time-lines (cf. Fig. 5) instantly signal, for each video, when the target emotion is displayed by the face. It also signals the periods where observers decode different emotions, that is, when a weak consensus exists between judges regarding what is displayed by the face (ambiguous expression).

The timelines demonstrate that facial expressions of emotions are rarely prototypical and that idiosyncratic characteristics of expressers are often salient elements. Therefore, DynEmo provides an expressive material near to natural social interactions (HHI communications).

10. Conclusion

So far, DynEmo is the most thorough publicly database available of sufficient size (notably for the development of algorithms, cf. Gross, 2005). 358 EFE videos associated with the expressers emotional self-reports are accessible. Out of these, 33 videos have been judged with real-time emotional assessments. The emotional expressive timeline of each of these 33 videos are also obtainable (the remaining videos are actually under judgment process). Thus, the affective characterization of all EFE is available. All methodological, contextual, etc., elements are also at the disposal of the scientific community. Free access to DynEmo (videos of dynamic and spontaneous emotional facial expressions and associated emotional data) are accessible at the following address, powered by Marvelig platform (LIG laboratory) purpose. A consent paper has to be signed to have a login on the site and download the data (https://dynemo.liglab.fr/)

11. Acknowledgements

Authors thank Gwenn BOUSSARD, Laurent ROUZÉ, Étienne BORDET and Mathieu RIVOIRE for their essential participation in this project. DynEmo project was financed by an ANR (-06CORP- 019-01) grant.

12. References

Back, E., Jordan, T, Thomas, S. (2009), The recognition of mental states from dynamic and static facial expressions, Psychology Press, Taylor Francis Group

Bould, E., & Morris, N. (2008). Role of motion signals in recognizing subtle facial expressions of emotion. *British Journal of Psychology*, 99, pp.167--189.

Cohn J.F., Zlochower A. J., Lien J. J & Kanade T. (1998). Feature point tracking by optical flow discriminates subtle differences in facial expression. *IEEE International Conference on Automatic Face and Gesture Recognition*, (pp. 396-401), Nara, Japan.

Douglas-Cowie, E. ,Cox, C., Martin, J-C., Devillers, L., R. Cowie, R., Sneddon, I., McRorie, M., Pelachaud, C., Peters, C., Lowry, O., Batliner, A., F. Hönig, F. (to appear). *The HUMAINE database*. Springer.

Goeleven, E., De Raedt, R., Leyman, L., & Verschuere, B. (2008). The Karolinska Directed Emotional Faces: A validation study. *Cognition & Emotion, 22*(6), 1094-1118.

Gross, R. (2005). Face Databases In S. Li and A. Jain, (Ed.). *Handbook of Face Recognition*. Springer, Verlag.

Gunes, H., & Piccardi, M. (2007). A Bimodal Face and Body Gesture. *Journal of Network and Computer Applications, 30,* 1334-1345.

Kanade, T., Cohn, J.F. & Tian, Y. (2000). Comprehensive Database for facial expression analysis. *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition* (FG'00) (pp. 46--53). Grenoble, France.

Maat, L., Sondak, R., Gaia, P., & Pantic, M. (2004). *MMI face database*. Thesis. Man-Machine Interaction Group, Delft University of Technology, Delft.

O'Toole, J., Harms, J., Snow, S. L., Hurst, D. R., Pappas,

M. R., Ayyad, J. H., & Abdi, H. (2005). Video Database of Moving Faces and People. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(5), 812-817.

Sim, T., Baker, S., & Bsat, M., (2003). The CMU Pose, Illumination, and Expression Database. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*, 1615-1618.

Tcherkassof, A., Bollon,T., Dubois, M., Pansu, P., & Adam, J. M. (2007). Facial expressions of emotions: A methodological contribution to the study of spontaneous and dynamic emotional faces. *European Journal of Social Psychology*, 37, pp.1325-1345.

Tong, Y., Liao, W., & Ji, Q. (2007). Facial Action Unit Recognition by Exploiting Their Dynamics and Semantic Relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(10), 1683-1699.

Zeng, Z., Pantic, M., Roisman, G. I., & Huan, T. S. (2009). A Survey of Affect Recognition Methods: Audio Visual, and Spontaneous Expressions. *IEEE transactions on pattern analysis and machine intelligence,* 31(1), pp.39--59.

A Filipino Multimodal Emotion Database

Jocelynn Cu[1], Merlin Suarez[1], Madelene Sta. Maria[2]
[1]Center for Empathic Human-Computer Interactions, [2]Department of Psychology, De La Salle University
2401 Taft Avenue, Malate, Manila, Philippines
E-mail: {jiji.cu, merlin.suarez}@delasalle.ph, madelene.stamaria@dlsu.edu.ph

Abstract

This paper describes a collection of Filipino expressions of emotions in several modalities. The database is composed of audio recordings of acted speech and videos recordings of spontaneous emotions. The acted emotional speech recordings range from 0.32 to 5.64 seconds, with a total of 10,500 clips collected. Actors and non-actors verbally express the six basic emotions using a set of control texts. The spontaneous emotion video clips range from 3 to 30 seconds, with a total of 400 clips collected from a television reality show. Labelling of spontaneous emotions was done by 20 annotators using discrete Filipino labels. The initial annotation process involves manually labelling the video sequences showing the facial expressions only. Results showed that 59% of the collected clips involve subjects expressing mixed or masked emotions, and 41% of the clips show the subjects expressing discrete emotions. The development of this database is to ensure availability of data for intercultural research on affective computing. Future efforts will focus on finishing the labelling process, adding useful clips to ensure a balanced database, and using the database to train a self-improving affect recognition system.

1. Introduction

Human computing is moving towards human-centered interactions. Instead of the traditional user models that describe cognitive processes and user preferences, human-centered interaction supports user models that can understand and estimate human behaviours like affective and social expressions. Such is the area of empathic computing, which aims to develop intelligent interactive systems with affect capability in a socio-emotional context.

For the past decade, research in empathic computing has been investigating systems that model human affect or synthesize human affect. The successes of these affective systems lie on meticulously designed emotional databases. Examples of such useful databases include that of Douglas-Cowie *et. al.* (2000), Zervas *et. al.* (2003), Zeng *et. al.* (2004), Busso *et. al.* (2004), You *et. al.* (2005), Abrillian *et. al.* (2005), Burkhardt *et. al.* (2005), Barra-Chicote *et. al.* (2008), Chitu *et. al.* (2008), and Gajsek *et. al.* (2009).

However, an affect system that is designed in one language may not work as successfully in another language (Abelin *et. al.*, 2000). According to studies in psychology (Eid and Diener, 2001; Altarriba, Basnight, and Canary, 2003; Matsumoto *et. al.*, 2005), cultural norms and social restraints, as well as interactions between the self and the other, influence how an individual experience and express emotions, especially if it is a negative emotion like anger. In the study conducted by Sta. Maria and Magno (2007), for example, it was found that Filipinos have a total of 34 negative emotions compared to 15 positive emotions that can be experienced with and for another person. This may suggest that negative social emotions are "hypercognized" emotions in the Filipino culture. This suggests that tensions that are felt by the Filipino in his or her social world, which are experienced and expressed through negative emotions, are more elaborated and enhanced compared to other types of emotions.

It has likewise been found that Filipino facial movements may indicate more than just an expression of emotion (Washington, 2007). For example, instead of verbally expressing "yes" or "no", Filipinos would raise or lower their eyebrows; or, instead of pointing to a location, Filipinos would pucker their lips to indicate location. Mapping out these facial movements in the realm of emotion expression will contribute to interpreting communication messages in cultures where nonverbal cues become critical components of interactions. Thus is the motivation to build our own multimodal emotion database, such that expressions of emotion through face, voice, or gestures are interpreted in their proper context, allowing us to develop affective systems with applications in education, health care, and communications.

This paper describes the FilMED, a Filipino multimodal emotion database. Section 2 reviews existing databases and highlights key points raised by experts in building an emotion database. Section 3 presents a detailed description of FilMED, including recording and annotation methods. Finally, we present an evaluation of the corpus and its use.

2. Existing Emotion Databases

Existing researches on affective systems use either acted (Zeng *et. al.*, 2004; Busso *et. al.*, 2004; Chitu *et. al.* 2008) or spontaneous (Douglas-Cowie *et. al.*, 2000; Abrilian *et. al.*, 2005; Gajsek *et. al.* 2009) emotion databases. Acted databases are those that are built through analyses of emotions acted out by a poser; while spontaneous databases are constructed from streams of previously recorded segments of action which depict a given emotion. In an acted database, data is easy to obtain and the content is controlled by the researcher. However, the emotional

expressions are very limited in terms of variations. With spontaneous databases, a wide variety of real-life emotions are captured; but, it is difficult to annotate and interpretation is heavily dependent on the availability of contextual information.

2.1 Affect recognition using acted databases

The Zeng *et. al.* (2004) study on bimodal affect recognition uses an acted audio-visual emotion database taken from 100 subjects, who are mostly students. The subjects were asked to display a specific emotion on request by posing the facial expression and speaking appropriate sentences. Eleven affect categories were used, including seven basic affects (i.e., happiness, sadness, fear, surprise, anger, disgust, and neutral) and four HCI-related affects (i.e., interest, boredom, confusion, and frustration).

The system developed by Busso *et. al.* (2004) was trained on an acted emotion database taken from an actor reading 258 sentences expressing the four emotions (i.e., sadness, happiness, anger and neutral).

Chitu *et. al.* (2008) used an acted emotion database that was recorded from 105 performances of 25 actors, each with approximately 15 minutes of recorded session. Each actor was asked to transpose himself/herself into the correct affective state while reading from a collection of stories that are deemed to carry strong emotional load. Twenty-one emotions were recorded as follows: admiration, amusement, anger, boredom, contempt, desire, disappointment, disgust, dislike, dissatisfaction, fascination, fear, fury, happiness, indignation, interest, pleasant surprise, unpleasant surprise, satisfaction, sadness, and inspiration.

2.2 Affect recognition using spontaneous databases

Douglas-Cowie *et. al.* (2000) specified four guiding principles to ensure the ecological validity of the collected data for spontaneous emotions. These are (1) genuine emotion, (2) emotion in interaction, (3) gradation, and (4) richness. According to their study, it is important that the emotion database should include materials generated by people experiencing genuine emotion, which is effectively derived when one is interacting with another person. This also entails that emotions typically expressed in everyday life could be mixed and may change over time. Such is the basis of the Belfast Database, which is a collection of audio-visual clips featuring 125 English-speaking subjects. The database includes 209 video clips, approximately 10 – 60 seconds per clip, taken from television programs and from studio recordings, showing subjects discussing topics that evoke strong emotions and one-on-one interactions between a field psychologist and the subject. To annotate the database, Douglas-Cowie *et. al.* used dimensional and categorical approaches. For the dimensional labelling, annotators used the FEELTRACE (Cowie *et. al.*, 2000) program to label the clips using the activation-evaluation space. Activation refers to the activity or passivity of the perceived emotion and evaluation reflects how positive or negative the perceived motion is. For the categorical labelling, annotators first labelled the clips using coarse description of emotion, consisting of 16 words, then they used a finer-grained description of emotion, consisting of 24 words. The intensity of the emotion is also recorded in the scale of 1 – 3. Materials are annotated as audio only, video only, and combination of audio and video.

Abrilian *et. al.* (2005) built the French multimodal emotion database with 51 video clips, collected from French TV news interviews on 24 different topics, 48 subjects, and 800 distinct words. Anvil (Kipp, 2001) was used to annotate the database based on context and emotion. Context is annotated using the following attributes: theme, degree of implication, to whom, what for, and causes of emotion. Emotion is annotated through dimensional and categorical labelling, which resulted to 176 emotion labels and later classified into 14 smaller sets of coarse-grained categories and then finally into 6 Ekman classes of basic emotions including the neutral state.

Gajsek *et. al.* (2009) built the Slovenian multimodal emotion database from 19 participants with approximately 30 hours of total recorded session. Annotators used the Transcriber (Barras *et. al.*, 2001) to label the database. Each recording was split into shorter utterances or sentences, and then the annotators freely assign emotion labels anywhere in the clip and for any duration.

3. FilMED Database Design

FilMED is a Filipino multimodal emotion database consisting of acted emotional speech and spontaneous multimodal emotional display.

3.1 FilMED1 – the acted emotional speech database

The emotional database has a total of 10,500 recordings of acted speech. Recorded clips range from 0.32 seconds to 5.64 seconds, for a total of 4.7 hours of recorded speech. A total of 7,000 speech recordings were taken from five actors (3 males and 2 females). The actors were screened to be fluent in Filipino, a resident in the urban area, aged 18 to 25 years old, with no speech defect. The rest of the 3,500 recordings are taken from five non-actors (2 males and 3 females), with the same qualifications as the actors.

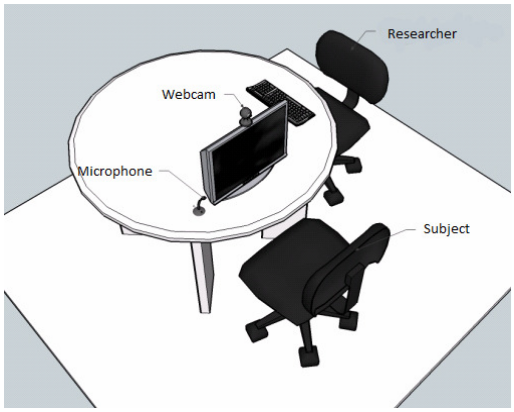


Figure 1: Physical setup of the recording session

Recording was done in a quiet room whose walls were padded with acoustic boards to reduce the effect of echo and noise. The subject was seated in front of a computer monitor, which acts as a prompter, with a webcam and a microphone. The subject was asked to verbally express each text in six basic emotional states identified by Ekman (Filipino labels used are given in parenthesis): *happiness* (kasiyahan), *sadness* (kalungkutan), *anger* (galit), *fear* (takot), *surprise* (gulat), and *disgust* (pandidiri). A seventh state, *neutrality*, was used to establish ground truth. The recording setup is shown in Figure 1.

The control texts, carefully chosen with the help of a linguist, consists of 2-syllabic words, 5-syllabic words, phrases, and sentences. Examples of control texts are:

2-syllabic words:
Kamay (hand)
Pagod (tired)

5-syllabic words:
Ipaliwanag (explain)
Maalikabok (dusty)

Phrases:
Malinis na kapaligiran (clean surrounding)
Patay man o buhay (dead or alive)

Sentences:
Ayaw kong kumain ngayon.
(I don't want to eat now.)
Salubungin natin ang bagong taon.
(Let us welcome the new year.)

For the purpose of modeling, useful prosodic features such as pitch, energy, and formants can be derived using PRAAT (Boersma and Weenink, 2009). An example of a speech vector taken from a recording is shown in Table 1.

Attribute	Value
Pitch Min	144.1619964
Pitch Max	561.7801564
Pitch Ave	273.205088
Energy	0.199735875
F1	1306.512904
F2	2878.744639

Table 1. An example of a speech vector

3.2 FilMED2 – the spontaneous multimodal emotion database

FilMED2 was built following the design principle introduced by Douglas-Cowie *et. al.* (2000) on building a spontaneous multimodal emotion database. The criteria in choosing television programs to be included in the database are: subject is in a realistic situation, emotion is clearly expressed by facial expression and/or voice, frontal facial regions are visible (i.e., left/right eyebrows, left/right eyes, nose, and mouth), and the subject is speaking in Filipino.

The main source for the television clips is the *Pinoy Big Brother Season One Collection*. This television series follows the Philippine version of the reality-TV show Big Brother. It follows the same premise as its foreign counterparts around the world, where 12 residents are forced to live with each other inside a house for 100 days. The house is built with 26 surveillance cameras positioned all over the house including the living room, kitchen, bedroom, and bathroom to monitor the occupants' every move. The choice of using the first season of this type of program is mainly due to its "reality" flavor and "non-actor" participants, in which we believe authentic emotions, is more likely to be displayed and captured on camera. To get a high quality recording of the show, we decided to get copies of the DVD version of the series. The collection is composed of 6 DVD's at standard DVD resolution of 720 x 480 pixels image. Video clips are manually segmented using Adobe Premiere CS4. Segmented clips feature conversations and interactions between two or more subjects, as well as interrogations engaged by the show's host. The clips range from 3 seconds to 30 seconds in length. As of this writing, around 400 clips showing different emotions had already been collected and 300 of which had been annotated. Annotators are screened to be fluent in Filipino, urban residents, and aged 18 to 25 years old. A total of 20 annotators participated in the labelling process. The clips were labelled as video only, audio only, and a combination of audio and video. Coarse-grained discrete Filipino labels are used. An example of the clip is shown in Figure 2.

Figure 2: Same person showing different emotions

Feature Points	x	y
1	82	221
2	137	218
3	109	217
4	109	235
5	34	45
6	63	47
7	48	42
8	48	51
9	48	26
10	63	29
11	97	48
12	127	47
13	112	44
14	112	52
15	97	31
16	112	28

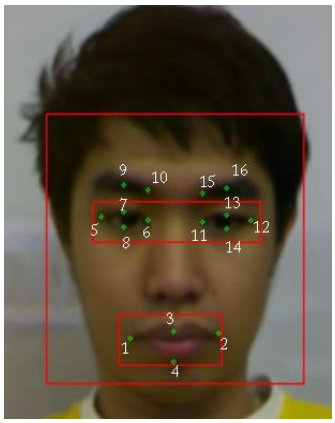


Table 2. An example of facial point vector

From FilMED2, speech vectors can also be derived much like in FilMED1. Furthermore, facial points on the eyes, eyebrows, and mouth can also be extracted to form a facial point vector. The OpenCV library can be used to automatically detect these facial regions and identify the facial points. An example of a facial point vector is shown in Table 2.

4. Evaluating the Database

FilMED1 and FilMED2 were evaluated separately given the different nature of the collected data.

4.1 Evaluation on FilMED1

From the recordings, it can be observed that:

- Neutrality and sadness are characterized by the same drop in pitch and volume. It seems that when the actors were asked to express a neutral emotion through their voice, they tend to sound. It was also noted that non-actors tend to mute their expression especially when they are conscious of being watched or recorded.
- Anger expressions of actors and non-actors were also observed to be consistently characterized by short burst of high energy at the onset of speech. This may suggest that when acted, actors need to exaggerate the emotion to ensure its being recognized and labelled as anger.
- Disgust is difficult for Filipinos to express if the English word label is used. For Filipinos, disgust may be interpreted as *suklam*, *yamot*, *suya*, *sama ng loob*, *rimarim*, at *pandidiri* among others. Not all have direct translations in English and all are scattered across the entire activation-evaluation space evident in the result of a multidimensional scaling technique conducted by Sta. Maria and Magno (2007). This is shown in Figure 3. Disgust, when translated into its Filipino equivalents, can therefore have a range of meanings. This indicates that a distinction needs to be made between lexical and linguistic equivalence (Altarriba *et. al.*, 2003). Translation equivalents do not guarantee equivalence in meanings.
- Happiness, surprise, and fear are similarly characterized by a wide pitch range and higher pitch, but differ in terms of duration of utterances. Further investigation should be carried out to clearly distinguish these three emotions from each other. Additional prosodic features may be needed for better characterization.

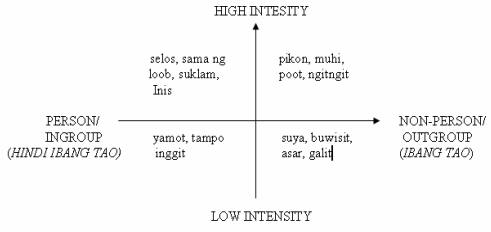


Figure 3. Dimensions of Filipino negative emotions (Sta. Maria and Magno, 2007)

4.2 Evaluation on FilMED2

Of the 400 recordings collected, 300 are manually labelled by 20 annotators so far. A kappa coefficient (Carletta, 1996) cutoff of 0.6 was used to come up with a unanimous label for each clip. The result of the labelling process is presented below:

- 177 clips contain either mixed (e.g., crying but actually happy about something) or masked (e.g., neutral but actually surprised) emotions
- 123 clips are classified as happy (46%), sad (20%), angry (7%), and neutral (27%)

Since the television program simulates the environment where a typical Filipino will be forced to interact in a social group, we believe that the clips were able to capture authentic and spontaneous emotional expressions.

Close inspection of the clips showing mixed or masked emotions suggests that Filipino emotional display is highly influenced by cultural values and traits. For example, the traits *pakikisama* (getting along well with people) and *pakikiramdam* (sensitivity to feelings) are just some of the important facilities ingrained in the Filipino psyche to discourage criticism and maintain good relationship (Gripaldo, 2005). These traits may be why a typical Filipino would automatically mute his or her emotional expression to ensure conformity with the group.

Although happiness is the most amplified among all emotions across all cultures, members of collectivistic cultures are more likely to exhibit increased positive emotions among in-group members than are members of individualistic cultures (Matsumoto *et. al.*, 2005). This may be evident among the Big Brother players who experienced prolonged interactions within a household. Their amplified expression of positive emotions may be characteristic of collectivistic culture members toward their in-group. The increased display of positive emotions, such as happiness, can also further suggest that laughter may be used to temper highly-charged emotional situations.

It was also observed that the kappa coefficient among annotators is highest (i.e., 1.0) if the clips presented to them contain more contextual information. When presented with video only clips, annotators rely heavily on the combined information given out by the subject's facial expression, hand gestures, shoulder movements, and head tilt. When presented with audio only clips, annotators rely equally on the linguistic content of the conversation and the "tone" of voice used by the subject. Moreover, previous research has shown that judgments of another's emotions by members of collectivistic cultures (i.e., Japanese) are influenced by the emotions of the surrounding persons in a given situation (Masuda *et. al.*, 2008). The annotators may have been similarly influenced and allowed more efficient strategies for identifying the emotions displayed on the clippings.

5. Concluding Remarks

This study developed a multimodal emotion database based on Filipino subjects. The annotators are also Filipinos to ensure that cultural and social norms in experiencing and expressing emotions are preserved. FilMED1, with a total of 10,500 audio clips, is collected from actors and non-actors verbally expressing the six basic emotions using a set of control texts. FilMED2, with a running total of around 400 video clips, is collected from a television program showing the subjects in a real-life situation. FilMED2 is manually annotated by 20 annotators and the kappa coefficient is computed to derive a common label for the clip. This database is still growing and annotation of spontaneous clips is still on-going. As shown in the previous section, only 50% of the Ekman basic emotion is currently available in FilMED2. There is still a need to collect data showing fear, surprise, and disgust.

As previously mentioned, Filipino laughter and body language may mean a lot of things and may convey various emotions. FilMED will eventually include collection of clips focusing on Filipino social behaviours like laughter, and the body postures and movements associated with it. The ultimate goal is to ensure that we have enough useful data for intercultural emotion research, development of self-improving multimodal affect recognition systems, user-specific behaviour prediction systems, and empathic agents that can provide appropriate empathic feedback to the user.

6. Acknowledgements

This project is sponsored by the Department of Science and Technology – Philippine Council for Advanced Science and Research (PCASTRD – DOST) under the program entitled "*Towards the Development of a Self-Improving and Ambient Intelligent Emphatic Space: Data-centric, Multimodal Emphatic Modeling from a Pluridisciplinary Perspective*" and the University Research Coordination Office of De La Salle University. Many thanks to the students, faculty, and staff of the Center for Empathic Human-Computer Interactions and the College of Computer Studies for the support and encouragement.

7. References

Abelin, A. and Allwood, J. (2000). Cross Linguistic Interpretation of Emotional Prosody. *In Proceedings of ISCA Workshop on Speech and Emotion*.

Abrilian, S., Devillers, L., Buisine, S., and Martin, J. C. (2005). EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. *In Proceedings of 11th HCI International*, Las Vegas, USA.

Altarriba, J., Basnight, D. M., and Canary, T. M. (2003). Emotion Representation and Perception Across Cultures. In W. J. Lonner, D. L. Dinnel, S. A. Hayes, and D. N. Sattler (Eds.), *Online Readings in Psychology and Culture*, Center for Cross-Cultural Research, Western Washington University, Bellingham, Washington, USA.

Barra-Chicote, R., Montero, J. M., Macias-Guarasa, J., Lufti, S. L., Lucas, J. M., Fernandez-Martinez, F., Dharo, L. F., San-Segundo, R., Ferreiros, J., Cordoba, R., and Pardo, M. (2008). Spanish Expressive Voices: Corpus for Emotion Research in Spanish. *In Proceedings of LREC 2008*.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001) Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production, *Speech Communication*, pp. 5 – 22.

Boersma, P. and Weenink, W. (2009). PRAAT: Doing Phonetics by Computer Vesion 5.1.05 [Computer program], Retrieved June 2009, from http://www.praat.org.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A Database of German Emotional Speech. *In Proceedings of Interspeech 2005*, Lisbon, Portugal.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U., and Narayanan, S. (2004). Analysis of Emotion Recognition Using Facial Expression, Speech, and Multimodal Information. *In Proceedings of 6th ACM International Conference on Multimodal Interfaces*, pp. 205 – 211.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, Vol. 22, No. 2, pp. 249 – 254.

Chitu, A., van Vulpen, M., Takapoui, P., and Rothkrantz, J. M. (2008). Building a Dutch Multimodal Corpus for Emotion Recognition. *Workshop on Corpora for Research on Emotion and Affect*, Vol. 6, pp. 53 – 56.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, S., and Schroder, M. (2000). FEELTRACE: An Instrument for Recording Perceived Emotion in Real Time, *ISCA Workshop on Speech and Emotion*, pp. 19 – 24.

Douglas-Cowie, E., Cowie, R., and Schroder, M. (2000). A New Emotion Database: Considerations, Sources and Scope. *In Proceedings of ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, pp. 39 – 44.

Eid, M. and Diener, E. (2001). Norms for Experiencing Emotions in Different Cultures Inter- and Intranational Differences. *Journal of Personality and Social Psychology*, Vol. 81, No. 5, pp. 869 – 885.

Gajsek, R., Struc, V., and Mihelic, F. (2009). Multimodal Emotional Database: AvID. *Informatica 33*, pp. 101 – 106.

Gripaldo, R. M. ed. (2005). *Filipino Cultural Traits*. Cultural Heritage and Contemporary Change Series IIID, Southeast Asia, Vol. 4.

Kipp, M. (2001). Anvil: A Generic Annotation Tool for Multimodal Dialogue, *Eurospeech*.

Masuda, T., Ellsworth, P., Mesquita, B., Leu, J., Tanida, S., and Van de Veerdonk (2008). Placing the Face in Context: Cultural Differences in the Perception of

Facial Emotion, *Journal of Research and Social Psychology*, 94: 565 – 381.

Matsumoto, D., Yoo, S. H., Hirayam, S., and Petrova, G. (2005). Development and Validation of a Measure of Display Rule Knowledge: The Display Rule Assessment Inventory, *Emotion*, Vol. 5, No. 1, pp. 23 – 40.

Sta. Maria, D. and Magno, C. (2007). Dimensions of Filipino Negative Social Emotions. *In Proceedings of 7th Conference of the Asian Association of Social Psychology*.

Washington, B. D. (2007). Understanding Nonverbal Communication of Filipinos: A Traditional Form of Literacy. *California State University East Bay Online Journal*, CFSJ-CSUEB-2007.

You, M., Chen, C., and Bu, J. (2005). CHAD: A Chinese Affective Database. *LNCS 3784, Springer Berlin*, pp. 542 – 549.

Zeng, Z., Tu, J., Liu, M., Zhang, T., Rizzolo, N., Zhang, Z., Huang, T. S., Roth, D., and Levinson, S. (2004). Bi-modal HCI-related Affect Recognition, *In Proceedings of 6th ACM International Conference on Multimodal Interfaces*, pp. 137 – 143.

Zervas, P., Fakotakis, N., Geourga, I., and Kokkinakis, G. (2003). Greek Emotional Database: Construction and Linguistic Analysis. *In Proceedings of 6th International Conference of Greek Linguistics*, Rethymno.

Automatic Segmentation of Spontaneous Data using Dimensional Labels from Multiple Coders

Mihalis A. Nicolaou[*], **Hatice Gunes**[*,‡] **and Maja Pantic**[*,†]

[*]Department of Computing, Imperial College London, U.K.
(michael.nicolaou08, h.gunes,m.pantic)@imperial.ac.uk

[‡] School of Computing and Communications, University of Technology Sydney, Australia
[†]Faculty of EEMCS, University of Twente, The Netherlands

Abstract

This paper focuses on automatic segmentation of spontaneous data using continuous dimensional labels from multiple coders. It introduces efficient algorithms to the aim of (i) producing ground-truth by maximizing inter-coder agreement, (ii) eliciting the frames or samples that capture the transition *to* and *from* an emotional state, and (iii) automatic segmentation of spontaneous audio-visual data to be used by machine learning techniques that cannot handle unsegmented sequences. As a proof of concept, the algorithms introduced are tested using data annotated in arousal and valence space. However, they can be straightforwardly applied to data annotated in other continuous emotional spaces, such as power and expectation.

1. Introduction

In everyday interactions people exhibit non-basic, subtle and rather complex mental or affective states like thinking, embarrassment or depression (Baron-Cohen and Tead, 2003). Accordingly, a single label (or any small number of discrete classes) may not reflect the complexity of the affective state conveyed by such rich sources of information (Russell, 1980). Hence, a number of researchers advocate the use of dimensional description of human affect, where an affective state is characterized in terms of a number of (continuous) latent dimensions (Russell, 1980),(Scherer, 2000).

Spontaneous data and their dimensional annotations, provided by multiple coders, pose a number of challenges to the field of automatic affect sensing and recognition (Gunes and Pantic, 2010). The first challenge is known as *reliability of ground truth*. In other words, achieving agreement amongst the coders that provide annotations in a dimensional space is very challenging (Zeng et al., 2009). In order to make use of the manual annotations for automatic recognition, most researches take the mean of the coders ratings, or assess the annotations manually. How to best model inter-coder agreement levels for automatic affect analyzers remain mainly unexplored. The second challenge is known as *the baseline problem*: having "a condition to compare against" in order for the automatic recognizer to successfully learn the recognition problem at hand (Gunes and Pantic, 2010). Automatic affect analyzers relying on audio modality obtain such a baseline by segmenting their data based on speaker turns (e.g., (Wollmer, M. and Eyben, F. and Reiter, S. and Schuller, B. and Cox, C. and Douglas-Cowie, E. and Cowie, R., 2008)). For the visual modality the aim is to find a frame in which the subject is expressionless and against which changes in subject's motion, pose, and appearance can be compared. This is usually achieved by constraining the recordings to have the first frame containing a neutral expression. Although expecting expressionless state in spontaneous multicue or multimodal data is a strong and unrealistic constrain, automatic affect analysers *depend* on the existence of such a baseline state (e.g., (Petridis et al., 2009; Gunes and Piccardi, 2009)). Moreover, a number of machine learning techniques such as (coupled) Hidden Markov Models and Hidden-state Conditional Random Fields cannot handle unsegmented sequences, they require the data to have a class label for the entire sequence. To date, many automatic affect recognizers using audio-visual data and utilizing the aforementioned techniques segment their data manually (e.g., (Petridis et al., 2009)).

This paper provides solutions to all of the aforementioned issues. It (i) produces ground-truth by maximizing inter-coder agreement, (ii) elicits the frames or samples that capture the transition *to* and *from* an emotional state (a baseline condition to compare against), and (iii) automatically segments long sequences of spontaneous audio-visual data to be used by machine learning techniques that cannot handle unsegmented sequences.

2. Data

As a proof of concept, the algorithms introduced are tested using data annotated in arousal (how excited or apathetic the emotion is) and valence (how positive or negative the emotion is) space to obtain sequences that contain either positive or negative emotional displays. We use the Sensitive Artificial Listener Database (SAL-DB) (Cowie et al., 2005; Douglas-Cowie et al., 2007) and the SEMAINE Database (SEMAINE-DB) [1] that contain audio-visual spontaneous expressions.

2.1. Data Sets and Annotations

Both for the SAL-DB and the SEMAINE-DB, spontaneous data was collected to the aim of capturing the audio-visual

[1]The Semaine Database: http://semaine-db.eu/

interaction between a human and an avatar with four personalities: Poppy (happy), Obadiah (gloomy), Spike (angry) and Prudence (pragmatic).

The SAL data has been annotated by a set of coders who provided continuous annotations with respect to valence and arousal dimensions using the FeelTrace annotation tool (Cowie et al., 2000; Cowie et al., 2005). Feeltrace allows coders to watch the audio-visual recordings and move their cursor, within the 2-dimensional emotion space (valence and arousal) confined to $[-1, 1]$, to rate their impression about the emotional state of the subject.

For SAL-DB, 27 sessions (audio-visual recordings) from 4 subjects have been annotated. 23 of these sessions were annotated by 4 coders, while the remaining 3 sessions were annotated by 3 coders. The SEMAINE-DB has also been annotated using FeelTrace along five emotional dimensions (valence, arousal, power, expectation and intensity) separately, by (up to) 4 coders.

2.2. Challenges

The time-based operation of Feeltrace presents us with the following challenges: (i) for the sessions coded, there is no one-to-one correspondence between the timestamps of each coder, (ii) throughout the annotation files, there are time intervals where annotations are not available, and (iii) annotations are not (always) synchronized with the audio-visual data stream.

We tackle the first issue by binning the annotations: annotations that correspond to one video frame are grouped together. The second point refers to missing annotations for some sets of frames. This could potentially be due to the following reasons: (i) the coder might not be certain about the annotation for that particular interval, (ii) the coder might release the mouse button for some other reason, (iii) the coders appear to stop annotating when the avatar is talking, and (iv) the CPU load may have an effect on the frequency of measurements being recorded. Finally, the third issue could possibly be due to the following: (i) the response time is expression dependent, i.e., positive expressions are perceived faster and more accurately than negative ones (Alves et al., 2008), and (ii) the lag caused by the CPU load may have an effect on the synchronization between the actual video played and the recording of the annotations.

Table 1: The inter-coder MSE after applying local normalisation procedures: normalizing to a standard deviation of one and a zero mean (GD), normalizing to zero mean (ZA) and no normalisation (NN).

	ZA$_{MSE}$	**GD**$_{MSE}$	**NN**$_{MSE}$
Valence	0.046	0.93	0.072
Arousal	0.0551	0.9873	0.0829

3. Methodology

In this section we address the challenges identified when working with databases annotated in continuous dimensional spaces.

Algorithm 1: Binning the annotations of the coders {set of bins, b} $\leftarrow Binning()$

```
1  //all members of any structures are considered to be zero
2  for each coder file c in the annotation files set do
3      for each annotation a in a coder file c with a timestamp of t do
4          Determine bin b where t ∈ b
5          b.val ← b.val + a.val
6          b.arsl ← b.arsl + a.arsl
7          b.annotCount ← b.annotCount + 1
8      end
9      for all bins b in the set of bins do
10         Average b.val and b.arsl by dividing with b.annotCount
11     end
12 end
```

Algorithm 2: Detecting crossovers in coder annotations: $\{PosCrossOver, NegCrossOver\} \leftarrow$ DetectCrossovers(coder c)

```
1  //bstr is the binned structure, every member is an annotation of A-V values at that
   frame by the specific coder
2  for each f in bstr do
3      if sign(bstr(f).val) ≠ sign(bstr(f − 1).val) then
4          if sign(bstr(f).val) > 0 then
5              Add f to PosCrossOver structure
6          end
7          else
8              if sign(bstr(f).val) < 0 then
9                  Add f to NegCrossOver structure
10             end
11         end
12     end
13 end
```

3.1. Annotation Pre-processing

This process involves determining normalisation procedures and extracting statistics from the data in order to obtain segments with a baseline and high inter-coder agreement.

Binning. Binning refers to grouping and storing the annotations together. As a first step the measurements of each coder c are binned separately. Since we aim at segmenting video files, we generate bins which are equivalent to one video frame f. This is equivalent to a bin of 0.04 seconds (SAL-DB was recorded at a rate of 25 frames/s). The basic binning procedure is illustrated in Algorithm 1. The fields with no annotation are assigned a "not a number" (NaN) identifier.

Normalisation. The arousal and valence (A-V) measurements for each coder are not in total agreement, mostly due to the variance in human coders' perception and interpretation of emotional expressions. Thus, in order to deem the annotations comparable, we need to normalize the data. Similar procedures have been adopted by other works using SAL-DB (e.g. (Wollmer, M. and Eyben, F. and Reiter, S. and Schuller, B. and Cox, C. and Douglas-Cowie, E. and Cowie, R., 2008)).

We experimented with various normalisation techniques. After extracting the videos and inspecting the superimposed ground truth plots, we opted for local normalisation (normalizing each coder file for each session). This helps us avoid propagating noise in cases where one of the coders is in large disagreement with the rest (where a coder has a very low correlation with respect to the rest of the coders). As can be seen from Table 1, locally normalizing to zero mean produces the smallest mean squared error (MSE) both

for valence (0.046) and arousal (0.0551) dimensions. Varying the standard deviation results in values which are outside the range of $[-1, 1]$ and generates more disagreement between coders.

Statistics and Metrics. We extract two useful statistics from the annotations, with a motivation of using them as measures of agreement amongst the annotations provided: correlation (COR) and sign-agreement (SAGR). We start the analysis by constructing vectors of pairs of coders that correspond to each video session, e.g., when we have a video session where four coders have provided annotations, this gives rise to six pairs. For each of these pairs we extract the correlation coefficient between the valence (val) values of each pair, as well as the percentage of sign-agreement in the valence values, which stands for the level of agreement in emotion classification in terms of positive or negative:

$$SAGR(c_i, c_j) = \frac{\sum_{f=0}^{|frames|} e(c_i(f).val, c_j(f).val)}{|frames|} \quad (1)$$

where c_i and c_j represent the pair of coders the sign-agreement metric is calculated for, and $c_i(f).val$ stands for the valence value annotated by coder c_i at frame f. Function e is defined as:

$$e(i, j) = \begin{cases} 1 & if(sign(i) = sign(j)) \\ 0 & else \end{cases}$$

The sign-agreement metric is of high importance for the valence dimension as it determines whether the coders agree on the classification of the emotional state as positive or negative. More specifically, such metrics provide information regarding the perception and annotation behaviour of the coders (i.e., to what degree data is annotated similarly by different coders). In these calculations we do not consider the NaN values to avoid negatively affecting the results.

After these metrics (agreement, correlation) are calculated for each pair, each coder is assigned the average of the results of all pairs that the coder has participated in. In other words, the averaged metric m'_{S,c_j} with respect to coder c_j for a specific metric m (i.e., correlation or agreement) is defined as follows:

$$m'_{S,c_j} = \frac{1}{|S| - 1} \sum_{i \in S, c_i \neq c_j} m(c_i, c_j) \quad (2)$$

where S is the relevant session annotated by $|S|$ number of coders, and each coder annotating S is defined as $c_i \in S$. Essentially, we calculate the averaged level of agreement of coder c_j with respect to the rest, by using the metric m. This is somewhat equivalent to the enumerator of the modified Williams Index, which would be obtained by dividing this enumerator by the averaged level of agreement of all the coders except c_j (Alberola-Lopez et al., 2004). Instead, we obtain the weighted average by using the m' as weights, as shown in line 28 of Algorithm 4. The automatic segmentation process is based on the correlation metric (cor') alone as correlation experimentally proved stricter than sign-agreement in providing better comparison between the coders.

Interpolation. In order to deal with the issue of missing values, similar to other works reporting on data annotated in continuous dimensional spaces (e.g., (Wollmer, M. and Eyben, F. and Reiter, S. and Schuller, B. and Cox, C. and Douglas-Cowie, E. and Cowie, R., 2008)), we interpolated the actual annotations at hand. We used piecewise cubic interpolation as it preserves the monotonicity and the shape of the data.

Algorithm 3: Match crossovers across coders for each session, maximizing the number of coders participating: $\{MatchedCO\} \leftarrow$ MatchCrossOvers($CrossOvers$)

```
1  for Each session s do
2      for i=4 to 2 do
3          //get as many coders as possible to agree (max. 4 and min. 2)
4          for Each crossover co in CrossOvers belonging to s do
5              currentlyMatched ← {co}
6              Find all crossovers co2 in CrossOvers which:
7              - Belong to s
8              - Are from different coders
9              - co2 ≠ co ∧ abs(co2.time − co.time) ≤ 0.5 seconds
10             Add the co2 to currentlyMatched
11             if length(currentlyMatched) = i then
12                 mark all crossovers in currentlyMatched as seen
13                 add currentlyMatched to MatchedCO
14                 remove currentlyMatched from CrossOvers
                   belonging to s
15             end
16         end
17     end
18 end
```

3.2. Automatic Segmentation

The automatic segmentation stage consists of producing negative and positive audio-visual segments with a temporal window that contains an offset before and after (i.e., the baseline) the displayed expression. This process is presented in Algorithm 4 that makes use of Algorithms 2 and 3.

Firstly, we describe the actual time window that the audio-visual segment is supposed to capture. For instance, for capturing negative emotional states, if we assume that the transition *from* non-negative *to* a negative emotional state occurs at time t (in seconds), we then have a window of $[t - 1, t, t', t' + 1]$ where t' seconds is when the emotional state of the subject returns to non-negative. The procedure is analogous for positive emotional states.

Detecting and Matching Crossovers. In Algorithm 2, for an input coder c, the crossing over from one emotional state to the other is detected by examining the valence values and identifying the points where the sign changes. Here a modified version of the sign function is used which returns 1 for values ≥ 0 (a value of 0 valence is never encountered in the annotations), -1 for negative, and 0 for NaN values. Algorithm 2 accumulates all crossover points for each coder, and returns the set of crossovers *to-a-positive* ($PosCrossOver$) and *to-a-negative* ($NegCrossOver$) emotional state. The output is then passed to Algorithm 3.

The goal of Algorithm 3 is to match crossovers across coders. For instance, if a session has annotations from 4 coders, due to synchronization issues discussed previously, the frame (f) where each coder detects the crossover is not the same for all coders (for the session in question). Thus, we have to allow an offset for the matching process.

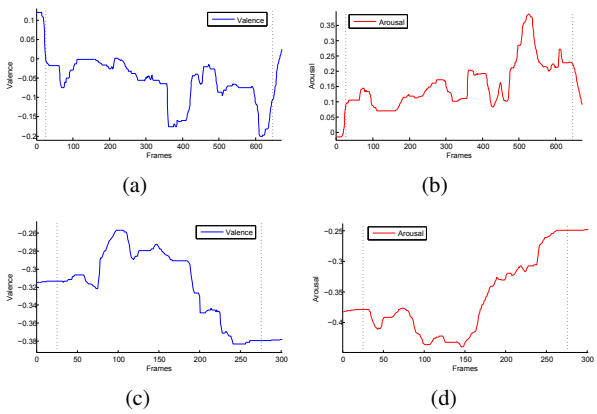


Figure 1: Two examples of interpolated valence ((a),(c)) and arousal ((b),(d)) plots from two individual segments produced by the segmentation procedure.

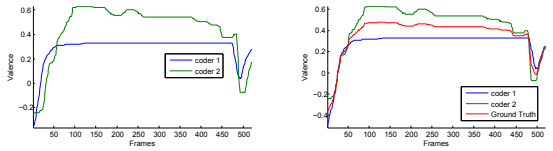


Figure 2: Valence annotations from two coders in SEMAINE-DB before and after applying pre-processing operations.

This procedure searches the crossovers detected by the coders and then accepts the matches where there is less than the pre-defined offset (time) difference between them. When a match is found, we remove the matched crossovers and continue with the rest. The existence of different combinations of crossovers which may match using the predefined offset poses an issue. By examining the available datasets, we decided to maximize the number of coders participating in a matched crossover set rather than minimizing the temporal distances between the participating coders. The motivations for this decision are as follows: (i) if more coders agree on the crossover, the reliability of the ground truth produced will be higher, and (ii) the offset amongst the resulting matches is on average quite small (less than 0.5 secs) when considering only the number of participating coders. Maximising the number of participating coders can simply be achieved by iterating over the entire set of crossovers. This is expressed by the loop beginning in line 2 of Algorithm 3. We disregard cases where only one coder detects a crossover due to lack of agreement between coders.

Segmentation Driven by Matched Crossovers : This procedure (illustrated in Algorithm 4) takes the output of Algorithm 3, and attains the sets of matched crossovers (Algorithm 3, lines 6-7). An iteration for all sets of matched crossovers for *to-Negative* transition is shown starting in line 7. $mcos$, $mcos(i).f$ and $mcos(i).c$ represent the current matched crossover, the frame where the i-th crossover (of the matched crossover) occurred, and the coder who detected the i-th crossover of $mcos$, respectively. $mcos(i).val$ is the vector of valence measurements for coder i participating in $mcos$. The crossover frame decision (for each member of the set) is made in lines 10:17, and the *start frame* of the video segment is decided. In order to capture 1 second before the transition window, the number of frames corresponding to the pre-defined offset are subtracted from the *start frame*. The ground truth values for valence are retrieved in lines 19:30 by incrementing the initial frame number where each crossover was detected by the coders. The procedure of determining combined average values continues until the valence value crosses again to a *non-negative* valence value. The endpoint of the audio-visual segment is then set to the frame including the offset after crossing back to a *non-negative* valence value.

Figure 3: Example frames from an automatically extracted segment from SEMAINE-DB capturing the transition from a negative to a positive emotional state and back.

The ground truth of the audio-visual segment consists of the arousal and valence (A-V) values described in lines 24 and 28 of the algorithm. If only two coders agree in the detection of crossovers, their contribution is weighted by using the correlation metric (cor', calculated as described in Equation 2).

4. Experiments and Results

As a proof of concept, the algorithms introduced have been extensively tested on SAL-DB.

We first present in Fig. 1 two segments extracted by using Algorithm 4, for a transition *to* a negative emotional state. The first dashed vertical line represents the transition *to* that state, and the second one *out of* that state. In the plots, we present the A-V values after the interpolation. Thus, at times no crossover may be observed in the valence values.

As performance evaluation is a significant issue for any automatic system, in Table 2 we attempt to provide meaningful performance results of the introduced algorithms on SAL-DB. The table presents the performance of the automatic audio-visual segmentation procedure in terms of: (i) how well it is able to utilise the actual number of frames (*# of frames*), (ii) using the given data, how many audio-visual segments it is able to produce (*# of segments*), and (iii) how much overlap there is (*overlap*) between the segments, and between the positive and negative classes. The goal of the automatic segmentation procedure is then to utilise as many frames as possible from the given data to produce a high number of meaningful segments. Too much overlap between the segments or between the classes is un-

Table 2: Evaluation of the introduced segmentation algorithms using SAL-DB. The table presents the actual number of frames together with the utilised number of frames (*# of frames*), the number of audio-visual segments produced (*# of segments*) using the data at hand, and the intra-class (percentage of frames included in more than one segment within the same class) and inter-class (percentage of frames included in both classes) overlap.

	subject #	1	2	3	4
total	**# of frames**	56162	80553	28583	88199
negative	**# of frames**	27389	46056	14554	43353
	# of segments	110	170	99	166
	intra-class overlap	6.42%	8.33%	4.53%	7.70%
positive	**# of frames**	23831	36034	13584	38589
	# of segments	110	149	91	174
	intra-class overlap	18.90%	14.18%	10.22%	11.60%
	inter-class overlap	6.16%	7.39%	14.37%	9.92%

Algorithm 4: Segment and produce ground truth: Segmentation()

1 **for** *each coder annotation file c* **do**
2 //capture a transition to and from a neg. state to a non-neg.
3 // use the correlation (cor') for weighting when match has 2 coders
4 $\{PosCrossOver, NegCrossOver\} \leftarrow DetectCrossovers(c)$
5 $MatchedPos \leftarrow MatchCrossOvers(PosCrossOver)$
6 $MatchedNeg \leftarrow MatchCrossOvers(NegCrossOver)$
7 **for** *each matched set of crossovers* $mcos$ *in* $MatchedNeg$ **do**
8 //average time (frame) of crossing over to negative valence
9 //0.5 second offset has been used
10 **if** *length(*$mcos$*)* $\geq$ *3* **then**
11 //agreement in 3 or 4 coders
avgFrm = $int\left(\frac{\sum_{i=0}^{|mcos|} mcos(i).f}{length(mcos)}\right)$
12
13 **end**
14 **else**
15 //2 coders agree, weight using correlation (cor')
avgFrm = $int\left(\frac{\sum_{i=0}^{|mcos|}(mcos(i).f * cor'(mcos(i).c))}{\sum_{i=0}^{|mcos|} cor'(mcos(i).c)}\right)$
16
17 **end**
18 $startFrm = avgFrm - 25$
19 incFrm $\leftarrow$ 0
20 **repeat**
21 incFrm $\leftarrow$ incFrm + 1
22 **if** *length(*$mcos$*)* $\geq$ *3* **then**
23 //agreement in 3 or 4 coders
24 avgValence =
$\frac{\sum_{i=0}^{|mcos|} mcos(i).val(mcos(i).f+incFrm)}{length(mcos)}$
25 **end**
26 **else**
27 //2 coders agree, weight using cor'
28 avgValence =
$\frac{\sum_{i=0}^{|mcos|}(mcos(i).val(mcos(i).f+incFrm) * cor'(mcos(i).c))}{\sum_{i=0}^{|mcos|} cor'(mcos(i).c)}$
29 **end**
30 **until** *sign(avgValence)=1 or avgValence is* NaN ;
31 //add offset after crossing back to non-negative (or NaN)
32 endFrm = (avgFrm + incFrm) + 25
33 //Video is segmented in the range [startFrm,endFrm]
34 //Ground truth (valence/arousal) is averaged
35 **end**
36 //the process is repeated analogously for "to-Positive" crossovers ($MatchedPos$) - line 7
37 **end**

intended and undesirable, but expected to some degree due to the offsets before and after the transitions. By observing Table 2 we conclude that the algorithm fulfills its goal.

As a final step we test the developed algorithms on the recently released SEMAINE-DB. Although the arousal and valence annotations of SEMAINE-DB do not contain NaN values, the steps to be followed for segmentation are similar.

Finally, a qualitative assessment of the proposed algorithms is provided by Fig. 2 and Fig. 3. Fig. 2 illustrates valence annotations from two coders in SEMAINE-DB before and after applying pre-processing operations (for synchronization). Fig. 3 shows example frames from an automatically extracted segment from SEMAINE-DB using the presented algorithms. Overall, the produced segment appears to well capture the transition from a negative emotional state to a positive one, and back.

5. Conclusion

This paper introduced efficient algorithms to the aim of (i) producing ground-truth by maximizing inter-coder agreement, (ii) eliciting the frames that capture the transition *to* and *from* an emotional state, and (iii) automatic segmentation of spontaneous multimodal data to be used by machine learning techniques that cannot handle unsegmented sequences. As a proof of concept, the algorithms introduced have been tested using SAL and SEMAINE data annotated in arousal and valence spaces. Overall, the automatic segmentation procedures introduced appear to work as desired and output segments that well capture the targeted emotional transitions.

6. Acknowledgments

The work of Mihalis A. Nicolaou and Maja Pantic is funded by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). The work of Hatice Gunes is funded by the European Community's 7th Framework Programme [FP7/2007-2013] under the grant agreement no 211486 (SEMAINE).

7. References

C. Alberola-Lopez, M. Martin-Fernandez, and J. Ruiz-Alzola. 2004. Comments on: A methodology for evaluation of boundary detection algorithms on medical images. *IEEE Transactions on Medical Imaging*, 23(5):658–660, May.

N. T. Alves, J. A. Aznar-Casanova, and S. S. Fukusima. 2008. Patterns of brain asymmetry in the perception of positive and negative facial expressions. *Laterality: Asymmetries of Body, Brain and Cognition*, 14:256–272.

S. Baron-Cohen and T. H. E. Tead. 2003. *Mind reading: The interactive guide to emotion*. Jessica Kingsley Publishers Ltd.

R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schroder. 2000. Feeltrace: An instrument for recording perceived emotion in real time. In *Proc. of ISCA Workshop on Speech and Emotion*, pages 19–24.

R. Cowie, E. Douglas-Cowie, and C. Cox. 2005. Beyond emotion archetypes: Databases for emotion modelling using neural networks. *Neural Networks*, 18:371–388.

E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, L. Lowry, M. McRorie, L. Jean-Claude Martin, J.-C. Devillers, A. Abrilian, S. Batliner, A. Noam, and K. Karpouzis. 2007. The humaine database: addressing the needs of the affective computing community. In *Proc. of Int'l Conf. on Affective Computing and Intelligent Interaction*, pages 488–500.

H. Gunes and M. Pantic. 2010. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, 1(1):68–99.

H. Gunes and M. Piccardi. 2009. Automatic temporal segment detection and affect recognition from face and body display. *IEEE Tran. on Systems, Man, and Cybernetics-Part B*, 39(1):64–84.

S. Petridis, H. Gunes, S. Kaltwang, and M. Pantic. 2009. Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities. In *Proc. of ACM Int'l Conf. on Multimodal Interfaces*, pages 23–30.

J. A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

K.R. Scherer, 2000. *The Neuropsychology of Emotion*, chapter Psychological models of emotion, pages 137–162. Oxford University Press.

Wollmer, M. and Eyben, F. and Reiter, S. and Schuller, B. and Cox, C. and Douglas-Cowie, E. and Cowie, R. 2008. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. of 9th Interspeech Conf.*, pages 597–600.

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

The UTEP-ICT Cross-Cultural Multiparty Multimodal Dialog Corpus

David Herrera*, David Novick*, Dusan Jan†, David Traum†

*The University of Texas at El Paso
†Institute for Creative Technologies, University of Southern California
herrera78@gmail.com. novick@utep.edu, {jan,traum}@ict.usc.edu

Abstract

To help answer questions about conversational control behaviors across cultures, a collaborative team from the University of Texas at El Paso and the Institute for Creative Technologies collected and partially coded approximately ten hours of audiovisual multiparty interactions in three different cultures and languages. Groups of four native speakers of Arabic, American English and Mexican Spanish completed five tasks and were recorded from six angles. Excerpts of four of the tasks were coded for proxemics, gaze, and turn-taking; interrater reliability had a Kappa score of about 0.8. Lessons learned from the multiparty corpus are being applied to the recording and annotation of a complementary dyadic corpus.

1. Introduction

Conversational corpora are important for a variety of purposes, including analysis of conversational behaviors, evaluation of theories about behavior, and training data for machine learning algorithms. We are particularly interested in comparing and contrasting conversational control behaviors across cultures. This effort includes each of the above corpus requirements. We need basic data for analysis of the differences in these behaviors across cultures. We will use the data to provide parameters for culture-specific models of virtual human conversation (Jan et al., 2007). Finally we will use the data to attempt to validate the models of virtual human behavior as well as theories from the literature and our own analysis.

Our current focus is on three different kinds of behavior that show cultural variation: proxemics, gaze, and turn-taking. In order to study these behaviors, a collaborative team from the University of Texas at El Paso (UTEP) and the Institute for Creative Technologies (ICT) at University of Southern California have collected and partially coded approximately ten hours of audiovisual multiparty interactions in three different cultures and languages.

In the next section, we elaborate on our goals and related existing corpora, none of which quite meet our needs. In sections 3 and 4, we discuss the design of our corpus and our annotations in the three areas of interest. In section 5, we describe issues that arose in the corpus recording and annotation. Finally, in Section 6, we describe dissemination plans and future directions.

2. Corpus Data Requirements

Part of this project involves understanding differences among language cultures with respect to proxemics, which is the pattern of physical distances people maintain from each other. There is some evidence in the literatures of sociolinguistics and cultural anthropology that these distances differ based on culture and context.

A second part of this project involves understanding differences among language cultures with respect to turn-taking, which is the amount of pause or silence that is typical between people's speech when a speaking turn changes. Thus, we sought to collect data of conversational interactions that are tuned for culture and context. For example, we are working on representing groups of people in the background of a scene, where these people are having small-group conversations. Their proxemics and turn-taking should be appropriate to their culture.

In a third part of the project, we seek to provide data for related conversational behaviors such as gaze, which is a factor in turn-taking. The relation of gaze to turn-taking, particularly considered across cultures, is the subject of open research questions (Rossano et al., 2009)

There are many conversational corpora, but it is still difficult to use these to study general cross-cultural conversational behavior. Many corpora record the speech only, which misses important information in face-to-face interaction. Dyadic conversation has been most studied, but this kind of dialogue has a simplified turn-taking scheme, in which actions such as releasing a turn and assigning a turn are not distinguishable. Moreover, addressee identification is trivial, and there is no distinction between individual or group addressing.

These factors affect the usefulness of existing corpora, including the AMI corpus (Carletta et al., 2005), the CUBE-G corpus (Rehm et al., 2008), and the UTEP CIFA corpus (Flecha-Garcia et al., 2008). The AMI corpus contains dialogues with four participants and audio-visual information, however it does not allow the study of proxemics, because the participants were given fixed locations in which to sit. Moreover, the participants were not balanced for cultural background, so it does not allow cross-cultural studies. Finally, the set of tasks is more rigid, with each participant assigned a specific unique role. Likewise, the CUBE-G corpus, while focused specifically on cultural differences for non-verbal conversational behaviors such as gaze and proxemics, has only dyadic dialogues. Moreover, one member of each of dyad was an actor trained by the researchers, so the corpus could be said to reflect individual rather than multiparty responses. The UTEP CIFA corpus also has limitations. While the participants were from different culture/language groups (American and Arab) and the recordings were made with multiple cameras to help with tracking

Figure 1a American Group 1 in Task 3.

Figure 1b Arab Group 3 in Task 4.

Figure 1: Comparison of American and Arab groups in Toy-related tasks. Note the difference in proxemics between the groups as shown by the dots on the carpet.

gaze, the participants were seated so that proxemics would be an independent variable, and all the conversations were dyadic.

3. Corpus Collection

To address research questions for which uni-cultural, dyadic and seated corpora were ill-suited, we designed the UTEP-ICT corpus with four-person groups, with the participants standing and free to move. The participants are selected from three different cultures: Arabs, Americans, and Mexicans, with each group consisting of members of the same culture. The participants were recruited from local churches, restaurants, on campus, and through networks of known members of each cultural group in the El Paso area, which borders Mexico and has, in part because of the university, many representatives of other nations and cultures. We have completed the recording and partial coding of twelve four-person groups. Four of the groups were composed of native speakers of Arabic, four of native speakers of American English, and four of native speakers of northern Mexican Spanish. In recruiting participants, we sought to obtain a mix of people, some of whom were strangers and some of whom knew each other. To facilitate analysis of culture as independent variable, most of the groups were male-only, but we had one group in each language condition with at least two female participants. In Arab group 1, there were two brothers, one friend (for three years), and one stranger. Arab group 2 comprised two brothers and two cousins. Participants from Arab group 3 belonged to the same English-as-a-second-language program (two friends for 15 years, the rest for few months), and Arab group 4 had two sisters and two strangers. In American group 1, there were two sisters and two strangers. Group 2 comprised a female and her friend (seven years). She was acquainted with a second male (two years), who in turn was acquainted with the third male (two years). Group 3 comprised three soldiers taking the same leadership course and a stranger. Group 4 was comprised of two males and two females. One male knew one of the females for 20 years and the other male for 16 years. The females knew each other for five years. In Mexican group 1, all four males were acquainted (three months). In Mexican group 2, two married couples were also friends (ten years). Mexican group 3 had all females (two were friends for two years) and group 4 had three females and one male (two females were friends for 18 years, the rest a few months).

Task 1	Describe your pet peeves
Task 2	Figure out which movies you've all seen and what were the best and worst parts
Task 3	Come up with a good name for a toy
Task 4	Tell a story about the toy
Task 5	Describe an inter-cultural experience

Table 1: Conversation tasks.

The experimenter asked each of the groups to complete five conversational tasks, which were designed to elicit a range of dialog behaviors. The tasks are listed in Table 1. Tasks 1, 4, and 5 are mainly narrative tasks, where the participants can take turns relating stories or reacting to the narratives of others. Task 1 was meant to "break the ice" and get people comfortable talking with each other. Tasks 2 and 3 are constructive tasks, in which the participants must pool their knowledge and work together to reach a group consensus. Tasks 3 and 4 were designed to have possible task-related gaze focus other than the participants themselves, so a plush toy was provided and tasks related to the toy (see Figure 1). This allows gaze patterns with a copresent referent to be contrasted with gaze patterns without this referent. Task 5 is meant to elicit subjective experiences of intercultural interaction, as a possible starting point for future investigation of higher-level cross-cultural differences.

Each task lasted 10 minutes; the recording sessions lasted about 50 minutes total per group. We wanted to make the interactions as natural as possible, so the participants were not given any special tracking equipment (other than wireless microphones), and the camera were made as unobtrusive as possible. The interactions were recorded with six Apple iMac computers, placed around the periphery of a large open room that serves as a computer lab for UTEP's College of Liberal Arts. We thus recorded six simultaneous views of the participants as they conversed, making it possible, with only rare exceptions, to code the participants' proxemics, gaze and turn-state.

The participants were free to move about the room and

Alternate view 2 of group in Figure 1a.

Alternate view 3 of group in Figure 1a.

Figure 2: Alternate views of Figure 1a.

Alternate view 2 of group in Figure 1b.

Alternate view 3 of group in Figure 1b.

Figure 3: Alternate views of Figure 1b.

stand where they liked. The floor of the computer lab was covered with carpet that had dots evenly spaced at one-foot intervals, which facilitated coding the participants' positions in the room. Audio was recorded at high quality with wireless microphones worn by the participants. Figure 1 shows frames from corpus recordings of an American group and an Arab group. In Figure 1a, the subject on the right holds the plush toy involved in tasks 3 and 4. In Figure 1b, the subject on the left holds the toy. Some of the differences in proxemics among groups can be seen by comparing interpersonal distances between Figure 1a and 1b. Figure 2 shows alternate camera angles for the same the American group, illustrating the ranges views provided in the parallel recordings. Figures 3 shows alternate angles for the same Arab group.

4. Data Coding

From the recordings, we produced time-aligned partial codings of each of the twelve conversations. Specifically, we coded two 30-second excerpts of each of the conversations for tasks 1 through 4 for proxemics, turn-taking, and gaze. For proxemics, a matrix was composed of the area created by the four members as points on a polygon. For turn-taking, the data consisted of a subject's state at each tenth of a second, where state could take a value of talk, pause, or laugh. Table 2 summarizes the range of the collected data and annotations that we have completed, and serves as a guide for identifying specific annotation.

Annotation was done using the ANVIL coding tool (Kipp, 2008). Figures 4, 5, and 6 present examples of annotations of the corpus. Figure 4 shows a segment of proxemics annotations for video file `mx1_t3_1.mov`, which represents the first Mexican group performing task 3 and of the first 30-second excerpt. The figure shows participants' proxemics positions from 17 seconds to 29 seconds. Although only a single numeric value is shown in each element, it consists of x,y-coordinate data in feet and inches. Figure 5 shows the gaze of each participant over the stretch of time from 0 to 6 seconds in Arab group 3 task 4, excerpt 2. D is the speaker during this segment and most of the other participants are looking at him. Figure 6 shows information on when each participant in American group 1, task 2 excerpt 2 was speaking, not speaking (indicated as "pause"), or laughing. In this segment we can see one section between 20 and 21 seconds where three participants are speaking simultaneously, as well as a small segment at 23 where no one is speaking, in between utterances by participant C. Figures 7, 8, and 9 show parallel annotations

Culture	American, Arab, or Mexican
Group	1–4 for each culture
Task	Tasks 1–5 from Table 1
Excerpt	One or two (near beginning or end of task)
Time	Range of time within the excerpt
Behavior	Proxemics, Turn-taking, Gaze

Table 2: Corpus Dimensions.

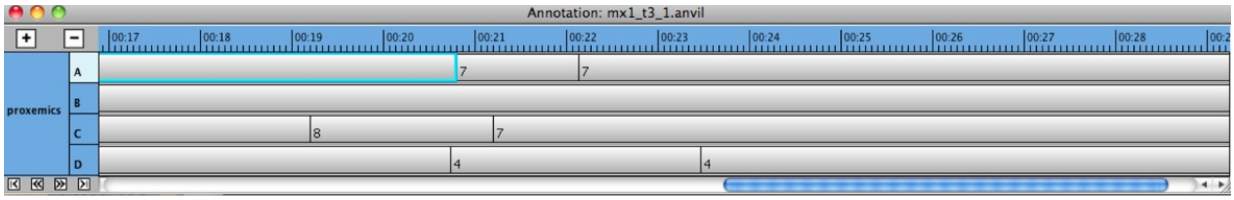


Figure 4: Proxemics coding of Mexican Group 1, Task 3, Excerpt 1.

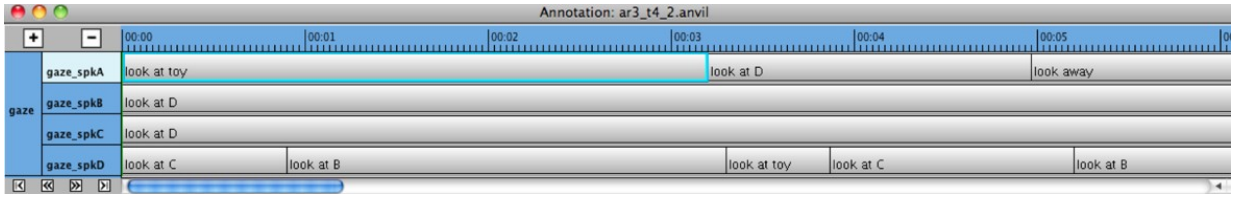


Figure 5: Gaze coding of Arab Group 3 Task 4, Excerpt 2.

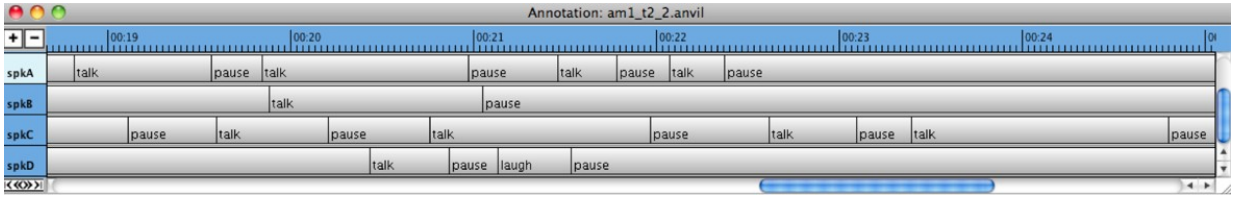


Figure 6: Turn-taking coding of American Group 1, Task 2, Excerpt 2.

for part of a conversation by American group 2 on task 3. In this figure we can see relation of proxemics, gaze and turn-taking behavior of the participants for this period.

5. Corpus Collection & Annotation Issues

In the course of the project, we have had to address a number of issues that arose in annotating the video and audio recordings. At the most basic level, we had to deal with equipment malfunction. Particularly, the software used to record four separate audio channels was sometimes unstable and crashed occasionally. When this occurred, the recovered file would only retain three of the four channels. Coders had to annotate turns for the fourth speaker using audio from the other three channels in conjunction with the video.

Another, more frustrating, problem was the difficulty of placing the cameras to catch the feet and gaze of speakers simultaneously. As speakers were allowed to stand anywhere in an area approximately 20 feet by 10 feet, camera angles could not always capture all gaze and proxemics simultaneously. It was important to capture where the conversants stood, at times sacrificing where the conversant gazed. A camera angle that captures the body of the conversant will not have such a detailed picture of the face making it sometimes hard to see the gaze direction. Additionally, with four conversants, bodies frequently occluded views of others' faces and, even with six camera angles, conversants' eye gaze was not always visible. Originally, we had selected a 30-second excerpts beginning two minutes and six minutes into the conversational task. However, after reviewing some of the videos, there were too many gaze occlusions to obtain useful data, and instead we relied on finding a 30-second excerpt free of occlusions in the first and last five minutes of each conversational task.

A final minor source of error occurred when raters annotated a subject's standing position. As no angle showed all floor marks at once but all marks were relative to the room's top left corner, raters had to determine mark number from one angle to the next by counting the marks. Occasionally, a rater would count incorrectly. Fortunately, these errors were easily spotted when comparing raters' annotations. Counting errors in this case were easily spotted and corrected.

Another difficulty is in recruiting the appropriate subjects. Ideally we would have subjects who had only mono-cultural experience in their native culture, and culture groups would be completely parallel as to their constituent participants, balancing such factors as gender, age, status, how well the participants are known to each other. For this study we were unable to provide such a balance, so it will be difficult to determine which findings are specific to the culture group and which to the particular social relationships of the participants. Broad tendencies across multiple groups (such as proxemic distance in Figure 1) can be attributed to culture group, but many factors will be more subtle, and thus further investigation with additional groups is required.

6. Dissemination and Future Work

We plan to disseminate the corpus to other researchers, subject to privacy-protection restrictions associated with the projects' IRB requirements, beginning January, 2011. Each

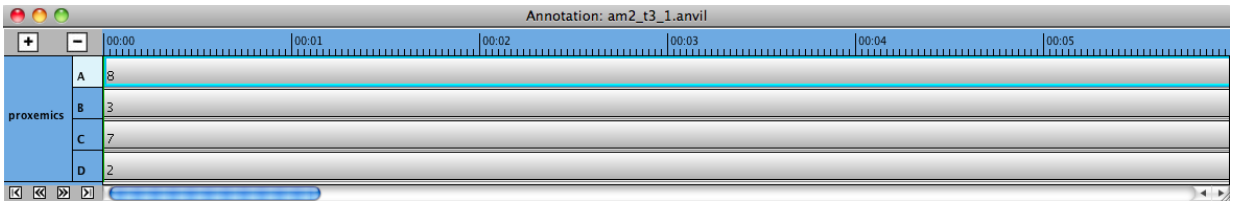


Figure 7: Proxemics coding of American Group 2, Task 3, Excerpt 1.

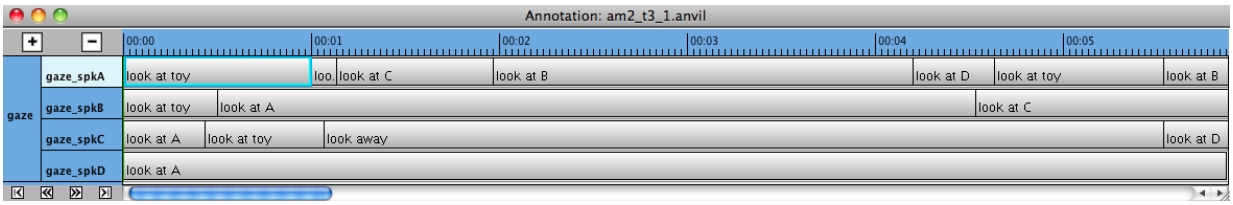


Figure 8: Gaze coding of American Group 2, Task 3, Excerpt 1.

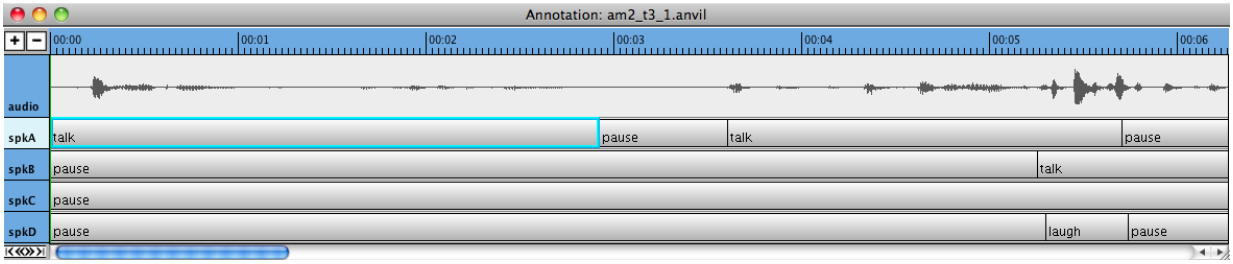


Figure 9: Turn-taking coding of American Group 2, Task 3, Excerpt 1.

file is a combined audio-visual recording. Because there are six cameras per session, the combined corpus files are large; we anticipate that distribution will be via hard-drive or, for subsets, flash drive.

In addition to analyzing patterns of proxemics, gaze, and turn-taking in the multiparty cross-cultural corpus, we are currently collecting a complementary corpus of dyadic conversations with the same tasks across the same cultures. The dyadic corpus should enable us to distinguish differences related to group size from those related to culture.

While six angles were sufficient to determine conversants' proxemic positions, they were not adequate for consistently reliable determination of participants' direction of gaze. For this reason, the corpus collection for dyadic conversations will rely on eight camera angles, although we do not expect as much occlusion as we encountered in the multiparty case. Additionally, the audio software instability problem is no longer anticipated as software for two-channel recording is more stable than that for four-channel recording.

We are also willing to share our partial annotations, which consist of time-aligned notations in ANVIL; we have not yet transcribed the participants' speech, since it was not a major factor in the analyses motivating collection of the corpus, and annotation budget for the initial project was limited. The cross-cultural multimodal phenomena on which our research focuses – proxemics, gaze and turn-taking – appear to be reasonably consistent within groups; our initial analysis suggests that the differences in behaviors between excerpts within groups is much smaller than the differences across groups.

Beyond adding realism to conversational agents in immersive environments, the analysis of the corpus may also help instructors of people who will be conducting conversations with people of different cultures. Because non-verbal behaviors often have different meanings within different culture groups, training in these conversational behaviors may enable conversants to avoid misunderstanding.

Acknowledgments

The work described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position of the United States Government, and no official endorsement should be inferred.

7. References

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain A. McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The ami meeting corpus: a pre-announcement. In *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI'2005*.

M Flecha-Garcia, D Novick, and N Ward. 2008. Differences between Americans and Arabs in the production

and interpretation of verbal and non-verbal dialogue behaviour. In *Speech and Face-to-Face Communication Workshop, Grenoble France*, pages 47–48.

Dusan Jan, David Herrera, Bilyana Martinovski, David G. Novick, and David R. Traum. 2007. A computational model of culture-specific conversational behavior. In *Proceedings of Intelligent Virtual Agents, 7th International Conference, IVA 2007*, pages 45–56.

Michael Kipp. 2008. Spatiotemporal coding in anvil. In *Language Resources and Evaluation Conference (LREC)*, May.

Matthias Rehm, Yukiko Nakano, Hung-Hsuan Huang, Afia Akhter Lipi, Yuji Yamaoka, and Franziska Grüneberg. 2008. Creating a standardized corpus of multimodal interactions for enculturating conversational interfaces. In *Proceedings of the IUI-Workshop on Enculturating Interfaces (ECI)*.

F. Rossano, P. Brown, and S. C. Levinson. 2009. Gaze, questioning and culture. In J. Sidnell, editor, *Conversation analysis: Comparative perspectives*, pages 187–249. Cambridge University Press.

The USC CreativeIT Database: A Multimodal Database of Theatrical Improvisation

Angeliki Metallinou [†], Chi-Chun Lee [†], Carlos Busso [‡], Sharon Carnicke [£], Shrikanth Narayanan [†]

[†] Electrical Engineering Department, University of Southern California, Los Angeles CA 90089,
[‡] Electrical Engineering Department, University of Texas at Dallas, Dallas TX 75080,
[£] School of Theater, University of Southern California, Los Angeles CA 90089
metallin@usc.edu, chiclee@usc.edu, busso@utdallas.edu, carnicke@usc.edu, shri@sipi.usc.edu

Abstract

Improvised acting is a viable technique to study human communication and to shed light into actors' creativity. The USC CreativeIT database provides a novel bridge between the study of theatrical improvisation and human expressive behavior in dyadic interaction. The theoretical design of the database is based on the well-established improvisation technique of Active Analysis in order to provide naturally induced affective, goal-driven interaction. The carefully engineered data collection and annotation processes provide a gateway to quantify and investigate various aspects of theatrical performance and human communication.

1. Introduction

Human interaction is a complex blend of intents, communicative goals and emotions, which are expressed, among others, through body language, prosodic cues, speech content. The study of human communication and expressive behaviors has attracted interest from multiple domains including psychology, social sciences, engineering, theater, etc. This paper describes the design, collection and annotation process of a novel, multimodal and multidisciplinary interactive database, the USC CreativeIT database. The database is a result of the collaborative work between the USC Viterbi School of Engineering and the USC School of Theater. The database is collected using cameras, microphones and motion capture and contains detailed audiovisual information of the actors' body language and speech cues. It serves two purposes. First, it provides insights into the creative and cognitive processes of actors during theatrical improvisation. Second, the database offers a well-designed and well-controlled opportunity to study expressive behaviors and natural human interaction.

The significance of studying creativity in theater performance is that improvisation is a form of real-time dynamic problem solving (Mendonca and Wallace, 2007). Improvisation is a creative group performance where actors collaborate and coordinate in real time to create a coherent viewing experience (Johnstone, 1981). Improvisation may include diverse methodologies with variable levels of rules, constraints and prior knowledge, concerning the script and the actor's activities. Active Analysis, introduced by Stanislavsky, proposes a goal-driven performance to elicit natural affective behaviors and interaction (Carnicke, 2008), and is the primary acting technique utilitized in the database. It provides a systematic way to investigate the creative processes that underlie improvisation in theater. The role of acting has been considered as a viable research methodology for studying human emotions and communication. Theater has been suggested as a model for believable agents; agents that may display emotions, intents and human behavioral qualities (Perlin and A.Goldberg, 1996). Researchers have advocated the use of improvisation as a tool for eliciting naturalistic affective behavior for studying emotions and argue that improvised performances resemble real-life decision making (Fig. 1, (Busso and Narayanan, 2008)). Furthermore, it has been suggested that experienced actors, engaged in roles during dramatic interaction may provide a more natural representation of emotions, avoiding exaggeration or caricatures (Douglas-Cowie et al., 2003).

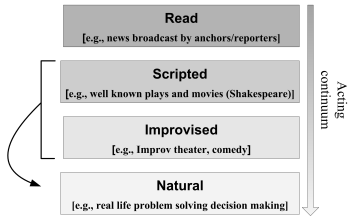


Figure 1: Acting continuum: From fully predetermined to fully undetermined (Busso and Narayanan, 2008)

A variety of acted emotional/behavioral databases exist in the literature. As argued in (Enos and Hirschberg, 2006) valuable emotional databases can be recorded from actors using theatrical techniques. Examples of databases which explore acting techniques include the audiovisual IEMOCAP database (Busso et al., 2008), which contains improvised and scripted acting, and the speech Genova Multimodal Emotion Portrayal (GEMEP) database (Banziger and Scherer, 2007). In (Anolli et al., 2005), authors describe the collection of a multimodal database where contextualized acting is used.

The USC CreativeIT database is a novel, multimodal database that is distinct and complemets most of the existing ones. Its theoretical design is based on the well-established theatrical improvisation technique of Active Analysis and results from a close collaboration of theater experts, actors and engineers. We utilize Motion Capture technology to obtain detailed body language information of the actors, in addition to microphones, video and carefully designed post-performance interviews of the participants. Annotation of the data includes continuous emo-

This work was supported in part by funds from NSF

tional descriptors (valence, activation) as well as theatrical performance ratings (naturalness, creativity) from various perspectives (e.g actor, expert, observer). The database aims to facilitate the study of creative theatrical improvisation qualitatively and provides a valuable source to study human-human communicative interaction.

The rest of this paper is organized as follows. Section 2 describes the theatrical methodology and design hypotheses, section 3 contains the experimental protocol and the technological equipment and section 4 describes the data annotation process. Finally, section 5 contains discussion of future research directions.

2. Theatrical Methodology

2.1. Active Analysis

In Active Analysis, the actors play conflicting forces that jointly interact. The balance of the forces determines the direction of the play. The scripts used in the case play the role of guiding the events (skeleton). The course of the play can be close to or different from the script. This degree of freedom provides an flexibity to work at different levels in the improvisation spectrum. A key element in Active Analysis is that actors are asked to keep a verb in their mind, while they are acting, which drives their actions. As a result, the interaction and behavior of the actors may be more expressive and closer to natural, which is crucial in the context of emotion recognition. For instance, if the play suggests a confrontation between two actors, one of them may choose the verb *inquire* while the other may choose *evade*. If the verbs are changed (e.g.*persuade*, *confront*) the play will have a different development. By changing the verbs, the intensity of the play can be modified as well (i.e.ask versus interrogate). As a result, different manifestations of communication goals, emotions and non-verbal behaviors can be elicited through the course of the interaction. This flexibility allows us to explore the improvisation spectrum at different levels and makes Active Analysis a suitable technique to elicit emotional manifestations.

2.2. Design of Data Collection

The USC CreativeIT database utilizes two different theatrical techniques, the two-sentence exercise and the paraphrase, both of which orginate from the Active Analysis methodology. We also perform a post-performance survey after the recording.

In the two sentence exercise, each actor is restricted to saying one given sentence with a given verb. For example, one actor may say "Marry Me" with verb *confront*, and another one may say "I'll think about it" with verb *deflect*. Given the lexical contraint, the expressive behaviors and the flow of the play will be primarily based on the prosodic and non-verbal behaviors of the actors. This type of controlled interaction can bring insights into how human/actors use their expressive behaviors, such as body language and prosody, to reach a communication goal. Also, this approach is suitable to study emotion modulation at a semantic level, since the same sentences are repeated different times with different emotional connotation.

In the paraphrase, the actors are asked to act out a given script with their own words and interpretation. Examples of plays that are used are "The Proposal" by Chekhov or "Taming of the Shrew" by Shakespeare. In this set of recordings, actors are not lexically constrained. As a result, the performance is characterized by a more natural and free-flow interaction between the actors which bears more ressemblance to real-life scenarios, compared to the two-sentence exercise. Therefore, behavioral analysis and findings on such sessions could possibly be extrapolated to natural human interaction and communication.

Finally we perform a brief interview of the actors right after each performance. Examples of the questions asked are 'What verbs did you and the other actor use?', 'What was the goal of your character?', 'How would you describe your and the other actor's emotion during the interaction?'. These questions are designed to help understand the cognitive planning process of the actors as they improvise on the scenes.

3. Data Collection

3.1. Session Protocol

An expert on Active Analysis (the 4th author of the paper) directed the actors during the rehearsal and the recording of the sessions. Prior to the scheduled data collection date, the actors had to go through a rehearsal with the director to become familiar with active analysis and the scene. Just before the recording of the paraphrase, there was another 5-minute session to refresh actors' memory and give the director a chance to remind actors of the essence of the script. A snapshot of an actor during the data collection is shown in Figure 2(a)

The data collection protocol consists of the following steps:

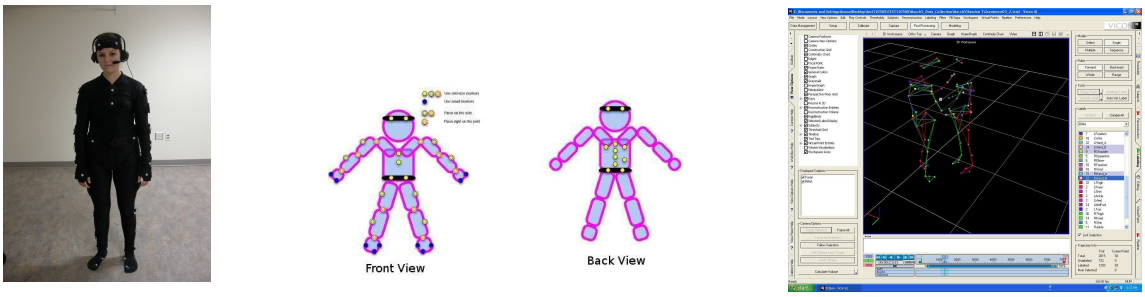
1. Two-Sentence Exercise (unknown verbs)
2. Two-Sentence Exercise, using the same sentences as previously but different verbs (known verbs)
3. Paraphrase of Script (known verbs)
4. Paraphrase of Script, using the same script as previously but different verbs (known verbs)
5. Two-sentence Exercise (unknown verbs)
6. Two-sentence Exercise, using the same sentences as previously but different verbs (known verbs)

Verbs are chosen either by the actors or the director prior to each performance. Some of the commonly chosen verbs are *to shut him out, to seduce, to deflect, to confront, to force the issue* etc, which introduce a large variety of communication goals. *Unknown verbs* indicate that actors are not aware of each other's verb prior to the performance. This setting provides a variety in the interaction dynamic of the two-sentence exercise. During the paraphrases the actor's verbs are always known to each other in advance.

3.2. Equipment and Technical Details

The following is the list of equipment that is utilized in the data collection:

- **Vicon Motion Capture System:** 12 motion capture cameras to record 45 marker's (x, y, z) position for each actor. The markers are placed according to Figure 2(b).
- **HD Sony Video Camcorder:** 2 Full HD cameras are places at each corner of the room to capture the performance of the actors.
- **Microphones:** Each actor has a close-up microphone to record actors' speech at 48KHz with 24 bits.



(a) Actor wearing microphone and markers

(b) Positions of markers

(c) Snapshot of the motion capture post-processing software

Figure 2: Snapshots of an actor during data collection, the marker positions and the post-processing software

3.3. Motion Capture Post-Processing

The first data post processing step is to map each of the markers captured into a subjects defined body model. There are two subjects each with 45 markers, and also there are about 5000 - 10000 frames per interaction session. Since actors are asked to be expressive with body language and gesture, occulusion of markers happens fairly often. Because of this, the computer software is unable to perform all the labeling automatically and accurately. For example, when two subjects are close to one another, one's hand marker may be labeled as the other person's shoulder if we rely on computer labeling. In order to obtain reliable and detailed marker information, the motion capture data was manually corrected frame by frame. The spline function was used to interpolate any missing markers. Such post-processing of one actor in one performance may require approximately 1 - 2 hours, which is a fairly time consuming task. Figure 2(c) shows a snapshot of the post-processing software.

3.4. Data Collection and Annotation Progress

The database contains the recording of nine full sessions, each of which contains approximately one hour of audiovisual data. In total we have recorded 40 two-sentence exercises and 19 paraphrases with 19 actors. The motion capture post-processing step is approximately 95% complete. One full session has been annotated by five different annotators.

4. Data Annotation

4.1. Annotation Attributes and Annotator Groups

The design of the annotation process depends on the collected data as well as the research scope of the database. During improvised dyadic sessions there is a continuous flow of body language and dialog and a diverse expression of emotions and intentions. In order to preserve this flow, we annotate the sessions using continuous labels instead of chopping them into sentences or other arbitrary chunks. Furthermore, since commonly used categorical emotional attributes (angry, happy etc) may not be applicable or sufficient for our data, we choose a more comprehensive set of attributes. These contain dimensional emotional descriptors (valence, activation, dominance) as well as theater performance ratings (interest, naturalness), which may facilitate future theatrical performance analysis.

The attributes that are annotated for each session are described in Table 1. For each attribute, it is mentioned whether the annotation is continuous or a discrete label per session, or both, and if the attribute is annotated per actor or per session as a whole.

All continuous annotations are performed by watching the session videos and using the Feeltrace software. Feeltrace is a publicly available emotional annotation tool, described in (Cowie et al., 2000), which we slightly modified to suit our purposes. The Feeltrace interface enables the user to continuously move the mouse along the computer screen so as to indicate the attribute value, ranging from -1 to 1. For the discrete annotations, annotators are asked to provide a label ranging from 1 to 5.

The annotated attributes are, to a large extent, subjective. In addition to using multiple annotators for the same videos, we are also interested in examining how diverse audience groups may perceive and rate a video, according to their expertise. We categorize the annotators into three groups; theater experts, actors and naive audience. The first group consists of professors of the USC theater school and experts in active analysis while the second consists of students of the theater school, who may or may not have performed in the session. Finally, the naive audience(observer) group consists of USC students who have no technical knowledge of theater.

4.2. Multiple Annotator Correlations

In order to examine the correlations between different annotators for a certain attribute, we performed statistical analysis of the annotations of one two-sentence exercise recording. An example is shown in Figure 3, where we present a segment of the annotation of the activation of an actor, annotated by 5 people; 3 students (naive audience) and 2 actors. Although various annotations differ, the correlations between them are evident. In order to examine linear relationships between the annotations, we computed the Pearson correlation coefficients between all pairs of annotations, which are all found significant at the 0.01 level (2-tailed). We also investigate the prediction success of an annotation using linear regression with the rest of the annotations as predictors. In Table 2, for each of the 5 cases, we present the adjusted R^2 as a measure of the goodness of fit of the linear regression model. The relatively large numbers of R^2 indicate that an annotation can be well-predicted using the rest annotations, suggesting linear relationships

Attribute	Definition	Type	Rating
Continuous Emotional Descriptors			
Valence	Positive vs Negative	continuous	per actor
Activation	Excited vs Calm	continuous	per actor
Dominance	Dominant vs Submissive	continuous	per actor
Theatrical Performance Ratings			
Interest	How interesting do you find the session	continuous and discrete	per session
Naturalness	How natural do you find the performance	continuous and discrete	per actor
Creativity	How creative,in terms of novelty, do you find the performance	discrete	per actor
Actor Verbs	How successful are actors in performing their verbs	discrete	per actor

Table 1: Annotated attributes

between the different annotations.

Furthermore, we computed the Pearson correlation coefficients between all pairs of annotations for all continuous emotional descriptors of the two-sentence exercise (activation, valence, dominance). They were all found statistically significant at the 0.01 level (2-tailed) and positive, which indicates consistency between different annotators despite their possibly different styles. All statistical tests were performed using the SPSS software.

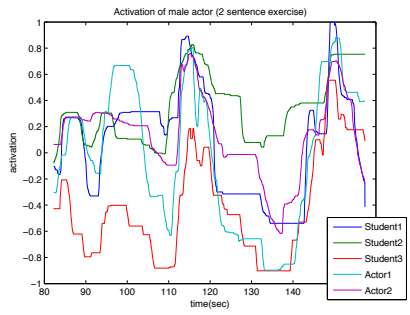


Figure 3: Continuous annotation of the activation of one actor in a two-sentence exercise recording

Linear Regression Fit		
Dependent	Predictors	adjusted R^2
student1	students:2,3, actors:1,2	0.731
student2	students:1,3, actors:1,2	0.561
student3	students:1,2, actors:1,2	0.501
actor1	students:1,2,3 actor:2	0.530
actor2	students:1,2,3 actor:1	0.650

Table 2: Adjusted R^2 values of linear regression for the annotations of the activation of an actor for a two-sentence exercise

5. Future Research Directions

The USC CreativeIT database is a novel, multimodal and multidisciplinary database which represents a unique opportunity to marry engineering methods with the theory and practice of acting. Future research directions that could be pursued using these data include:

- Analysis of prosody and nonverbal behaviors of the actors, such as facial expression and body language. Investigation of how these behaviors are affected by the communication goal, which is specified by the improvisation verb.
- Analysis of the interaction flow and possible synchronization patterns between the actors during the performance, in relation to the pair of improvisation verbs used.
- Analysis of the theatrical performance ratings and possible differences in ratings among the different evaluator groups. Investigation of the body language and expressive choices that may lead to higher overall performance ratings.
- Application of the insights gained from the database analysis to the design of affect-sensitive believable agents.

6. References

L. Anolli, F. Mantovani, M. Mortillaro, A. Vescovo, A. Agliati, L. Confalonieri, O.Realdon, V. Zurloni, and A. Sacchi. 2005. A multimodal database as a background for emotional synthesis, recognition and training in e-learning systems. In *ACII 2005, Beijing*.

T. Banziger and K. R. Scherer. 2007. Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus. In *Int'l Conference on Affective Computing and Intelligent Interaction (ACII)*.

C. Busso and S. Narayanan. 2008. Recording audio-visual emotional databases from actors: a closer look. In *In Language Resources and Evaluation (LREC 2008), Marrakech, Morocco*, pages 17–22, May.

C. Busso, M. Bulut, C-C Lee, A.Kazemzadeh, E. Mower, S. Kim, J. Chang, S.Lee, and S.Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

S. M. Carnicke. 2008. *Stanislavsky in Focus: An Acting Master for the Twenty-First Century*. Routledge, UK.

R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schroeder. 2000. Feeltrace: An instrument for recording perceived emotion in real time.

E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. 2003. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40:33–6, April.

F. Enos and J. Hirschberg. 2006. A framework for eliciting emotional speech: Capitalizing on the actor's process. In *LREC Workshop on Corpora for Research on Emotion and Affect, Genova, Italy*.

K. Johnstone. 1981. *Impro: Improvisation and the Theatre*. Routledge / Theatre Arts, New York.

D. Mendonca and W. Wallace. 2007. A cognitive model of improvisation in emergency management. *IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans*, 37(4):547–561.

K. Perlin and A.Goldberg. 1996. Improv: A system for scripting interactive actors in virtual worlds. In *Proceedings of the 23rd Annual Conference on Computer Graphics*.

Exploiting Motion Capture for Virtual Human Animation

Data Collection and Annotation Visualization

Alexis Heloir[1], **Michael Neff** [2], **Michael Kipp**[1]

[1]DFKI, Germany [2]UC Davis, USA
[1]firstname.surname@dfki.de [2]neff@cs.ucdavis.edu

Abstract

Motion capture (mocap) provides highly precise data of human movement which can be used for empirical analysis and virtual human animation. In this paper, we describe a corpus that has been collected for the purpose of modelling movement in a dyadic conversational context. We describe the technical setup, scenarios and challenges involved in capturing the corpus, and present ways of annotating and visualizing the data. For visualization we suggest the techniques of motion trails and animated re-creation. We have incorporated these motion capture visualization techniques as extensions to the ANVIL tool and into a procedural animation system, and show a first attempt at automated analysis of the data (handedness detection).

1. Motivation

Video has been the technology of choice for empirical movement analysts since it faithfully records the movements, facial expressions and spatial surroundings of the recorded subject. However, video has obvious limitations: the view angle cannot be changed after recording and any automatic analysis must use computer vision techniques to extract meaningful information like hand/face locations.

Motion capture is becoming an increasingly widely available resource for recording human movement. It allows researchers to supplement audio and video recordings with 3D reconstructions of a performer's movements. Normally, motion capture techniques reconstruct body movement as a stick figure skeleton, yielding angle data at each joint in the skeleton. The amount of data that can be captured is a function of the number and resolution of the cameras available, the number of subjects, the range of movement allowed and the amount of time available for cleaning and reconstructing the data. Nonetheless, motion capture provides a more precise 3D view of a subject's movement and also supports automated analysis of the data.

Although motion capture offers numerous advantages for motion analysis, new tools and visualization techniques are needed to fully exploit the potentials of this technology. In this paper, we present some of the trade-offs involved in building a corpus of conversational interactions that includes motion capture. Our intended application is virtual character models that can both talk and gesture. We discuss issues related to both capturing data and analysis. We also illustrate how an existing motion annotation tool ANVIL (Kipp, 2001; Kipp, 2010b; Kipp, 2010a) can be extended in order to take advantage of such data.

Our final corpus contains audio, video and motion capture data. Each modality provides different, important information for the analysis and synthesis process. Audio data provides the text that was spoken and the word timings. Motion capture data provides a 3D reconstruction of the motion, but in many standard applications such as ours, this reconstruction is at the fidelity of a stick figure. It does not capture the surface deformations of the performer, including facial expressions, muscle bulges and breathing. Video helps provide these missing pieces. Shooting from two angles, we can capture facial expressions of both interlocutors and also subtleties of body movement that may be missed in the motion capture.

2. Motion Capturing Dyadic Conversations

Building a corpus begins by determining the goals for its intended use, and from that, planning a set of scenarios to record, and choosing appropriate subjects. Our goals were to perform early, exploratory studies on gestures analysis and generation for two person (dyadic) conversations. This required obtaining a wide set of gesture variations. In this section, we describe one particular session in building our corpus.

2.1. Scenarios

We decided to use improvised scenarios as they placed less demand on our subjects by not requiring them to learn lines and also avoided introducing the bias of a pre-selected script. We chose subjects with extensive movement experience, both subjects had dance training and performance experience. Both were trained in Laban Movement Analysis. In general, we feel experience with verbal improvisation and physical acting is important for this kind of session, and offers the following benefits:

- subjects can better cope with the disturbing garment/setup required by motion capture,
- subjects can improvise coherent stories and interaction with minimal guidance,
- subjects can take directions well and adjust their performance to yield the desired data,
- the addtional training these subjects had in Laban Movement Analysis (LMA) (Laban, 1988) allowed them to be given directions in terms of LMA parameters, which allows precise changes in movement to be requested.

We recorded 23 separate sequences, each having a length of 1-2 minutes. In 19 of the 23 sequences, both actors were interacting. Performers were given minimal improvisation instructions, each focused on particular aspects of interaction:

- social status and levels of dominance, as suggested by Johnstonne (Johnstone, 1981),
- *valence* of the interaction,
- amount of *arousal* in the interaction,
- discussions where subjects *agree or disagree*.

The recording started with a warm-up sequence where subjects were told to talk about what they did the day before. Subsequent sequences are summarized and briefly described in the following table:

Dominance	
Corrupt judge and briber	judge is low status
Corrupt judge and briber	judge is high status
New neighbors meeting	both high status
New neighbors meeting	both low status
Boss fires employee	boss is high status
Boss fires employee	boss is low status
Valence	
Old friends meet	they are happy to meet
Uncomfortable meeting	they dislike each other
Arousal	
Sketch of the dead parrot	high arousal
Sketch of the dead parrot	low arousal
Agree/disagree	
Coffee is better with a cigarette	disagree
Brad Pitt should be president	agree
Mac computers are better	disagree
Jay Leno is an alien	agree
The best way to eat an egg (small end/big end first)	disagree

2.2. Technical Setup

We were interested in capturing the following modalities for two subjects simultaneously: speech, posture and gesture, including hand shape. Our corpus was recorded in a motion capture lab equipped with 12 optical Vicon MX 40+ cameras and two digital HD video cameras. This system tracks the 3D locations of refelctive markers attached to the subjects. Speech and facial expressions were captured using digital video cameras aimed at each subject. Motion capture was used to record both body motion and hand shape, the latter being particularly challenging. The difficulty of recording finger motion using optical motion capture, especially with a limited number of cameras, comes from the high probability of visual occlusions between crossing and/or overlapping fingers. Full body and hand capture can be attempted using three different strategies:

One strategy consists of recording the hand motion and the body motion separately. The performer must wear different markers for each capture session and the two sets of data must be spliced together afterwards using temporal warping algorithms. This method has been successfully demonstrated by Majkowska et al. for choreographed Mudras dance (Majkowska et al., 2006). Unfortunately our scenarios heavily rely on improvised performances in which the hand poses are unknown. Indeed, one of our goals is to study changes in hand shape. Therefore, this technique didn't fit our requirements.

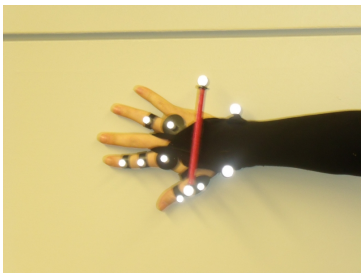


Figure 1: Marker constellation used for hand shape and finger orientation.

A second strategy uses a combination of an optical motion capture system for the body motion and a glove equipped with bend sensors for the hands. This technique has the advantage of being robust to finger occlusions and has been successfully employed for recording Sign Language sequences (Heloir et al., 2005). However, data gloves have several drawbacks: they record motion at lower frequency than optical system (approx. 60Hz vs 120Hz), the sensors have non-linear behavior when approaching flexion limits, the gloves need to be recalibrated at regular intervals, they are expensive and many systems require wires.

A third strategy consists of using a limited set of optical markers on the hand to capture a portion of its movement, and then inferring the remainder of the hand shape. This technique has been used extensively in the motion picture industry and has proved to give acceptable, although not optimal, results. Recent research work took advantage of the joint inter-dependencies of the human hand to perform hand motion capture with a limited set of markers for grasping tasks (Chang et al., 2007). The third method was chosen because it made use of existing equipment and allowed for the simultaneous capture of hand and body movement.

After some experiments, we found that seven markers on the hand were enough to provide a faithful reconstruction of the hand's overall shape in most instances. The marker constellation for one hand is depicted in Fig. 1. We used two markers for the thumb, two markers for the ring finger and three markers for the index.

2.3. Lessons Learned

The recording of the 23 sequences took six to seven hours. Two hours were necessary to brief and prepare the two subjects. Once recorded, postprocessing of the motion capture data took one week for a single person working full time. Not surprisingly, the reconstruction of the hand motion required the most manual correction. Only for some sparsely occurring intervals (approx. 3% of the time), hand motion reconstruction could not be achieved due to occlusion.

3. Annotation, Analysis and Visualization

In our work, we are concerned with the phase structure of gesture (Kita et al., 1998). A given gesture can be broken down into a set of phases: preparation, hold, stroke and retraction. The whole gesture is considered the next level of

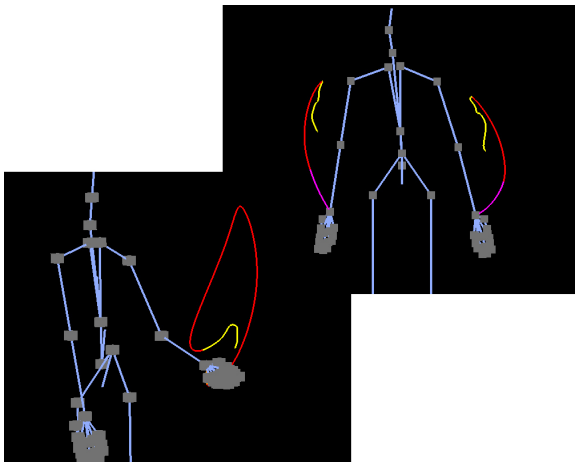


Figure 2: Two examples of gesture trails. Yellow indicates the preparation phase, red the stroke and magenta the retraction.

analysis and called a *phrase* in the literature. It is useful to break these phases out both for analysis and generation. From the perspective of analysis, the *stroke* phase is considered the meaning carrying portion of the gesture, so it is helpful to separate it from the total gestural movement. Another important phase is the *independent hold* if a gesture has no movement at all (e.g. the proverbial raised index finger). Both strokes and independent holds are called the *expressive phase* of a gesture. Other factors like the occurrence and length of holds can help define a particular individual's gesture style. From the perspective of generation, the phase structure provides a convenient framework for specifying animation. A system can solve for the poses at the phase boundaries and interpolate in between to create continuous gesture animation.

We manually annotate gesture phases using the tool ANVIL which has recently been extended to visualize motion capture data using a 3D skeleton (Kipp, 2010b). ANVIL allows users to view synchroized video, 3D skeleton data and time-aligned annotations. Additionally, ANVIL can visualize the position, velocity and acceleration of the hands as curves (either x, y, z separately or as total value) on separate tracks (Figure 3).

3.1. Automated Handedness Analysis

Since motion capture data offers more information than plain video, providing nearly continuous 3D data, it offers increased potential for automatically deriving meaningful descriptions of the movement. In manual annotation, it can be a challenge to arrive at high inter-coder agreement for phenomena like gesture phase annotation, since this can be quite subjective, especially for spontaneous gestures. If more of these tasks can be successfully automated (even if partially so, combined with human corrections), it will increase inter-coder agreement and reduce coding effort. Detecting the hand used for a gesture (LH, RH, 2H) is one annotation task that lends itself to automation.

To detect handedness on the phrase level (i.e. for a whole gesture) we first find the corresponding *expressive phase*

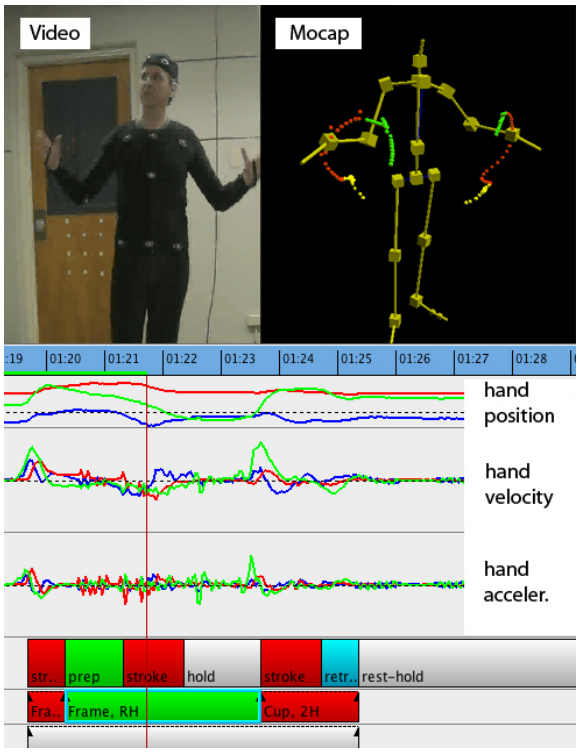


Figure 3: In the ANVIL tool, movement is usually encoded in terms of timeline-based annotations (bottom: colored boxes) and video. Mocap data allows for the display of hand position, velocity and acceleration curves. Motion trails visualize the hands' path in 3D, viewable from all angles.

on the phase track. The expressive phase is either a stroke or an independent hold (Kita et al., 1998). This phase is marked red in Fig. 2, note how the difference in length gives a clear cue of handedness. Therefore, we take the length of the path travelled by left hand L_{RH} and right hand L_{LH} respectively during this expressive phase (in meters), and normalize it by the duration d of the phase (in seconds). If the normalized difference $\frac{|L_{RH}-L_{LH}|}{d}$ is below the threshold of $0.12\frac{m}{s}$, we label it a bihanded gesture (2H), otherwise we label it right-handed if $L_{RH} > L_{LH}$, or left-handed (LH) if $L_{RH} < L_{LH}$. On an annotated corpus of 269 phrases, we achieved 83% correct annotations with this algorithm.

3.2. Annotation Visualization with Motion Trails

Previous approaches (Neff et al., 2008; Kipp et al., 2007) for annotating gesture have included positional data by estimating the wrist positions at the start and end of a stroke. This provides a sparse description of the gesture sequence. One of the chief advantages of motion capture is that by capturing over 100 samples per second, it approximates a more continuous representation of the motion. This allows us to visualize the overall form of a gesture.

We suggest a new visualization technique that draws the 3D movement of the speaker's hand as a "trail" through space, shown either as a continuous line or by discrete sphere (Figures 2 and 3). This allows one to closely examine the actual path of a gesture from all angles, revealing the smoothness

or edginess of the curve and even giving an impression of the velocity profile which is reflected in the spacing of the spheres.

The trail feature has been incorporated into both a standalone animation package and the ANVIL annotation tool. The gesture trails are color coded to indicate the phases of the gesture, as shown in Figure 2. We can play an animation of the trail data with its actual timing, scrub through the trail and also view it from any direction in 3D. This allows for more carefull study of the gesture form and the transitions between the phases, to examine features like the continuity across phases. One insight we gained with respect to phase boundaries is that changes in hand shape may play a significant role in defining these boundaries because judging from the trails alone (no hand shape information visible!) boundaries would often have been placed a little earlier or later.

3.3. Validation by Recreation

In both ANVIL and our standalone animation system, we can simultaneously playback the motion capture data with the gesture trail over top of it. This provides an easy and effective method for validating an annotation. If the animated character performs a gesture, but there is no accompanying gesture trail, this indicates an error in the annotation. This makes it very easy to detect errors such as marking the incorrect hand, missing a gesture, or annotation errors in the timing of the gesture.

We can also use the motion capture data as input to a procedural motion generation system. For instance, the systems presented in (Neff et al., 2008; Heloir and Kipp, 2009) use the positions at the start and end of the stroke in order to generate animation. This position data can be automatically calculated from the motion capture data and then used as input to the procedural systems. We can overlay both the motion capture and generated animations and produce gesture trails for each. This allows for direct comparison between the form of the gesture created by the procedural system and the form of the original gesture. It provides a way to evaluate and improve the procedural generation system so that it can better match the captured data.

4. Summary and Outlook

We have described the recording of a motion capture corpus involving two speakers interacting in various improvisational scenarios. We showed that a standard optical motion capture setup was sufficient to provide a faithful reconstruction of the body and hand motion of both subjects. We found that 7 hand markers, strategically placed, were suffcient to reconstruct hand shape.

We also presented a basic annotation scheme in terms of gesture phases and two visualizaton helpers designed in order to reduce annotation errors and to increase inter-coder agreement. The first one, motion trails, shows the 3D path of the hands colored according to the movement phase annotation. The second one, re-creation, animates a stick figure according to extracted information which allow direct visual feedback concerning the quality of the animation algorithm.

In the future we plan to pursue two major lines of inquiry: first, determining methods to automatically derive movement phases from motion capture data and second, analyzing the particular interactions between two speakers in terms of timing, rhythm and imitation.

Acknowledgements

We would like to thank the performers for contributing to our corpus. Many thanks Quan Nguyen (DFKI) for implementing motion trails in ANVIL. Financial support for this work was provided in part by NSF grant IIS 0845529, by DAAD grant D/08/09945 (PPP USA) and by the DFG-funded Excellence Cluster *Multimodal Computing and Interaction* (MMCI).

5. References

Lillian Y. Chang, Nancy Pollard, Tom Mitchell, and Eric P. Xing. 2007. Feature selection for grasp recognition from optical markers. In *Proc. of the 2007 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS 2007)*, pages 2944 – 2950.

Alexis Heloir and Michael Kipp. 2009. EMBR - a realtime animation engine for interactive embodied agents. In *Proc. of the 9th Int. Conf. on Intelligent Virtual Agents (IVA-09)*.

Alexis Heloir, Sylvie Gibet, Franck Multon, and Nicolas Courty. 2005. Captured motion data processing for real time synthesis of sign language. In *Proc of GW-05*.

Keith Johnstone. 1981. *IMPRO: Improvisation and the Theatre*. Methuen Drama.

Michael Kipp, Michael Neff, and Irene Albrecht. 2007. An annotation scheme for conversational gestures: How to economically capture timing and form. *Journal on Language Resources and Evaluation*, 41(3-4):325–339, December.

Michael Kipp. 2001. Anvil – a generic annotation tool for multimodal dialogue. In *Proc. of Eurospeech*, pages 1367–1370.

Michael Kipp. 2010a. Anvil: The video annotation research tool. In Jacques Durand, Ulrike Gut, and Gjert Kristofferson, editors, *Handbook of Corpus Phonology*. Oxford University Press.

Michael Kipp. 2010b. Multimedia annotation, querying and analysis in ANVIL. In Mark Maybury, editor, *Multimedia Information Extraction*, chapter 21. MIT Press.

Sotaro Kita, Ingeborg van Gijn, and Harry van der Hulst. 1998. Movement phases in signs and co-speech gestures, and their transcription by human coders. In Ipke Wachsmuth and Martin Fröhlich, editors, *Proc. of GW-97*, pages 23–35, Berlin. Springer.

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

Anna Majkowska, Victor B. Zordan, and Petros Faloutsos. 2006. Automatic splicing for hand and body animations. In *SCA '06*.

Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Trans. on Graphics*, 27(1):5:1–5:24, March.

Linking Conversation Analysis and Motion Capturing: How to robustly track multiple participants?

Karola Pitsch[1,2], Bernhard Brüning[3], Christian Schnier[1], Holger Dierker[3], Sven Wachsmuth[1,3]

[1]Applied Informatics, [2]CoR-Lab & [3]CITEC
Bielefeld University, Faculty of Technology, P.O.Box 100 131, 33501 Bielefeld, Germany
E-mail: {kpitsch}{bbruenin}{cschnier}{hdierker}{swachsmu}@techfak.uni-bielefeld.de

Abstract

If we want to model the dynamic and contingent nature of human social interaction (e.g. for the design of human-robot-interaction), analysis and description of natural interaction is required that combines different methodologies and research tools (qualitative/quantitative; manual/automated). In this paper, we pinpoint the requirements and technical challenges for constituting and managing multimodal corpora that arise when linking Conversation Analysis with novel 3D motion capture technologies: i.e. to *robustly* track *multiple* participants over an *extended* period of time. We present and evaluate a solution to by-pass the limits of the current standard Vicon system (using rigid bodies) and ways of mapping the obtained coordinates to a human skeleton model (inverse kinematics) and to export the data into a format that is supported by standard annotation tools (such as ANVIL).

1. Introduction: Detecting interactional patterns across disciplines

In recent years, a range of initiatives has begun to enable robots and other technical systems to engage in more naturalistic forms of interaction with the human user. After important advances have been made both in detecting/sensing human conduct and creating human-like forms of system output, a central challenge today consists in enabling technical systems to participate in and deal with the dynamic nature of human social interaction: Systems need to observe – on a micro-level – human multimodal conduct, interpret it as meaningful in terms of the interactional organisation and react appropriately. While there is a longstanding tradition in the field of Ubiquitous Computing and Computer Supported Cooperative Work (CSCW) to include qualitative approaches, such as Ethnography and/or Ethnomethodological Conversation Analysis (EM/CA), into the development cycle of technical systems (e.g. Dourish, 2009; Luff et al., 2009), only recently researchers have begun to scoop from these same sources for the design of robot systems (Nishida et al., 2007; Kuzuoka et al., 2008; Pitsch et al., 2009). In particular, for the design of robot systems, EM/CA – with its fine-grained analysis of video data – is able to provide insights into the sequential organisation of interaction, reveal patterns of social conduct and investigate how one person's multimodal conduct both reacts to and shapes their co-participants' actions. On the one hand, this offers a rich basis for modelling the dynamic and contingent nature of social interaction; on the other hand, the ways in which a qualitative, video-based EM/CA is able to present its findings do not always match the sort of quantifiable information that is required for building computational algorithms. Against this background, we argue that interactional corpora – combining video recordings and new motion capture technologies – are required that allow researchers to use different methodologies and research tools (qualitative/ quantitative; manual/automated) on the same data set (cf. Chen et al., 2006). However, with such an integrated methodological approach a range of new technical challenges arise regarding the constitution and management of multimodal corpora.

In this paper, we pinpoint the requirements and technical challenges that a combined approach brings to light with regard to establishing multimodal corpora (section 2), present our solution to solve these problems (section 3) and evaluate seemingly 'unnatural' aspects of our approach (section 4).

2. Corpus: Requirements and technical challenges

When planning and establishing a corpus that is designed to investigate multimodal turn-taking and other aspects of interactional organization in a group of two vs. three participants with a mixed approach of qualitative/ quantitative and manual/automated analysis, we have been largely informed by analytical experience from another ongoing interdisciplinary project (iTalk). We will use examples from this study to point out the requirements that the new corpus would need to fulfil.

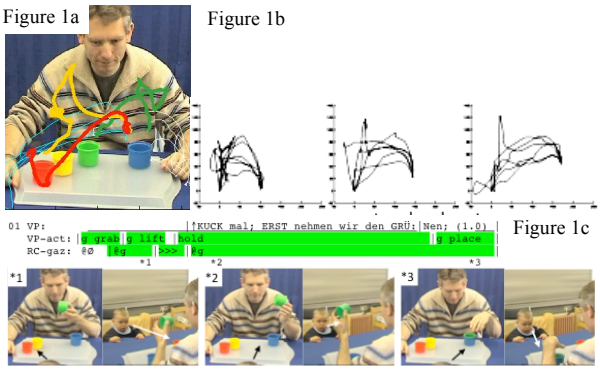


Figure 1: Parent demonstrating 'stacking cups' to his infant. (a) Video still with overlayed hand trajectories; (b) Normalized hand trajectories overlayed of several participants; (c) Transcript and stills from two cameras.

The iTalk project (www.italkproject.org) aims at

enabling robots to learn within and from the interaction with a human partner. Given the analogy of limited cognitive capabilities both in robots and young infants, our starting point consists in understanding the ways in which parents demonstrate actions to their young infants as a model for the design of the robot system (Rohlfing et al., 2006). In order to build the robot system, we need to know e.g. how participants structure their actions, which features are constitutive for tutoring, how the recipients react to the demonstration and how this, in turn, influences on the presenter's demonstration (Pitsch et al., 2009; Vollmer et al. 2009). Therefore, we have built and are analysing – with different research methodologies – a video corpus, in which 128 parents are demonstrating a set of actions to their infants aged 8 to 36 months. In this line of research, Ethnomethodological Conversation Analysis offers an interactional perspective on the task and is able to reveal with its qualitative-manual analysis the strategies and methods used by the participants, to uncover relevant multimodal features/cues and to find interactional patterns and systematic relationships between the co-participants' actions. At the same time, this approach is limited e.g. in describing the presenter's manual actions in terms of the concrete shape of the hand trajectory performed in a given demonstration. Interestingly, these shapes differ considerably in the corpus, which becomes visible once a semi-automatic computational 2D hand tracking is applied to the video data delivering time-stamped x,y-coordinates of the parent's hand motions (Fig. 1a, 1b). While EM/CA is able to reveal the interactional causes and effects of the variability in the hand trajectories (linked to the child's focus of attention, Fig. 1c), mathematical and statistical methods can describe these trajectories in a way that they become suitable for building computational algorithms that allow a robot to distinguish certain types of actions. At the same time, relevant interactional categories evolve from CA-analysis, which, then, can be systematically transcribed/annotated with corpus tools (such as ELAN, ANVIL) and be subject to a computational investigation of correlations between the different interactional variables on the entire corpus (Pitsch et al., 2009).

Not only does this example give a case for closely interrelated qualitative-quantitative analysis, but it also provides us with central *requirements* when establishing a new interactional corpus that is designed for the same area of research: We need to be able to capture (i) the timely interplay of *several* (two or more) participants, (ii) their talk, gaze, body posture, gestures, head, arm and body motions, and (iii) interactional episodes that take about 30 minutes of time. As – for the parent-child-corpus – we only dispose of video recordings of the interaction, we had to develop a motion tracker in order to be able to precisely describe and analyze the hand trajectories. While this has proven extremely useful for our case (and might be oriented towards the sensors that current robot systems are equipped with), analysis is limited with regard to the features that can be tracked robustly and by the fact that it can only deliver 2D information (information about depths is missing). Thus, (iv) for the new corpus both video and 3D motion capture data are required.

However, if we attempt to use current state of the art 3D motion capture technologies for recording data with the requirements presented above, we are facing a crucial *technical challenge*: How can we *robustly* track *multiple* participants over an *extended* period of time?

Existing optical motion capture technologies, such as the Vicon system, have been originally developed for capturing human motions in the fields of sports and health sciences or for animating virtual characters in movies and computer games. Small reflective markers (spheres) are attached to particular places of the human body, tracked simultaneously by a set of (at least 10) infrared cameras and mapped to a generic model skeleton. In these cases, generally *one* single participant is recorded for a *short* period of time. In recent times, researchers have begun to use such systems also for recording multi-party interaction (Chen et al. 2006; Battersby et al. 2008). However, once we attempt to use the system to track two or three participants during an interaction period of e.g. 30 minutes, we encounter a range of problems: (i) Due to visual obstruction, the system easily loses the individual markers during the recording process. (ii) This leads either to incomplete and thus problematic data or an extensive post-processing phase is required, in which markers need to be re-assigned and labeled. We have conducted a set of internal trials, which revealed that 1 minute of recording time requires about 60 minutes of post-processing for one participant – impossible to handle for large corpora.

3. The "Obersee Corpus": Suggestions for robustly tracking multiple participants

When establishing our corpus designed to investigate multimodal interactional organisation with a mixed methodological approach, we needed to find ways to by-pass the limits imposed by the current Vicon system, i.e. to *robustly* track *multiple* participants over an *extended* period of time. In what follows, we present our solution which involves both changes in hardware and new algorithms for transforming the raw motion capture data.

3.1 Study Design

As the corpus should allow for investigating a range of different aspects of multimodal interactional organization, we choose a semi-experimental set-up that would engage groups of participants in (a) a free conversation and (b) a task-related interaction which requires the use of material objects and gesticulation. At the same time, we needed to both control the situation in a way to allow for comparison between different groups of participants and to be open enough to allow for spontaneous social interaction. Therefore, we invited in total 15 groups of participants (6 dyades, 9 triades) to

engage in a 20 minutes conversation, in which they were supposed to discuss and come up with a solution for a redesign of the local lake (the "Obersee") into a new recreation area. We asked them to each assume a certain role (financial investor, local mayor, Greenpeace activist) and provided them with a map of the area as well as a set of toy objects (such as inline skater, diver, quad, barbeque) that they could use for inspiration and (re-)position on the map. Afterwards, they were asked to remain seated while the experimenter had to check the recording, get the questionnaires to be filled out, which provided us with further 10 minutes of free conversation.

3.2 Technical Setup

We recorded these interactions with four HD video cameras, ten Vicon T20 cameras and an additional microphone hanging from the ceiling (Fig.2). While the video footage was stored individually, the Vicon data was (i) firstly gathered and processed by a Vicon MX Giganet server, (ii) then sent to a PC using the Vicon Nexus software V1.4.112 to detect (patterns of) Vicon markers and to calculate their position and orientation and (iii) finally sent to another PC for saving the data. This setup allowed us to by-pass the limits of recording time and amount of data imposed by the Vicon Nexus software while using its pattern recognition facilities.

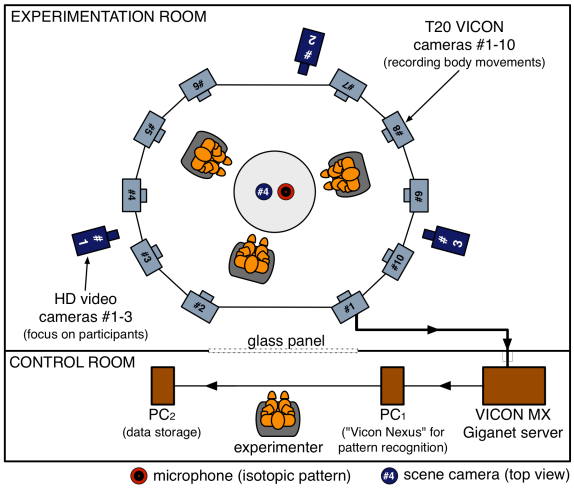


Figure 2: Technical setup

With this approach, we lose the function provided by the Vicon system of producing synchronized video and 3D motion capture data. We compensate for it by making one participant clap a slate at the beginning of the session, which creates a distinctive signal that can – afterwards – be automatically detected in the different media sources. In addition, a visual calibration pattern was positioned in the middle of the scene, so that we are able to calculate 3D information from the video footage.

3.3 Rigid bodies for robustly tracking three participants

In order to deal with the problem of losing markers and a resulting extensive post-processing, we decided to use – instead of individual markers – so-called "rigid bodies". A rigid body consists of a pattern of several markers that are spatially arranged in a particular way and can be distinguished from other rigid bodies (Fig. 3 and 4). It has a unique ID assigned by the marker, which, in turn, denotes the corresponding body part, so that it can be assigned to a position and an orientation in 3D space. The main advantage resides in the fact that – in case markers get lost – they can be automatically reassigned to each body limb by the system. Also, in the case of marker loss, chances are high that at least one or two markers (out of the set of five) are continuously tracked, so that limited information about the whereabouts of that particular body part will still be available. Consequently, no extensive manual post-processing is required.[1]

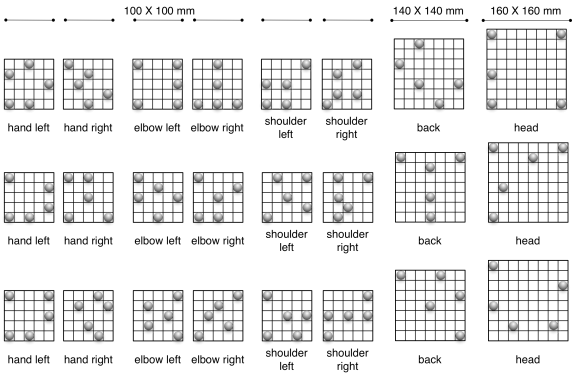


Figure 3: Three Sets of 8 rigid bodies worn by the participants

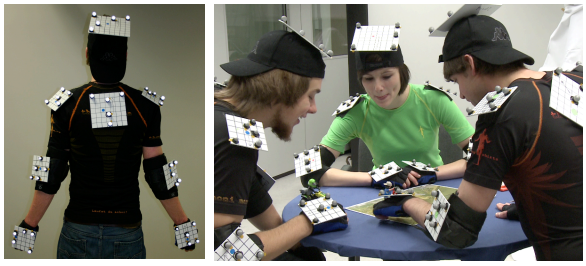


Figure 4: Participants wearing rigid bodies

In order to capture the most central movements of the human body, which are supposed to be interactionally relevant in a seated face-to-face setting, we used eight rigid bodies per person. These were attached to the head, back, left/right shoulders, left/right elbows and left/right hands (Figure 3). As we wanted to robustly track three participants simultaneously, we had to provide a set of 24 rigid bodies that were clearly distinguishable from each other. While we started with a systematic arrangement of markers on a 10 cm x 10 cm grid (allowing for a 5 by 5 grid), we soon had to increase the grid size to 16 cm x 16 cm (8 by 8 grid) to be able to create enough patterns that the Vicon Nexus software

[1]Systematic evaluation of this approach will be undertaken.

would robustly recognize as distinct. [2]
While the size of the rigid bodies was determined by the technical feasibility and robustness, we were concerned to keep their size as small and unobtrusive for the participants as possible. Initial pre-trials suggested that participants would rather tolerate the larger rigid bodies attached to their back and the top of their head, and could cope with the 10cm x 10cm grids at the other positions if they were fixed appropriately (e.g. by using thin fingerless biker gloves). Being aware that rigid bodies could potentially influence the participant's "natural" conduct in the experiment, we used a questionnaire to evaluate their experience of our setup (Section 4).

3.4 Skeleton representation and inverse kinematics

While our approach to by-pass the limits of the standard Vicon system (rigid bodies, "Vicon Nexus" software for detecting patterns of markers and giving their location and orientation, external data storage) allows us to robustly capture three participants over a long period of time, we have to find ways to map the rigid body's coordinates to a human skeleton model to calculate the joint angles.

To calculate the joint angles of the tracked person, we use a mathematical representation of the human skeleton based on the Denavit-Hartenberg-Convention developed and used in the field of robotics. It describes the transformation of a single joint with one degree of freedom to the next adjacent joint. For this, it uses four elementary transformations: $A_i = R_{z\,i-1}, * T_{zi-1} * T_{xi} * R_{xi}$

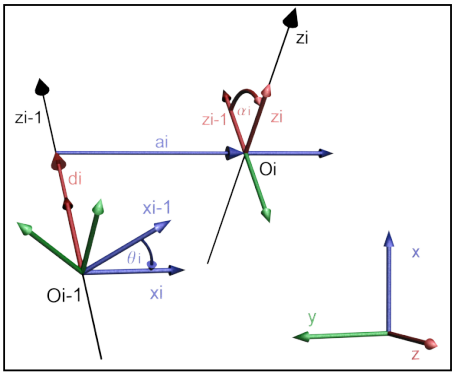


Figure 5: The four elementary transformation from one joint axis to the next adjacent joint

The transformation A_i contains a rotation around the previous z-axis, a translation along the previous z-axis, another translation along the current x-axis and a rotation around the current x-axis. Such a transformation can be used to model either a complete human skeleton or a single arm etc. (Fig.5). A single rotation joint is represented as a cylinder and has the ability to rotate around the z-axis which is parallel to the height of the cylinder (see Fig.6 where e.g. the shoulder has got three joints represented as cylinders).

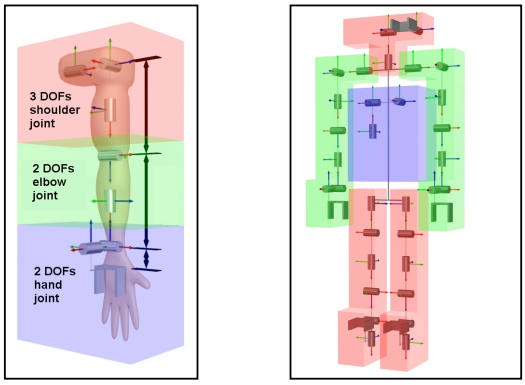


Figure 6: (a) Arm and (b) complete body representation in the Denavit Hartenberg Convention

Based on the mathematical description of the skeleton, we have developed algorithms that firstly calculate the positions of the joints out of the rigid body coordinates. Secondly, we proceed with inverse kinematics, in which the angles of each joint are calculated using the tangent = sinus/cosine = adjacent/opposite = y0/x0 (Spong et al., 2006; Brüning et al., 2008).

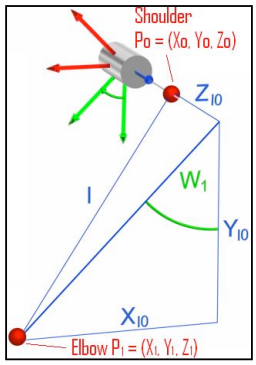


Figure 7: Inverse kinematics – Calculation of a single joint angle from a local joint coordinate system

From these calculations, we obtain the angles for one joint, which we then have to apply for all joints for each individual participant. When applying this procedure, we start by localizing the human body in space (i.e. the marker attached to the participant's back) and from there proceed by calculating step by step each further joint.

[2] While the rigid bodies and their locations were robustly tracked, initial investigation of the motion capture data showed slight problems for four markers, where – at moments – the orientation of the marker could not be precisely tracked. This can be caused by different factors (positioning of cameras, obstruction, the marker itself) and more detailed analysis of the causes will be required. At the current state, we used entire plastic plates as the basis for the rigid bodies. In a next iteration, we might consider cutting out the 'unused' space to reduce their obtrusiveness for the user. This, however, will need further consideration regarding mirror-invariance in the patterns.

3.5 Displaying data and integration into existing annotation tools

Once we have obtained the angles for all joints, we can display a skeleton of the human participant showing its posture at a given moment in time during the interaction (Fig.8a). The motion capture data also allows us to display and analyze in 3D the motion trajectories that the participants perform (Fig.8b).

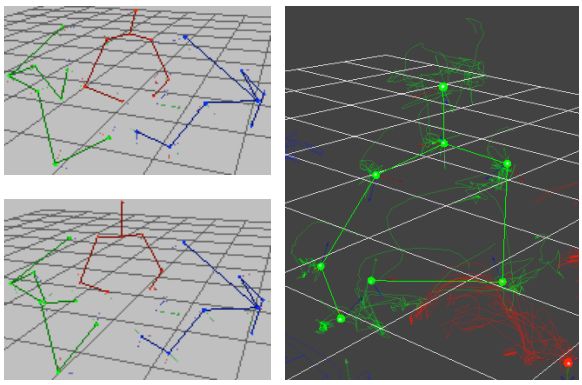


Figure 8: (a) Representations from the current scene and (b) including motion trajectories of one participant

In order to link the motion capture data with the video and sound files, we export the data obtained into a format that is supported by standard annotation tools (such as ANVIL) which are used by Conversation Analysts for transcription and annotation. To do so, we have developed a program that exports the motion capture data to the BVH (biovision hierarchy) format which is supported in the current version of the ANVIL annotation tool (Kipp et al., 2010). This file format consists of two main parts: one containing a description of the hierarchical order of the joints that describe the skeleton with the offsets from one joint to another; the other one comprising the angles of all joints written in the order of their hierarchical arrangement.

However, at the current state, ANVIL only supports motion capture data displaying *one* human; extensions will be required to also include the appropriate display of the interactional organization between *multiple* participants.

4. User Experience: How obtrusive are rigid bodies for the participants?

When developing our approach of using rigid bodies we were concerned with the question to which extent these objects might be – when being attached to the human participants – uncomfortable to wear and obtrusive for interacting or grabbing objects. While initial pre-trials suggested this approach to be acceptable, we wanted to evaluate the participants' experience more systematically. Therefore, after the experiment, we asked all participants to fill out a short questionnaire collecting information about their experience with regard to participating in (semi-)experimental studies, being videotaped and having used motion capture systems before.

In particular, two aspects are of interest here. We asked whether the participants felt disturbed during their interaction (i) by being videotaped and (ii) by the rigid bodies attached to their different body parts. Analysis reveals that in general, participants feel only 'slightly' disturbed by the recording equipment with a *similar* distribution between (i) being video-taped and (ii) having rigid bodies attached to their bodies (2x s-field- X^2-Pearson´s chi square test, with α=0,05). This result confirms our initial observations from the pre-trials and suggests that the rigid bodies do not seem to create more a unauthentic situation than video recordings – with the latter being recognized as a standard method of data acquisition in research.

Considering the answers for the motion capture in detail, we find that the participants' disturbance with regard to hand and elbow markers shows a tendency for slight disturbance while they feel hardly, i.e. 'less than slightly', disturbed with head, shoulder and back markers (Fig. 9).

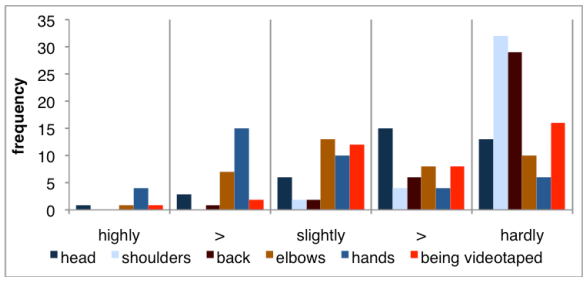
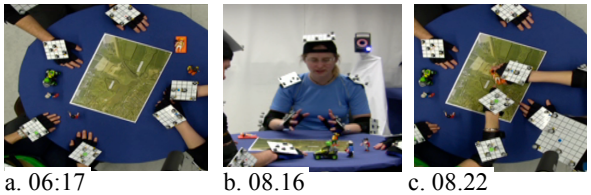


Figure 9: How disturbed do participants feel when (i) being videotaped and (ii) using rigid body markers?

These results and their analogy with the video recording situation suggest that our approach of using rigid bodies for overcoming the problem of robustly tracking multiple participants might able to generate – both technically and socially – valid interactional data.

In addition to asking users about their experience, close examination of the video recordings should allow to explore in more detail the rigid bodies' impact on the users' conduct in situ and the potential form of disturbance they might cause. Initial analysis of one group reveals that participants, at the beginning of the experiment, appear to position their hands rather flat on the table and without much manual actions or motions (Fig. 10a). This, however, changes step by step as the interaction unfolds. Around 8 minutes in the recording – when the participants are immersed in their roles and tasks – the first instance of gesticulation can be observed (Fig. 10b), and participants begin to bring their hand (and markers) close to some body else's hand (and markers) while manipulating objects on the plan (Fig. 10c). At this time, also vertical hand positions begin to occur, which suggests that participants are not particularly concerned (any more) with the question of the rigid bodies' adherence or trackability (Fig. 10d). After 17 minutes, participants can be seen to approach their hands even closer to the co-participant's hands (Fig. 10e) and to also reach to the other side of the table while

crossing their co-participants' arms and markers (Fig. 10f).



a. 06:17 b. 08.16 c. 08.22

d. 08.25 e. 17.52 f. 17.54

Figure 10: Participants' gestures and manipulation of objects changing over time

These observations suggest that during the first 5 to 8 minutes of an experiment participants seem to use more controlled hand motions and gestures, while after this initial period both their individual motions and their collaboration with others become more vivid. However, the question whether this effect is linked to the general situation of being observed or whether it might be caused specifically by the rigid bodies needs further consideration. Further analysis will also need to include other groups of participants, to investigate the motion of different body parts and begin to link the participants' motions to the concrete interactional tasks being carried out.

5. Conclusion and Future Work

In this paper, we have presented a system that is able to *robustly* track and record *multiple* participants over an *extended* period of time (30 minutes) with a 3D motion capture system. Linking this data to four HD video recordings, we are able to establish a multimodal corpus that is suitable for a combined qualitative/quantitative corpus analysis. The recorded data from the different sources can be analyzed using both Conversation Analysis and mathematical/statistical methods.

Our approach consists of by-passing the limits of the current standard Vicon system (using rigid bodies) and ways of mapping the obtained coordinates to a human skeleton model (inverse kinematics) and to export the data into a format that is supported by standard annotation tools (such as ANVIL). With regard to traditional motion capturing the following main differences can be summarized as follows:

Aspects	**Motion Capturing**	
	Traditional	**Rigid bodies**
Preparation (w/o subject)		- Build rigid bodies
Preparation (with subject)	- Attach 18 markers per user - Map markers to the body parts	+ Attach 8 rigid bodies per user
Comfort for subjects	+ More comfortable	- Less comfortable
Stability of tracking	- Markers lost easily - Once marker is lost, the system doesn't know the position of that body part until the post processing	+ Set of 5 markers more stable to track + Once rigid body is lost, it can be automatically reassigned to each body limb by the system
Data saving	- After recording. Time consuming	+ Real-time
Post processing	- Map marker (that got lost) to the corresponding body part	+ None. Rigid body is always attached to a specific body part

A first evaluation of the setup suggests that the use of rigid bodies does not create more an unauthentic situation than do video recordings.

Next steps consist in further evaluating the impact of the rigid bodies on the user's conduct, and we aim to establish automated ways of detecting typical motions to allow for more automated ways of corpus annotation.

6. Acknowledgements

The authors gratefully acknowledge the financial support from the Cluster of Excellence "Cognitive Interaction Technology" and the EU-funded project "iTalk".

7. References

Battersby, S.A., Lavelle, M., Healey, Patrick G.T. & McCabe, R. (2008): *Analysing Interaction: A comparison of 2D and 3D techniques.* In: Proceedings Workshop on Multimodal Corpora (LREC 2008), 73-76.

Brüning, B. (2008): *Entwicklung eines Motion Capture Recorders für einen virtuellen Agenten auf der Basis eines optischen Trackingsystems*, Bielefeld University: Diploma thesis.

Brüning, B., Latoschik, M. E., Wachsmuth, I. (2008), *Interaktives Motion Capturing zur Echtzeitanimation virtueller Agenten*, In VRAR.

Chen, L., Rose, R. T., Qiao, Y., McNeill, D., & Harper, M. (2006). *VACE multimodal meeting corpus*. In S. Renals & S. Bengio (Eds.), Machine learning for multimodal interaction. (pp. 40-51). Heidelberg: Springer.

Dourish, P. (2001): *Where the Action Is*. Foundations of Embodied Interaction. MIT Press.

Kipp, M. (2010) *Multimedia Annotation, Querying and Analysis in ANVIL*. In: M. Maybury (ed.) Multimedia Information Extraction, Chapter 19, MIT Press.

Kuzuoka, H., Pitsch, K., Suzuki, Y., Kawaguchi, I., Yamazaki, K., Kuno, Y., et al. (2008). *Effects of restarts and pauses on achieving a state of mutual gaze between a human and a robot.* In CSCW 2008, 201-204.

Luff, P., Pitsch, K., Heath, C., Herdman, P., & Wood, J. (2009). *Swiping paper and the second hand: Mundane artefacts, gesture and collaboration.* Journal of

Personal and Ubiquitous Computing, 213-224.
Nishida, T. (2007). *Conversational informatics: An engineering approach*. Wiley.
Pitsch, K., Kuzuoka, H., Suzuki, Y., Süssenbach, L., Luff, P., & Heath, C. (2009). *"The first five seconds". Contigent stepwise entry into an interaction as a means to secure sustained engagement.* In RO-MAN 2009, 985-991.
Pitsch, K., Vollmer, A.-L., Fritsch, J., Wrede, B., Rohlfing, K., & Sagerer, G. (2009). *On the loop of action modification and the recipient's gaze in adult-child-interaction.* In GESPIN 2009. Poznan, Poland.
Rohlfing, K., Fritsch, J., Wrede, B., & Jungmann, T. (2006). *How can multimodal cues from child-directed interaction reduce learning complexity in robots?* Advanced Robotics, 20(10), 1183-199.
Spong, M. W., Hutchinson, S., Vidyasagar, M. (2006). Robotik Modeling And Control, Wiley.
Vollmer, A.-L., Lohan, K., Fischer, K., Nagai, Y., Pitsch, K., Fritsch, J., Rohlfing, K.J. & Wrede, B. (2009): *People Modify Their Tutoring Behavior in Robot-Directed Interaction for Action Learning*. In. IDCL 2009.

3D Vision Technology for Capturing Multimodal Corpora: Chances and Challenges

Gabriele Fanelli[1], Juergen Gall[1], Harald Romsdorfer[2], Thibaut Weise[3], Luc Van Gool[1]

[1]Computer Vision Laboratory, ETH Zurich, Switzerland
[2]Signal Processing & Speech Communication Laboratory, Graz University of Technology, Austria
[3]Laboratoire d'Informatique Graphique et Géométrique, EPFL Lausanne, Switzerland
{gfanelli, gall, vangool}@vision.ee.ethz.ch, romsdorfer@tugraz.at, thibaut.weise@epfl.ch

Abstract

Data annotation is the most labor-intensive part for the acquisition of a multimodal corpus. 3D vision technology can ease the annotation process, especially when continuous surface deformations need to be extracted accurately and consistently over time. In this paper, we give an example use of such technology, namely the acquisition of an audio-visual corpus comprising detailed dynamic face geometry, transcription of the corpus text into the phonological representation, accurate phone segmentation, fundamental frequency extraction, and signal intensity estimation of the speech signals. By means of the example, we will discuss the advantages and challenges of integrating non-invasive 3D vision capture techniques into a setup for recording multimodal data.

1. Introduction

Multimodal corpora are an important resource for studying and analyzing the principles of human communication. Besides speech, the visual modality encodes probably the most important cues for humans to perceive communicative behavior like hand gesture, facial expression, or body posture. The recent efforts to collect audio-visual corpora are reflected in the literature, see (Zeng et al., 2009; Douglas-Cowie et al., 2007; Cowie et al., 2005), for instance.

The most labor-intensive part for acquiring a multimodal corpus is the annotation of the data, in particular for the visual modality. Although there are coding schemes like the Facial Action Coding System (FACS) (Ekman and Friesen, 1978) or annotation tools like ANVIL (Kipp, 2008) supporting the manual labeling of video sequences, the commonly used 2D video recordings inherently lead to a loss of information by projecting the 3D content onto the 2D image plane and discarding the depth information. Fortunately, 3D displays and capture technologies have emerged as commercial products. One example are invasive methods that place markers on the human body or face to capture the movement. However, the attachment of markers is not only time-consuming and often uncomfortable for the subject to wear, it can also significantly change the pattern of motion (Fisher et al., 2003). In general, invasive methods cannot be used for the acquisition of authentic data. Non-invasive vision systems provide a valuable alternative solution. Similar to the human vision system, they capture the scene with two or more 2D sensors to obtain the depth information. Current systems estimate the 3D surface deformation of the human body (Gall et al., 2009) or the face (Weise et al., 2009a) with minimal manual effort, as shown in Figure 1. In contrast to discrete annotation schemes like FACS which divides the state space into 32 facial muscles activation units, the continuously captured 3D data streams do not suffer from quantification artifacts as they capture the strength of the deformation accurately and consistently over time.

In this paper, we give an example for the use of 3D vision technology for the acquisition of an audio-visual corpus that comprises detailed dynamic face geometry, transcription of the corpus text into the phonological representation, accurate phone segmentation, fundamental frequency extraction, and signal intensity estimation of the speech signals. By means of the example, we will discuss the advantages and challenges of integrating non-invasive 3D vision capture techniques into a setup for recording multimodal data. The corpus will be released for research purposes at the end of the year.

Figure 1: From left to right, the image shows the 3D reconstruction of a person's face, the corresponding texture mapped on it, and the spatio-temporal consistent representation of the face surface.

2. 3D Face Capture System

Our goal is to capture the face geometry of a speaker over time and represent the data such that it serves as annotation for the visual modality. The first step requires the reconstruction of the depth map of the subject's face for each frame, as shown in Figure 1. To this end, we employ the 3D scanner described in (Weise et al., 2007). The system combines stereo and active illumination based on phase-shift for robust and accurate 3D scene reconstruction. Stereo overcomes the traditional phase discontinuity problem and motion compensation is applied in order to remove artifacts in the reconstruction. The system consists of two high-speed grayscale cameras, a color camera, and a DLP projector without the 4-segment color wheel (RGBW), so that it

projects three independent monochrome images at 120 Hz (sent as the R, G, and B channel). The two monochrome cameras are synchronized and record the three images. The color camera is also synchronized, but uses a longer exposure to integrate over all three projected images and thus capture the texture. In our current setting, the system is very stable at 25 Hz although the recording could be performed at up to 40 Hz. The acquisition rate is a limitation of many vision systems since higher frame rates require short exposure times and thus very bright light. Movements which are faster than the acquisition rate cannot be captured, nevertheless, the system accurately captures texture and the depth map of the face, as shown by the first two images in Figure 1.

The depth map of a face is a cloud of 3D points that cannot be directly used for analyzing changes in the face geometry as this requires full spatial and temporal correspondences of the 3D data. For example, a point of the left eyebrow should be also part of the left eyebrow for all captured faces, i.e., 3D points should maintain their semantic meaning among different scans. To achieve this goal, we use the two-step procedure introduced in (Weise et al., 2009a): First, a generic template mesh is warped to the reconstructed, expressionless, 3D model of the speaker. Second, the resulting personalized template is automatically tracked throughout all recorded sequences of the same speaker. The template ensures spatial consistency over time and between different subjects.

Preprocessing In order to build a person-specific face template, each speaker is asked to turn the head with a neutral expression and as rigidly as possible in front of the real-time 3D scanner. The sequence of 3D scans is registered and integrated into one 3D model using the online modeling algorithm proposed in (Weise et al., 2009b). Small deformations arising during head motion violate the rigidity assumption, but in practice do not pose problems for the rigid reconstruction. Instead of using the reconstructed 3D model directly, a generic face template is warped to fit it. Besides enabling a hole-free reconstruction and a consistent parameterization, using the same generic template has the additional benefit of providing full spatial correspondence between different speakers.

Warping the generic template to the reconstructed 3D model is achieved by means of non-rigid registration, where for each mesh vertex $\mathbf{v}_i$ of the generic template a deformation vector $\mathbf{d}_i$ is determined in addition to a global rigid alignment. This is formulated as an optimization problem, consisting of a smoothness term minimizing bending of the underlying deformation (Botsch and Sorkine, 2008), and a set of data constraints minimizing the distance between the warped template and the reconstructed model. As the vertex correspondences between generic template and reconstructed model are unknown, closest point correspondences are used as approximation similarly to standard rigid iterative closest point registration. A set of manually labeled correspondences are used for the initial global alignment and to initialize the warping procedure. The landmarks are mostly concentrated around the eyes and mouth, but a few correspondences are selected on the chin and forehead to match the global shape. The manual labeling needs to be done only once per speaker and takes at most a couple of minutes. The resulting personalized template accurately captures the facial 3D geometry of the corresponding person.

Figure 2: Recording setup: one speaker sits in front of the 3D scanner in the anechoic room while watching a video for instructions on the screen.

The diffuse texture map of the personalized template is automatically extracted from scans where the subject moves the head rigidly, by averaging the input textures. The face is primarily illuminated by the 3D scanner, and we can therefore compensate for lighting variations using the calibrated position of the projection. Surface parts that are likely to be specular are automatically removed. The reconstructed texture map is typically oversmoothed, but sufficient for the tracking stage.

3D Tracking The personalized face template is used to track the facial deformations of each performance. For this purpose, non-rigid registration is employed, in a similar manner as during the template creation. In this case, the distances between the template vertices and the 3D scans are minimized. To ensure temporal continuity, optical flow constraints are also included in the optimization, as the motion of each vertex from frame to frame should coincide with the optical flow constraints. During speaking, the mouth region deforms particularly quickly, and non-rigid registration may drift and ultimately fail. This can be compensated for by employing additional face-specific constraints such as explicitly tracking the chin and mouth regions, making the whole process more accurate and robust to fast deformations. Figure 1(right) shows a personalized model adapted to a specific frame of a sequence.

3. Example: Audio-Visual Corpus

In order to acquire and annotate data for an audio-visual corpus, we have integrated the 3D face capture system into a setup for recording audio data. To obtain clean audio data, the setup is placed in an anechoic room with walls covered by sound wave-absorbing materials. For the audio recordings, we use a studio condenser microphone placed in front of the speaker. Since the cooling fans of the projector for the 3D face capture system become very noisy, we have built a wooden enclosure, open at the back to allow the heat out and equipped with a non-reflecting glass at the front to emit the light. The enclosure reflects most of the noise

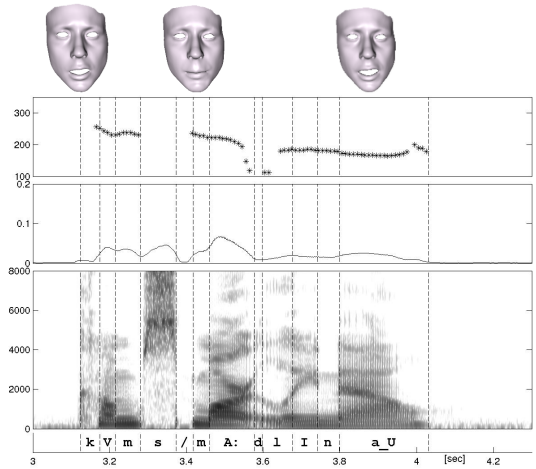


Figure 3: Phone sequence, spectrogram, signal intensity contour, and fundamental frequency contour of the speech signal, plus sample faces are shown from bottom to top.

vowels	i: I U u: e @ q 3 3: V O: A A: Q
diphthongs	@_U a_I a_U e_I E_@ I_@ O_I o_U U_@
consonants	p p_h b t t_h d k k_h g m n N r f v T D s z S Z x h j w l
affricates	t_S d_Z
pauses	c_u c_v /

Table 1: Segment types of English phones and speech pauses used for transcription of the speech data of the audio-visual corpus.

to the back where it is absorbed by the walls. The signal-to-noise ratio (SNR) of the recorded audio stream is high enough to perform automatic speech segmentation, which is described in Section 3.2. Figure 2 shows the setup, with a volunteer being scanned while watching a video for instructions on the screen. The computers for processing the data are placed in a separate room, from where the system can be fully controlled.

3.1. Corpus

Our corpus currently comprises 40 short English sentences spoken by 14 native speakers (8 females and 6 males, aged between 21 and 53) who volunteered to have their voice and facial movements recorded. Each person was required to sit alone in the anechoic room and asked to pronounce each sentence. For synchronization purposes, the volunteers clapped their hands in front of the cameras before uttering the sentences. The clapping is automatically detected in the audio stream and used for cutting the 3D video stream. This introduces a maximum temporal discrepancy of 20 ms between the audio and video stream. A hardware synchronization could improve the accuracy. We have also recorded a corpus for affective speech, described more in detail in (Fanelli et al., 2010).

3.2. Audio Annotation

The auditory modality of the corpus is best annotated by means of the speech prosody and the sequence of phones. Speech prosody can be described at the perceptual level in terms of pitch, sentence melody, speech rhythm, and loudness. The physically measurable quantities of a speech signal are the following acoustic parameters: fundamental frequency (F_0), segment duration, and signal intensity. F_0 correlates with pitch and sentence melody, segment duration with speech rhythm, and signal intensity with loudness. Figure 3 shows an example of the annotated audio-visual corpus.

The annotation process necessary for obtaining the physical prosodic parameters of the utterances includes a number of steps: First, the sentence's text is transcribed into the phonological representation of the utterance. Then accurate phone segmentation, fundamental frequency extraction, and signal intensity estimation are achieved by analyzing the speech data. For the extraction of these prosodic quantities, we applied fully automatic procedures provided by SYNVO Ltd. In the following, we give an overview of the extraction procedures for fundamental frequency, signal intensity, and segment duration.

Transcription The phonological representation contains the sequence of phones for the sentences in the audio-visual corpus, the stress level of syllables, the position and strength of phrase boundaries, plus the indicators of phrase types. Initial phonological representations of the sentences are obtained by applying the transcription component of the SYNVO text-to-speech synthesis system to the text version of the corpora. See (Romsdorfer and Pfister, 2007) for a description of such a transcription component. These initial phonological representations contain the standard phonetic transcription of the sentences (Table 1).

The phonological information (phrase type, phrase boundary, and sentence accentuation) of these automatically generated representations is then adapted to the speech signals. Neural network-based algorithms are employed for automatic phrase type, phrase boundary, and syllable accent identification. Detailed information on this procedure can be found in (Romsdorfer, 2009).

Fundamental Frequency Extraction F_0 values of the natural speech data of the prosody corpus are computed every 10 ms using a pitch detection algorithm based on combined information taken from the cepstrogram, the spectrogram, and the autocorrelation function of the speech signal, cf. (Romsdorfer, 2009). Signal sections judged as unvoiced by the algorithm are assigned no F_0 values.

Signal Intensity Extraction Signal intensity values of the natural speech data are computed every 1 ms. The root mean square value of the signal amplitude calculated over a window of 30 ms length is used.

Segment Duration Extraction An accurate extraction of phone and speech pause durations requires an exact segmentation of the natural speech data of the audio-visual corpus into adjacent, non-overlapping speech or pause segments, and a correct assignment of labels to these segments indicating their type.

Since the phonological representation contains the standard phonetic transcription of an utterance, it is convenient to use such transcription for automatic segmentation and labeling. However, a close phonetic transcription, indicating pronunciation variants made by the speaker, results in a much better segmentation and labeling.

Segment Types Segment types correspond to the phone types determined in the transcription. Plosives are additionally segmented into their hold and burst parts. While the burst part of a plosive is denoted by the same symbol used for the plosive phone type, a "ᴄ" denotes the hold part, also called closure or preplosive pause. Speech pauses corresponding to phrase boundaries are labeled with the symbol "/". For a plosive following a speech pause, no preplosive pause is segmented. Table 1 lists all segment types used for the transcription of natural speech data.

4. Chances and Challenges

State-of-the-art 3D vision technology opens new opportunities for acquiring and annotating the visual modality of multimodal corpora with minimal manual effort. We have shown by means of an example that 3D capture devices can be integrated into a setup for multimodal data acquisition. The problem of inference between the devices can often be solved sufficiently. In our example for instance, the microphone was placed outside of the field of view of the cameras and the visual capture devices were modified for noise reduction. A hardware synchronization between the capture devices for the various modalities is desirable although it is currently not implemented in our example where only the vision components are synchronized.

In contrast to marker-based systems, advanced multi-camera vision systems are non-invasive and consequently better suitable for capturing authentic data. Compared to 2D labeling schemes, they automate most of the annotation process, significantly reducing the costs for corpora acquisition. Due to the current trend of developing 3D hardware technology like cameras and displays for consumer applications, it's very likely that the prices for ready-to-use systems will drop over the next few years. The probably most appealing property of 3D vision technology is the ability to capture continuous surface deformations accurately and consistently over time. This provides a richer source of information than discrete annotation schemes.

However, there are also several limitations that need to be pointed out. The accuracy of the vision components usually degenerates at bad lighting conditions. In particular, arbitrary outdoor environments are very challenging for most systems. The speed of motion that can be captured is limited to the acquisition frame rate, which is typically in the range of 25 Hz and 40 Hz. The accuracy of non-invasive methods does not match yet the accuracy of marker-based systems. For instance, the 3D reconstruction shown in Figure 1 is sometimes noisy for the eyelids and the teeth which can result in errors around the eyes and for the estimated mouth shape. Resolving these problems is a challenging task for the future.

5. Acknowledgements

This work was partly funded by the SNF fund "Vision-supported speech-based human machine interaction".

6. References

M. Botsch and O. Sorkine. 2008. On linear variational surface deformation methods. *IEEE Trans. on Visualization and Computer Graphics*, 14:213–230.

R. Cowie, E. Douglas-Cowie, and C. Cox. 2005. Beyond emotion archetypes: Databases for emotion modelling using neural networks. *Neural Networks*, 18(4):371–388.

E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis. 2007. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *Int. Conf. on Affective Computing and Intelligent Interaction*.

P. Ekman and W. Friesen. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press.

G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool. 2010. Acquisition of a 3d audio-visual corpus of affective speech. Technical Report 270, ETH Zurich, January.

D. Fisher, M. Williams, and T. Andriacchi. 2003. The therapeutic potential for changing patterns of locomotion: An application to the acl deficient knee. In *ASME Bioengineering Conference*.

J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. 2009. Motion capture using joint skeleton tracking and surface estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition*.

M. Kipp. 2008. Spatiotemporal coding in anvil. In *Int. Conf. on Language Resources and Evaluation*.

H. Romsdorfer and B. Pfister. 2007. Text analysis and language identification for polyglot text-to-speech synthesis. *Speech Communication*, 49(9):697–724.

H. Romsdorfer. 2009. *Polyglot Text-to-Speech Synthesis. Text Analysis and Prosody Control*. Ph.D. thesis, No. 18210, Computer Engineering and Networks Laboratory, ETH Zurich (TIK-Schriftenreihe Nr. 101), January.

T. Weise, B. Leibe, and L. Van Gool. 2007. Fast 3d scanning with automatic motion compensation. In *IEEE Conf. on Computer Vision and Pattern Recognition*.

T. Weise, H. Li, L. Van Gool, and M. Pauly. 2009a. Face/off: Live facial puppetry. In *Symposium on Computer Animation*.

T. Weise, T. Wismer, B. Leibe, and L. Van Gool. 2009b. In-hand scanning with online loop closure. In *IEEE Int. Workshop on 3-D Digital Imaging and Modeling*.

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

Multi-level Annotations of Nonverbal Behaviors in French Spontaneous Conversations

Ning TAN[1], Gaelle FERRE[2], Marion TELLIER[3], Edlira CELA[4], Mary-Annick MOREL[4], Jean-Claude MARTIN[1], Philippe BLACHE[3]

[1] LIMSI-CNRS BP 133
91403 Orsay Cedex France
[2] Centre International de Langues - Département d'Etudes Anglaises - Chemin de la Censive du Tertre BP 81227
44312 Nantes Cedex 3 France
[3] Université de Provence UFR LACS - Département de FLE 29 avenue Robert Schuman
13621 Aix-en-Provence Cedex 1 France
[4] Université Paris 3 UFR de Littérature et Linguistique Françaises et Latines, CLF, 13 rue Santeuil
75005 Paris France
E-mail: {ntan, martin}@limsi.fr, gaelle.ferre@univ-nantes.fr, marion.tellier@univ-provence.fr, edlira.cela@yahoo.fr, marym@edilang.com, blache@lpl-aix.fr

Abstract

During spontaneous conversations, multiple modalities such as gesture, posture and gaze are combined in sophisticated ways for different functions such as spatial references. In the French research community, there is a lack of spontaneous multimodal corpora for the French language. This paper describes the multi-level scheme that we have defined for the annotation of gesture, posture and gaze. We explain how we applied it for the annotation of a corpus of spontaneous French conversations. Two types of analyses were made on the resulting annotations. Firstly, intercoder agreement was computed on the annotations of gesture space. Secondly, we clustered chunks of postures into categories. Such research will enable the joint study of multiple nonverbal modalities such as the relations between gesture semantics and posture.

1 Introduction

During spontaneous conversations, multiple modalities such as speech, gesture, posture and gaze are combined in sophisticated ways. Several studies considered how speech and gestures co-occur. Yet, studies on the way in which the different nonverbal modalities interact are deficient, partly due to the lack of accessible and relevant resources. For example, the multimodal corpora that are currently available for the French research community are limited in terms of accessibility, spontaneity and levels of annotation. In addition, studying relations between those modalities requires the definition of reliable coding schemes, describing multiple levels from the phonological to the gestural andpostural levels.

The present work lies within the framework of the OTIM project (Blache et al., 2009), which aims at building such French digital corpus of multimodal behaviors occurring during spontaneous conversations. The project makes use of the Corpus of Interactional Data (CID) which is an audio-video database of spontaneous spoken French (Bertrand et al., 2008). Eight pairs of native French speakers take part separately in a one-hour dyadic spontaneous spoken conversation, in which each speaker tells unexpected personal experiences (Figure 1). The CID corpus is one of the first multimodal corpora in French.

In this paper we focus on a subset of nonverbal modalities which are relevant for studying *spatial references* occurring during conversations. This function is shared by several nonverbal modalities such as gesture, gaze and posture. The speech and linguistic aspects are beyond the scope of this paper (Bertrand et al., 2008). After specifying a coding scheme for gesture, postures and gaze (sections 2-4), we measure the inter-coder agreement between three independent coders on a subpart of the annotations (section 5.1). Finally, we study the relations between posture and gesture (section 5.2).

Figure 1: The CID French spontaneous conversation corpus.

2 Gesture

Different typologies have been adopted for the classification of gestures, based on the work by Kendon (1980) and McNeill (1992, 2005). The formal model we use for the annotation of hand gestures is adapted from the specification files created by Kipp (2003) and from the MUMIN coding scheme (Allwood et al., 2005). Both models consider McNeill's research on gestures (1992,

2005). The gesture types we are using are mostly taken from McNeill's work. *Iconics* present "images of concrete entities and/or actions", whereas *Metaphorics* present "images of the abstract", they "involve a metaphoric use of form" and/or "of space". (McNeill, 2005: 39). *Deictics* are pointing gestures and *Beats* bear no "discernible meaning" and are rather connected with speech rhythm (McNeill, 1992: 80). *Emblems* are conventionalized signs and *Butterworths* are gestures made in lexical retrieval. *Adaptors* are non verbal gestures that do not participate directly in the meaning of speech since they are used for comfort. Although they are not linked to speech content, we decided to annotate these auto-contact gestures since they give relevant information on the organization of speech turns.

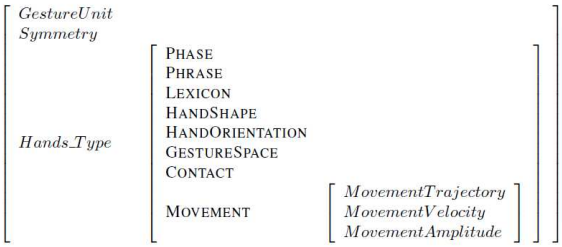


Figure 2: Formal model for the annotation of hand gestures.

We used the Anvil tool (Kipp 01) for the manual annotations. The changes we made concerned rather the organization of the different information types and the addition of a few values for a description adapted to the CID corpus. For instance, we added a separate track *'Symmetry'*. In case of a single-handed gesture, we coded it in its *'Hand_Type'*: left or right hand. In case of a two-handed gesture, we coded it in the left *Hand_Type* if both hands moved in a symmetric way or in both *Hand_Types* if the two hands moved in an asymmetric way. For each hand, the scheme has 10 tracks, enabling to code phases, phrases (the semiotic types being given in Table 1) (Figure 2). We allowed the possibility of a gesture pertaining to several semiotic types using a boolean notation. A gesture phrase (i.e. the whole gesture) can be decomposed into several gesture phases i.e. the different parts of a gesture such as the preparation, the stroke (the climax of the gesture), the hold and the retraction (when the hands return to their rest position) (McNeill, 1992). The scheme also enables us to annotate the gesture lemmas (Kipp, 2003:237), the shape and orientation of the hand during the stroke, the gesture space (where the gesture is produced in the space in front of the speaker's body, McNeill, 1992:89), and contact (hand in contact with the body of the speaker, of the addressee, or with an object). We added the three tracks to code the hand trajectory (adding the possibility of a left-right trajectory to encode two-handed gestures in a single Hand_Type, and thus save time in the annotation process), gesture velocity (fast, normal or slow) and gesture amplitude (small, medium and large). A gesture may be produced away from the speaker in the extreme periphery, while has very small amplitude if the hand was already in this part of the gesture space.

75 minutes of the CID involving 6 speakers have been coded for hand gestures. The annotation yielded a total number of 1477 gestures. The numbers of hand gestures per semiotic type are listed in the table below.

Adaptors	334
Beats	166
Butterworth	36
Deictics	137
Emblems	147
Iconics	286
Metaphorics	371
total	1477

Table 1: Number of hand gestures annotated in the CID corpus.

3 Gaze

Previous studies observed that the speaker gazes away from the listener before the beginning of his speech turn (Tabensky, 1997; Bouvet & Morel 2002), and that he returns his gaze towards the listener slightly before the end of the turn (Bouvet, 1997; Tabensky, 1997; Cuxac, 2000). There are different explanations for gaze shift: not only is it a way for the speaker to show the conversational partner that he is entitled to proceed to his turn at speech, but also that avoiding eye contact facilitates lexical retrieval, and lastly that gaze direction allows the speaker to better visualize the various referents in space (Bouvet, 1995; Cuxac, 2000; Bouvet & Morel, 2002). When the speaker gazes back at the listener, he hereby acknowledges that the listener has been addressed (Cuxac, 2000), while softening the expression of a personal standpoint. It may also be interpreted by the listener as a possible turn taking, although he may decline the offer of a turn by simply adding a verbal or gestural backchannel.

Gaze mobility reveals the fluidity of the dialogue: when the speaker gazes too long at the listener, he may become menacing and that would underline the uncooperative aspect of the interaction (Tabensky, 1997). Conversely, when the speaker gazes away from the listener too long, his gaze avoidance may be interpreted as disinterest as regards the listener's feedback.

At last, the shift between mutual eye contact between participants and either gaze towards particular reference points in the gesture space gives it a deictic value (Cuxac, 2000), or towards the speaker's own hands in which case the shift reveals to the listener the strong personal involvement of the speaker (Bouvet, 1997).

Gaze refers to three separate productive organs: brows, eyelids and eyes. Poggi (2001) established a set of

formational parameters to analyze gaze, roughly including eyebrows (inner part, medial part and outer part), eyelids (upper or lower), wrinkles, and eyes (humidity, reddening, pupil dilation, eye position and eye direction). Other existing gaze coding scheme focus on the direction of the speakers' gaze as mutual gaze or non-mutual (Cerrato, 2005), eye squinting, eyebrow raising or lowering (Foster et al., 2007), coarse gaze checking whether the person is looking at the whiteboard or at another person (Carletta, 2007).

Based on these studies, we defined a gaze annotation scheme enclosing '*gaze direction* (*side)'* and '*global gaze'*. The '*gaze direction (side)'* describes eye movement orientation (*up, down, sideways, left, right, around*) and its deictic target (*interlocutor/object*). The '*global gaze'* helps to capture a global view of gaze directions and the most common gaze poses (Wallhoff et al., 2006).

4 Posture

Posture shift may be linked to interpersonal attitudes (Argyle, 1988), emotions (Sherer, 2001), communicative styles (Richmond, 1995), and personalities (Ibister & Nass, 2000). One can reduce human postures to three categories: standing, sitting (including squatting and kneeling) and lying (Argyle, 1988).

Methods of describing postures are relatively domain-related. They vary from grid-based observational studies to technical measures. Ergonomists (Corlett et al., 1986) define postures as *conditions of the body*. Due to the three-dimensional character within the three-dimensional space, they use three methods to measure and describe postures by appropriate definable geometrical parameters (three coordinates of individual joints, or adjustment levels of the long axes of separate parts of the body relative to the surrounding absolute space, or those to the axis of the preceding body part). Although these *interval-scale-oriented* parameters can cover all possible rotational movements in any joint, only the third type takes anatomic facts into consideration (e.g. limited range of movements). For practical applications, *the ordinal-scale-oriented methods* are more commonly used than those *interval-scale-oriented* methods described above. It is due to the advantage of their being independent (or semi-independent) from technical aids for posture recording while keeping accuracy or reproducibility (*e.g.* using a body diagram with limb displacement segments for recording postures). Potential displacement can be assessed by the segments within the range of movement. The *nominal-scale methods* place the description of postures according to posture typologies, which are generally profession-related. Human anatomists (White & Folkens, 1991; Platzer & Kaltes, 2004) classify body motion as flexion vs. extension (the act of bending or straightening), abduction vs. adduction (movement away from or toward the median plan), internal rotation vs. external rotation (movement around an axis), and elevation vs. depression (movement adjusting the height). The terms describe the act of performing body movements as well as the body position after being moved. Psychologists (Scherer & Ekman, 1983) distinguish posture behavior from action behavior. The posture behavior refers to overall postures (sitting, standing, lying), frontal orientation of trunk (facing, turned away), trunk lean (forward, straight, backward, sideways), arm and leg position (folded arms, uncrossed legs) and feet (flat on floor, under chair, on other knee). In the Posture Scoring system (Bull, 1987), any movement which is taken up and maintained for at least one second is annotated as a posture. This system covers head, arms, trunk and legs. Bull also proposed a second system using a dynamic approach, which describes the posture in terms of a series of movements rather than static positions.

The previous annotation scheme for the CID corpus (Bertrand et al., 2008) only considered chest movements at trunk level. Aiming at extending the postural sphere, we added a set of tracks and attributes relevant to sitting positions met in the CID corpus. It is based on the Posture Scoring System (Bull, 1987) and the Annotation Scheme for Conversational Gestures (Kipp et al., 2007). Our scheme covers four body parts: arms, shoulders, trunk and legs. With respect to arm, Bull's system mainly distinguishes whether the hand touches or not; while Kipp's scheme covers four spatial dimensions to capture it. We made a trade-off decision between the two systems: we kept the four dimensions from Kipp's coding scheme with respect to the height, distance radial orientation and swivel degree of the arm, and created a new track describing the hand touching objects to get back to the ideas of Bull's system (Table 2). Also, we added two dimensions to describe respectively the arm posture in the sagittal plane and the palm orientation of the forearm and the hand. With respect to the leg posture (Table 2), we added three dimensions: the height of the feet, the orientation of the thigh and the way in which the legs are crossed (only suited to sitting positions).

Arms	*Legs*
Height	Height
Distance	Distance
Radial Orientation	Radial Orientation
Sagittal Orientation	Sagittal Orientation
Arms Swivel	Leg-to-leg Distance
Forearm and Hand Orientation	Crossed Leg
Touch	

Table 2: Attributes to encode the spatial configurations of arms and legs.

We annotated postures on 15 minutes of the corpus involving one pair of speakers.

The proposed coding scheme leads to annotating separately the positions of the different body parts. However, we prefer to highlight the postures, which the speakers commonly take up. To this end, we proceed in two steps. First, we export data from Anvil and retrieve them into a set of significant postures according to the

following criteria:

- exclude the frames in which there is no annotation in any track;
- exclude the repetitive and successive frames.

These criteria are relevant to those of the Posture Scoring System (Bull, 1987), in which the author emphasized that if the speaker moves from the current posture without establishing a different one and then returns to the original posture, the time spent moving should be excluded from the total duration of that posture. For this reason, we exclude the frames in which there is movement in any body parts. Second, we apply a simple Hierarchic Clustering (Euclidean distance) on the extracted data. We analyzed all parts of the video data to initialize the categorization process. This method enables us to make a global-view analysis of the posture annotations and cluster different posture combinations into similar categories, which is consistent with the idea of establishing a set of posture lexicons linking body spatial locations to posture communicative functions.

Three whole body posture types have been found for one of the speakers during the first 15 minutes of the recording. In Figure 2 we selected three representative frames to illustrate the most common postures. Over 89 extracted postures, type #1 represents 44% of the postures, type #2 represents 32.5%, and type #3 represents 15.7%. Posture type #2 occurs mostly at the opening of a turn by the speaker, while posture #1 occurs at the closing sequence; posture #3 occurs mostly in the continuing sequence for the listener.

#1 #2 #3

Figure 2: Illustrations of three typical postures occurring during the first 15 minutes of conversation for one speaker.

5 Multimodal annotations

5.1 Gesture Space

The Gesture Space is a "shallow disk in front of the speaker" (McNeill, 1992: 86) where most gestures are performed. It is divided into four regions (*center-center*, *center*, *periphery* and *extreme periphery*) and eight coordinates (*no coordinate, right, left, left-right, upper right, upper left, lower right, lower left, upper, lower, upper left-right, lower left-right)*. We used McNeill's (1992: 89) diagram for the coding. The *left-right* coordinate was useful whenever a gesture was produced with both hands.

We performed a measure of inter-reliability for three independent coders for Gesture Space. The measure is based on Cohen's corrected kappa coefficient for the validation of coding schemes (Kita & Ozyurek, 2003).

Three coders have annotated three minutes for *GestureSpace* including *GestureRegion* and *GestureCoordiantes*. Annotations of these two dimensions are based on the point of maximum extension of the gesture. When annotating *GestureRegion*, the gesture that occurred in the axis of the armrests of the chair is judged in *periphery*. When it is outside, we annotate *extreme periphery*. The landmark is the position of the wrist.

Kappa (k)	*Gesture Region*	*Gesture Coordinates*
RightHand.*GestureSpace*	0,649	0,257
LeftHand.*GestureSpace*	0,674	0,592

Table 3: Average kappa values of intercoder agreement measures for each categorical item of *GestureSpace* for three coders

Strong agreement has been achieved for *GestureRegion*, which attained 64.9% agreement for the right hand; and 67.4% for the left hand. The *GestureCoordinates* indicated 25.7% agreement for the right hand, and 59.2% for the left hand. Such low agreement of *GestureCoordinates* might be due to several factors. First, the number of categorical values is high. Second, three minutes might be limited in terms of data to run a kappa measure. Third, *GestureRegion* affects *GestureCoordinates*: if the coders disagree about *GestureRegion*, they are likely to also annotate *GestureCoordinates* in a different way. For instance, it was decided that *no coordinate* would be selected for a gesture in the *center-center* region, whereas there is a coordinate value for gestures occurring in other parts of the *GestureRegion*. This means that whenever coders disagree between the *center-center* or *center* region, the annotation of the coordinates cannot be congruent.

Yet, strong agreement has been attained between one pair of coders for both *GestureRegion* and *GestureCoordinates* (Table 4).

Kappa (k)	*Gesture Region*	*Gesture Coordinates*
RightHand.*GestureSpace*	0,748	0,609
LeftHand.*GestureSpace*	0,755	0,681

Table 4: Average kappa values of intercoder agreement measures for *GestureSpace* for two coders

5.2 Posture and gesture

We also investigated how posture overlaps with significant gestures. We chose to perform an association analysis between gesture phrases and arm postures (Kipp & Martin, 2009), and made use of all the annotated data without emphasizing individual characteristics and handedness. 18 different values derived from five spatial

location dimensions, are associated with six values of *gesture phrases*. We observed that, most of the metaphoric gestures co-occurred with the following postures (only at arm level):

- place hands in the height of abdomen;
- locate arm at some distance away from the body;
- let upper arm slightly hang on the side with semi-flexed forearm;
- keep arm in front of the body.

This effect that has been observed only with metaphoric gestures may be due to its high production rate compared to other semiotic gesture types, as described in Table 1.

6 Conclusions

In this study, we proposed a coding scheme that describes the nonverbal signals produced by two sitting speakers in spontaneous conversations in French. Establishing such a multi-level nonverbal coding scheme makes it possible to understand how different nonverbal modalities co-work in multimodal information production.

We see two approaches to continue the validation work. The first is to apply agreement measures with respect to gaze direction, hand shape, orientation, trajectory and velocity, with which we can run a number of association investigations about relationships between gaze and gesture. Another approach consists in validating each individual scheme using similar annotation features. For example, both *GestureSpace* in the gesture scheme and *ArmDistance* in the posture scheme, describe the arm/hand position in the body median plan. We intend to measure agreement based on these similar annotations to validate the related coding schemes.

These annotations of nonverbal behaviors will be jointly studied with previous speech and linguistic annotations of the same data as well as future annotations planned about communicative functions.

7 Acknowledgements

The work described in this paper is supported by the French Agence Nationale de la Rechcere (ANR) under the project OTIM (ANR BLAN08-2_349062). The authors would like to thank Roxane BERTRAND from the University of Provence, and Michael KIPP for the Anvil tool.

8 References

Allwood, J., Cerrato, L., Dybkjaer, L., Jokinen, K., Navarretta, C., Paggio, P. (2005) The MUMIN Multimodal Coding Scheme, NorFA yearbook 2005. http://www.ling.gu.se/~jens/publications/B%20files/B70.pdf

Argyle, M. (1975). *Bodily Communication*. London: International Universities Press.

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., Rauzy, S. (2008). Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. *Traitement Automatique des Langues*, vol. 49, no. 3. p. 105-134.

Blache, P., Bertrand, R., Ferré, G. (2009). Creating and Exploiting Multimodal Annotated Corpora: The ToMA Project. In Kipp M. (eds.) *Multimodal Corpora*. Berlin: Springer-Verlag. 2009, vol.LNAI 5509, p. 38-53.

Bouvet, D. (1997). Le corps et la métaphore dans les langues gestuelles. *A la recherche des modes de production des signes*, Paris, L'Harmattan, coll. « sémantiques », p. 120.

Bouvet, D., Morel, M.-A. (2002). Le ballet et la musique de la parole. *Geste et intonation dans le geste et l'intonation dans le dialogue oral en français*. Paris-Gap: Ophrys.

Bull, P. (1987). *Posture and Gesture*. Oxford: Pergamon Press.

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41(2), 181-190.

Corlett, E. N., Wilson,John R. Manenica. I. (1986). Chapter 18: Influence Parameters and Assessment Methods for Evaluating Body Postures. In Ergonomics of Working Postures: Models, Methods and Cases: The *Proceedings of the First International Occupational Ergonomics Symposium*, Zadar, Yugoslavia, 15-17 April 1985: CRC Press.

Cuxac, Ch. (2000). La langue des signes française, *Faits de Langues*, Paris-Gap : Ophrys, p. 14-15.

Foster, M., Oberlander, J. (2007). Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation*, 41(3), 305-323.

Isbister, K., Nass, C. (2000). Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies,* 53(2), 251-267.

Loredana, C. (2005). On the acoustic, prosodic and gestural characteristics of m-like sounds in Swedish. Göteborg, SUEDE: University of Göthenburg, Department of Linguistics.

Kendon, A. (1980). Gesticulation and Speech : Two Aspects of the Porcess of Utterance. In M.R. Key (ed.), *The Relationship of Verbal and Nonverbal Communication*, The Hague: Mouton, p. 207-227.

Kipp, M. (2004). Gesture Generation by Imitation - From Human Behavior to Computer Character Animation. Boca Raton, Florida, Dissertation.com.

Kipp, M., Neff, M., Albrecht, I. (2007). An annotation scheme for conversational gestures: how to economically capture timing and form. *Language Resources and Evaluation,* 41 (3):325-339.

Kipp, M., Martin, J.-C. (2009) Gesture and Emotion: Can basic gestural form features discriminate emotions? In: *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII-09)*, IEEE Press.

Kita, S., Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation

of spatial thinking and speaking. *Journal of Memory and Language,* 48: 16-32.

McNeill, D. (1992). Hand and Mind. *What Gestures Reveal about Thought*, Chicago: The University of Chicago Press.

McNeill, D. (2005). *Gesture and Thought*, Chicago, London : The University of Chicago Press.

Platzer, W., Kahle W. (2004*). Color Atlas and Textbook of Human Anatomy*, Thieme.

Poggi, I. (2001). The Lexicon and the Alphabet of Gesture, Gaze, and Touch *Intelligent Virtual Agents* (pp. 235-236).

Richmond, V. P., McCroskey J. C., Hickson M. L. (1995). *Nonverbal Behavior in Interpersonal Relations*. Pearson Ed: Allyn & Bacon.

Scherer, K.R., Ekman, P. (1982). *Handbook of methods in nonverbal behavior research*. Cambridge University Press.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theroy, Methods, Research*. A. S. K. R. Sherer, & T. Johnstone. New York and Oxford, Oxford University Press, p. 92-120.

Tabensky, A. (1997). *Spontanéité et interaction. Le jeu de rôle dans l'enseignement des langues étrangères*, Paris : L'Harmattan.

Wallhoff, F. M., Ablaßmeier, M. and Rigoll, G. (2006) Multimodal Face Detection, Head Orientation and Eye Gaze Tracking. In: *Proceeding of IEEE International Conference on Multisensor Fusion and Integration (MFI),* Heidelberg.

White, T. D., Folkens, P. A. (1991) *Human Osteology*. San Diego: Academic Press, Inc.

A multimodal corpus for gesture expressivity analysis

G. Caridakis[1], J. Wagner[2], A. Raouzaiou[1], Z. Curto[3], E. Andre[2], K. Karpouzis[1]

[1]Image, Video and Multimedia Systems Laboratory
National Technical University of Athens
Iroon Polytexneiou 9, 15780 Zografou, Greece
[2]Multimedia Concepts and their Applications Laboratory
Augsburg University
Universitatsstr. 6a, 86159 Augsburg
[3]Humanware S.r.l.
via Garofani 1,Pisa - 56125 Italy
{gcari,araouz,kkarpou}@image.ntua.gr, {johannes.wagner,andre}@informatik.uni-augsburg.de, z.curto@hmw.it

Abstract

This work presents the design and implementation of corpus recording sessions along with some preliminary processing results. Captured modalities include speech and facial expressions but the focus is on hand gesture expressivity. Thus, this is the primary modality and is recorded using three methods: bare hands, Nintendo Wii remote controls and datagloves. Such a setup allows for multimodal affective analysis and potentially provide quantitative parameters for synthesis of systems affectively aware and able to convey affect, such as Embodied Conversational Agents. Additionally, comparative studies of gesture expressivity based on different recording techniques could be based on the introduced corpus. Cross cultural affective behavior issues are also incorporated since the experiment was performed in three countries i.e. Greece, Germany and Italy.

1. Introduction

An abundance of research within the fields of psychology and cognitive science related to the non verbal behavior and communication stress out the importance of qualitative expressive characteristics of body motion, posture, gestures and in general human action during an interaction session (Trenholm and Jensen, 2007). Although such research work studies primarily and mainly context of human to human interaction such approach can be extended to human computer interaction. Some work has incorporated gesture expressivity in HCI context but the vast majority concentrates on the expressively enhanced synthesis of gestures by virtual agents and ECAs (Pelachaud, 2009). Currently, research on the automatic analysis of gesture expressivity is still immature and this fold of human action analysis is asymmetrically studied with reference to the synthesis counterpart.

Recent psychology studies suggest that body language does constitute a significant source of emotional information. Nevertheless, it is hard to identify specific characteristics of body language that could help us assess a user's emotional state. First of all, there is no clear mapping from gestures to emotional states. Secondly, the use of gestures differs from person to person and from situation to situation. How people express bodily emotions depends on a variety of factors including the social context, people's personality and their cultural background. Corpus studies of human behavior may provide useful insights in human behavioral patterns. In the social sciences, corpus studies have a long tradition as a basis to analyze human behavior. In computer science, corpora of human behavior have been employed both for generation and analysis tasks. First, a number of attempts have been made to extract human behavioral patterns from corpora to guide the design of virtual agents. Secondly, corpora have played an important role in recognition tasks where classifiers have been trained based on labeled data of human behavior. The acquisition of corpora enables us to ground research on gesture generation and recognition in empirical data. Unfortunately, corpora with annotated communicative gestures expressing emotional content are rare. In particular, there are no corpora available that enable us to investigate culture-specific aspects of gestural emotions. As a consequence, we decided to collect our own corpus.

The objective of our work is threefold: First of all, we aim at studying the use of emotional gestures in combination with other modalities, such as facial expressions and speech. As a consequence, data need to be collected in a synchronized manner. Secondly, we investigate how bodily emotions are expressed in different cultures. As a first step, we focus on three European countries: Greece, Italy and Germany. Thirdly, we are interested in finding out how the use of interaction devices influences people's gestures and the robustness of the recognition process. In particular, we focus on: video-based gesture recognition, Wiimote-based gesture recognition and gesture recognition based on a data glove. All these three technologies come with their own advantages and disadvantages. Computer vision is the less obtrusive means to capture information about the user's body movements. However, it is rather sensitive against lighting conditions. Gesture recognition using a data glove or the Wiimote does not suffer from these problems. However, it is much more obtrusive. In particular, using a Wiimote for performing affective gestures is rather unnatural since users have to carry a device in their hands which might influence their way of gesturing.

2. Related work

Designing, recording and labeling human affective expressions is a prerequisite in designing affective aware systems. Many aspects are included in the above mentioned processes involved in creating a affective corpus. Behavior spontaneity, recorded modalities, labeling are merely a few of the aspects that have to be taken under consideration when creating multimodal, affectively enriched corpora aiming to be used for affective analysis. Naturalistic behavior are considered ideal for validating real life affective analysis systems, although such behavior is relatively rare,short lived,and filled with subtle context-based changes. On the other hand there is a large number of issues and internal processes that influence the final result involved in the affective elicitation methods. Additionally, the selection of the recorded modalities incorporates intrusion issues and/or content based, modality related issues. Finally, the adopted emotion representation, annotation and labeling scheme should be predefined since these decisions are extremely important to both automatic affect recognition and user perception tests.

Besides the implications reported above the necessity for creating reusable databases consisting affectively enriched human behavior has resulted in a number of attempts for creating multimodal corpora. The importance of each corpus is determined both by the effort and reasoning for each decision involved in the database creation as well as the research work performed from the automatic analysis view using the specific corpus.

The Belfast database (Douglas-Cowie et al., 2003) was constructed by the Queen's University of Belfast and mainly consists of sedentary interactions, from chat shows, religious programs and discussions between old acquaintances. It consists of 125 English speaking subjects experiencing a wide range of positive and negative emotions and of emotional intensities. The FeelTrace (Cowie et al., 2000) tool was used for labeling the corpus recording the perceived emotional state via dimensional rating. The EmoTV corpus (Abrilian et al., 2005) is another corpus, which is in French and also draws material from TV interviews, but uses episodes with a wider range of body postures and more monologue, such as interviews on the street with people in the news. EmoTV uses ANVIL (Kipp, 2001) as a platform and the coding scheme uses both verbal categorical labels and dimensional labels (intensity, activation, self-control and valence). A corpus construction attempt (Kessous et al., 2009) was also performed within the HUMAINE EU-IST project framework during its Third Summer School held in Genova in 2006. While the previous corpora consisted of real life interviews, the Genoa corpus included acted human behavior induced using a process similar to the one adopted in the GEMEP corpus (Bänziger et al., 2006). Ten participants participated in the recordings representing 5 nationalities, incorporating cross cultural issues, and data on facial expressions, body movement, gestures and speech were simultaneously recorded. A pseudo-linguistic sentence was pronounced by the participants while acting through eight emotional states uniformly distributed in valence-arousal space (two emotional states per quadrant). The GEMEP (Geneva Multimodal Emotion Portrayals) corpus (Bänziger et al., 2006) constitutes a repository of portrayed emotional expressions. The researchers argue that portrayals produced by appropriately instructed actors are analogue to expressions that do occur in selected real life contexts as opposed to induced or real-life sampled emotional expressions that display expressive variability and therefore constitute excellent material for the systematic study of nonverbal communication of emotions. Ten professional French-speaking actors portrayed 15 affective states under the direction of a professional stage director, recording audio, facial expressions and head orientations, body postures and gestures from two viewpoints (perspective of an interlocutor and sideways).

This corpus construction innovation lies in its focus on gesture expressivity, the inclusion of multiple cultures and multiple human behavior capturing techniques, i.e. video, Wiimote and Datagloves. The introduced multicultural corpus allows for intercultural affective analysis, while the variety of technologies used to record human body behavior supports studies on their obtrusiveness effect. So, the motivation for this experiment is threefold. German, Greeks and Italians, while speaking, use their hands in a different way. The described experiment is providing us with the means to compare the expressiveness not only between the different cultures, but also between the different capturing techniques and the different emotional characterizations. Furthermore, the data is synchronized, so analyzing the affective behavior of the user allows us to extract conclusions for the correlation of gesture expressivity with acoustic prosody and facial expressions.

3. Corpus construction

3.1. Affective immersion

The adopted emotion elicitation method was inspired by the Velten mood induction technique (Velten, 1998) where people had to read aloud a number of sentences that put them in particular emotional state. First, we displayed a sentence with a clear emotional message and gave the user sufficient time to read it silently. Then the projection turned blank and the user was asked to express the according emotion through gesture and speech. The users were encouraged to use their own words as long as they helped them feel a particular emotion. The sentences were shown in three coherent blocks with first positive, then neutral and finally negative sentences in order to put the users gradually into the desired mood. We selected in total 120 sentences (40 for each target class) such as:

Table 1: Example of the used Velten sentences per emotion category.

The hike was fantastic! You won't believe it! But we made it to the top!	positive
The names on the mailing list are alphabetically ordered.	neutral
Sometimes I wonder whether my effort is all that worthwhile.	negative

We decided to choose the order positive-neutral-negative in order not have to switch directly between the two emotional

extremes. Furthermore, users usually feel less motivated towards the end of the experiment and it would be harder to put them into a positive emotional state.
Each of the three blocks is again divided into three sections, during which we equip the user with different interaction tools. During the first 20 sentences subjects are wearing a data glove by HumanWare. The next 10 sentences the glove is exchanged by two Wii remote controls, which the users hold in their hands. Finally, the remaining 10 sentences were performed with free hands.

3.2. Hardware setup

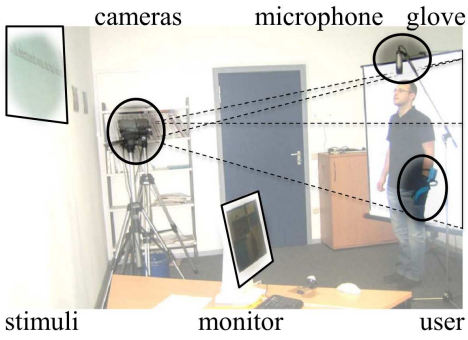


Figure 1: Picture of the experimental setting in which the capture devices are highlighted. The monitor in the front is used by the experimenter to overview the recording and run the stimuli script.

During the recordings the user stands in front of a neutral background. The stimuli, i. e. the Velten sentences, is projected on a screen in front of him. The projection is adjusted in a way that the user can read the displayed text without the need to turn his head. Below the projection, in a distance of two meters and approximately at the height of the user's face, two high-quality cameras (720x576 pixels, 25 fps, 24 bit colour depth) are placed. The first camera is setup to capture the user's complete body including arm gestures, while the second camera aims at the user's face and captures a close-up of shoulder and head. In addition the whole scene is captured at a lower resolution with a webcam, primarily for annotation and monitoring purpose. Audio is recorded with an USB microphone (Samson C01U, 16kHz, mono, 16 bit). To avoid occlusions in the videos a stand is used to locate the microphone on top of the user's head. Each recording is divided in three parts characterized by different interaction modes. During the first mode the user is wearing a data-glove on one hand. The data-glove is provided by HumanWare and is used primarily to record finger movements during the experiment, to verify whether (and how much) users gesticulate with their hands and fingers. The dataglove records 26 signals at a sampling rate of 50Hz: 15 signals for flexions of all fingers on one hand, 2 signals for flexion and ad/abduction of the wrist and since it embeds an IMU (inertial measurement unit) in the forearm it also records a 3-axial magnetic field, 3-axial acceleration and 3 angular velocities. The data transfer is done over a wireless Bluetooth connection. During the second interaction mode the user holds Nintendo's Wii remote control in each hand, which measures 3D acceleration. The last interaction mode is freehand.

3.3. Recording Software

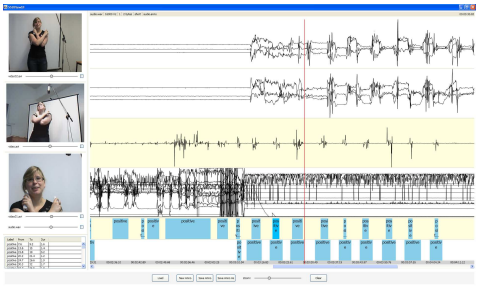


Figure 2: Reviewing the recordings using the SSI viewer tool features multiple annotation layers and simultaneous playback of modalities.

During the analysis of the corpus it is our purpose to not only look at modalities individually, but also investigate the relations between the different channels and explore ways to fuse the different kind of information. This, however, requires a proper synchronization between the modalities. To obtain synchronized recordings we use Smart Sensor Integration (SSI), a software framework for multi-modal signal processing in real-time, developed at the University of Augsburg (Wagner et al., 2009). In SSI synchronization is achieved by constantly updating the incoming signal streams to a global clock. As SSI already supports common sensor devices including webcam/camcorder, microphone and Wii remote control, integration of the modalities is straight-forward. For the data glove, which is not supported by default, we fall back to the possibility to capture a signal from a socket stream. The accomplishment of the experiment is supported by a graphical user interface, which allows the experimenter to display a sequence of HTML documents containing the according stimuli sentences.
Recordings from all devices are synchronized throughout the whole session and directly stored to disk. To avoid artefacts no compression is used on the high-quality video streams, which leads to a data transfer of $\approx$ 30 MB/s per video. For this reason the two videos are recorded on a separate pc, which is synchronized by a broadcast message sent from the main machine. This way it becomes possible to distribute recordings on any number of machines. During a recording SSI is used to detect parts of the signals with high activity, e.g. when a user is talking or performing a gesture. Based on these events preliminary annotations are generated for each signal. Signals and annotations can be reviewed and adjusted using a graphical tool shown in Figure 2. This way, SSI not only supports the recording, but also the analysis of the corpus.

3.4. Procedure

Apart from the setup of the equipment, participants were necessary for our experiment. Each partner was responsible to recruit local subjects and run the experiment at his lab. An according translation of the same set of Velten sentences

was used. In Greece, 11 subjects (6 male and 5 female) between 23 and 40 years old took part in the experiment, while in Germany 21 subjects (11 male, 10 female) were following our scenario. Their age was varying between varied between 20 and 28 years old, while in Italy 19 (11 males and 8 females) took part in the experiment, between 24 and 48 years old well distributed. In case of the German experiment, subjects were given an allowance of 20 €. So far, 15h:14m:24s of interaction has been recorded in the three countries (which includes parts when users were changing devices).

Subjects training Before the experiment we recorded a video with the whole procedure. A trained person was executing the gestures with Wii-mote, glove or bare hands. The possible participants were offered the opportunity to watch the video and/or read the Velten sentences and to pose any questions they want regarding the experiment. Once they agreed to participate, they were given a consent form to sign. The issues covered in this form are described in Section 3.4.. During the experiment we presented to every participant what she should do and she could re-watch the video or be present while another person was following the procedure. We were at her disposal to settle any queries and two persons were constantly present during the experiment to guarantee its success and to help the participants using the Wiimote, wearing the glove etc.

Ethical issues As already mentioned, the participants should sign a consent form which ensures that they are informed about the scope of the experiment, their involvement and that they can assess the risks that might occur from the processing of data. The data is stored, so they should have in mind that, although the samples collected are anonymous, the voice or the face of a subject might be recognized. The consent form gives them the right to ask for erasure or blocking of the data that concerns her/him and to withdraw from the experiment at any time.

4. Corpus preliminary analysis

4.1. Speech

To analyse the expressivity in speech we use EmoVoice [1], a tool for real-time recognition of emotions from acoustic properties of speech (Vogt et al., 2008). The acoustic features used by EmoVoice are mainly based on short-term acoustic observations, including pitch, signal energy, Mel-frequency cepstral coefficients, spectral and voicing information, and the harmonics-to-noise ratio. Overall, a set of 1316 features is obtained from each speech segment[2]. For evaluation purpose a Naïve Bayes classifier is trained and tested using on two-fold cross-validation.

So far, the German and Italian sentences have been proceeded and some of the results are listed in Table 2. The quoted recognition rates were obtained by taking the sentences of all subjects: first separately for each country and then combined. In addition, the last two columns list the worst and best performance of a speaker-dependent classification. In both cases, a difference of more than 20% proves the high variability between users. Compared to that the drop from 53.83% (German) and 51.16% (Italian), respectively, to 48.91% (combined) appears rather small. This, for instance, suggests that differences between individuals are actually more relevant than differences between the two cultures.

nation	positive	neutral	negative	average	min	max
GER	36.36%	67.11%	58.01%	53.83%	54.17%	75.83%
IT	36.39%	60.83%	56.25%	51.16%	49.17%	74.17%
BOTH	31.81%	53.60%	61.31%	48.91%	-	-

Table 2: EmoVoice classification results

4.2. Gestures

4.2.1. Gesture expressivity features

Behavior expressiveness is an integral part of the communication process since it can provide information on the current emotional state and the personality of the interlocutor (Mehrabian, 2007). Many researchers have studied characteristics of human movement and coded them in binary categories such as slow/fast, restricted/wide, weak/strong, small/big, unpleasant/pleasant in order to properly model expressivity. We adapted six expressivity dimensions described in (Hartmann et al., 2005), as the most complete approach to expressivity modeling, since it covers the entire spectrum of expressivity parameters related to emotion and affect. Derived from the field of expressivity synthesis five parameters have been defined, consisting a subset of the six expressivity dimensions: Overall activation, Spatial extent, Temporal, Fluidity and Power.

Overall activation is considered as the quantity of movement during a dialogic discourse and, given the above definitions, is formally defined as the sum instantaneous quantities of motion of the two hands. Spatial extent is expressed with the expansion or the condensation of the used space in front of the user (gesturing space). In order to provide a strict definition of this expressivity feature spatial extent is considered as the maximum value of the instantaneous spatial extent during a gesture. The temporal expressivity parameter denotes the speed of hand movement during a gesture and dissociates fast from slow gestures. On the other hand, the energy expressivity parameter refers to the movement of the hands at during the stroke phase of the gesture. This parameter is associated qualitatively with the acceleration of hands during a gesture. Fluidity differentiates smooth/elegant from the sudden/abrupt gestures. This concept attempts to denote the continuity between hand movements and is suitable for modeling modifications in the acceleration of the upper limbs.

4.2.2. Bare hands

Regarding the hand and head detection and tracking problem which is a required step for extracting expressivity features from a gesture several approaches have been reviewed. Amongst them only video based methods were considered since motion capture or other intrusive techniques would interfere with the person's emotional state which is a crucial issue in this kind of analysis. The major factors taken under consideration are computational cost and robustness, resulting in an accurate near real-time skin

[1] https://mm-werkstatt.informatik.uni-augsburg.de/EmoVoice.html

[2] Here: the utterance of one Velten sentence

detection and tracking module. The overall process is described in detail in (Caridakis et al., 2007) and includes creation of moving skin masks and tracking the centroid of these skin masks among the subsequent frames of the video depicting a gesture. The described algorithm is lightweight, allowing real time processing and indicative results and intermediate steps can be seen at 3.

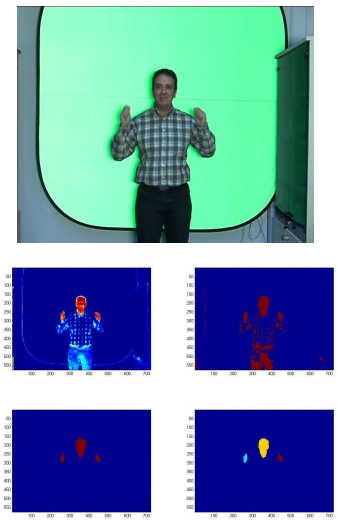


Figure 3: Image processing intermediate steps and results

Extraction of the expressivity features based on the coordinates of the hands, as detected using the image processing described above, is actually a process based on calculations over the motion vectors resulting from the coordinates. Overall activation is computed as the sum of the motion vectors' norm, Spatial extent is modeled by the maximum Euclidean distance of the position of the two hands and Fluidity is the variance of the norms of the motion vectors. Power is actually identical to the first derivative of the motion vectors norms, calculated previously.

4.2.3. Glove

Regarding finger movement analysis, our approach is based on the extraction of some features (or expressivity dimensions) very similar to the ones seen before. In particular we have extracted: Overall activation, Power, Spatial extent and Fluidity.

The dataglove measures the angles of finger joints. The angles are then normalized according to a multi-user transformation function, accessible through a calibration procedure. We can define the finger's motion energy as a sum of the fingers joint angular velocities and use this value as an overall activation feature. The power of one gesture is approximated by the sum of the root mean square of all the finger joint angular accelerations while a measure of fluidity of the gesture can be determined using motion jerk, i.e. the derivative of acceleration. Finally, the spatial extent of the movement is given by the maximum distance a finger joint has moved during the sentence.

5. Conclusions

The work presented here discusses issues related to the design and implementation of an experiment, resulting in a multimodal corpus of affective behavior, incorporating acoustic prosody, facial expressions and gesture expressivity, and briefly introduces the methods that will be used in the future to process the resultant corpus. During corpus affective analysis, that is considered ongoing and future work, significant conclusions are expected to be drawn, especially for the analysis of gesture expressivity, its correlation with other modalities and related cross cultural and interaction obtrusiveness issues. The ambition of this research work is that the constructed multimodal corpus, once synchronized and formatted, will be established as a benchmark multimodal corpora standard focused on gesture expressivity. Feature extraction, multimodal analysis and synchronization and fusion techniques from the involved research teams will be applied to the corpus and, hopefully, will provide reference point for future attempts within the affective computing community.

6. Acknowledgements

This work has been funded by the FP6 IP Callas (Conveying Affectiveness in Leading edge Living Adaptive Systems), Contract Number IST-34800.

7. References

S. Abrilian, L. Devillers, S. Buisine, and J.C. Martin. 2005. Emotv1: Annotation of real-life emotions for the specification of multimodal affective interfaces. In *HCI International*.

T. Bänziger, H. Pirker, and K.R. Scherer. 2006. GEMEP–GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions. In *The Workshop Programme Corpora for Research on Emotion and Affect Tuesday 23 rd May 2006*, page 15. Citeseer.

G. Caridakis, A. Raouzaiou, E. Bevacqua, M. Mancini, K. Karpouzis, L. Malatesta, and C. Pelachaud. 2007. Virtual agent multimodal mimicry of humans. *Language Resources and Evaluation 41 (3-4), Special issue on Multimodal Corpora, pp. 367-388, Springer.*

R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder. 2000. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Citeseer.

E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. 2003. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1-2):33–60.

B. Hartmann, M. Mancini, S. Buisine, and C. Pelachaud. 2005. Design and evaluation of expressive gesture synthesis for embodied conversational agents. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, page 1096. ACM.

L. Kessous, G. Castellano, and G. Caridakis. 2009. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces, Springer, DOI 10.1007/s12193-009-0025-5*.

M. Kipp. 2001. Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*. ISCA.

A. Mehrabian. 2007. *Nonverbal communication*. Aldine.

C. Pelachaud. 2009. Studies on gesture expressivity for a virtual agent. *Speech Communication*, 51(7):630–639.

S. Trenholm and A. Jensen. 2007. *Interpersonal communication*. Oxford University Press, USA.

E. Velten. 1998. A laboratory task for induction of mood states. *Behaviour Research and Therapy*, 35:72–82.

T. Vogt, E. André, and N. Bee. 2008. Emovoice - a framework for online recognition of emotions from voice. In *Proceedings of Workshop on Perception and Interactive Technologies for Speech-Based Systems*, Kloster Irsee, Germany, June. Springer.

J. Wagner, E. André, and F. Jung. 2009. Smart sensor integration: A framework for multimodal emotion recognition in real-time. In *Affective Computing and Intelligent Interaction (ACII 2009)*.

Timing Relationships between Speech and Co-Verbal Gestures in Spontaneous French

Gaëlle Ferré
Laboratoire de Linguistique de Nantes (LLING)
Université de Nantes
Chemin de la Censive du Tertre, BP 81227
44312 Nantes cedex 3 -- FRANCE
E-mail: Gaelle.Ferre@univ-nantes.fr

Abstract

Several studies have described the links between gesture and speech in terms of timing, most of them concentrating on the production of hand gestures during speech or during pauses (Beattie & Aboudan, 1994; Nobe, 2000). Other studies have focused on the anticipation, synchronization or delay of gestures regarding their co-occurrence with speech (Schegloff, 1984; McNeill, 1992, 2005; Kipp, 2003; Loehr, 2004; Chui, 2005; Kida & Faraco, 2008; Leonard and Cummins, 2009) and we would like to participate in the debate in the present paper. We studied the timing relationships between iconic gestures and their lexical affiliates (Kipp, Neff et al., 2001) in a corpus of French conversational speech involving 6 speakers and annotated both in Praat (Boersma & Weenink, 2009) and Anvil (Kipp, 2001).
The timing relationships we observed concerned the position of the gesture stroke as compared to that of the lexical affiliate and the Intonation Phrase, as well as the position of the gesture Phrase as regards that of the Intonation Phrase. The main results show that although gesture and speech are co-occurring, gestures generally start before the related speech segment.

1. Introduction

These last years, a major effort has been made by the international community to extend the number, variety and size of annotated multimodal corpora in several languages, especially since it has been shown by McNeill (1992) among others that gestures play a role in communication. Nowadays, some tools even allow automatic recognition of body movements (Campbell, 2009) which will save time in the annotation of the less interpretative gesture configurations (eyebrow rise, hand trajectory, for instance). Yet, manual annotation is still needed for features that involve interpretation (type of hand gesture, etc). Manual annotation is also needed for corpora recorded before the development of special tools.
This is why the OTIM project (Bertrand et al., 2008; Blache et al., 2009) is based on the annotation of several hours of conversational speech in French. Part of the annotation process is automatic (transcription and alignment of words and phonemes, annotation of syntactic clauses and morphological categories), but the rest is manual (gesture and body movements and postures, prosodic phenomena, discourse units). These annotations (whether automatic or manual) are made with the annotation tool Praat (Boersma & Weenink, 2009) for speech and Anvil (Kipp, 2001) for gestures, which is not the case of every study concerning gesture-speech relationships (for instance, the studies of Chui, 2005, and Kida & Faraco, 2008, were not based on alignment of speech transcription and gesture annotation). This does not mean that linguistic studies which are not based on time-aligned annotations are of no value, but simply that temporal alignment of annotations adds precision to otherwise more intuitive observations.
Among the studies concerned with co-verbal gestures, a few of them described the timing relationships between gesture and speech. Beattie & Aboudan (1994) and Nobe (2000), for instance, analyzed gesture production co-occurring with speech or with silent pauses. Others like Schegloff (1984), McNeill (2001, 2005), Kipp (2003), Loehr (2004), Kranstedt et al. (2006) and Leonard & Cummins (2009) for English, Chui (2005) for Chinese, Kida & Faraco (2008) for French, and Rochet-Capellan (2008) for French and Portuguese, concentrated on the timing of gesture with regards accompanying speech or parts of speech. These studies all show the interest of developing annotated corpora and timing relationships will also be the object of the present paper, in which we will compare the timing of the gesture stroke with regard to the lexical affiliate, and the gesture phrase with regard to the Intonation Phrase, after having briefly presented the corpus and the annotations made.

2. Corpus and data

This study is based on analysis of the data in a subset of the CID video corpus, fully described in Bertrand et al. (2008) and Blache et al. (2009). This corpus is still under the annotation process (OTIM project), but we were able to work on the hand gestures annotated in 75 minutes of speech, involving 6 speakers in dialogues of spontaneous French.

2.1 Speech transcription and annotations used to establish timing relationships

The paper is based on a semi-automatic transcription and its alignment with the sound file in Praat, which were then corrected manually. Intonation Phrases (IPs) as defined by Selkirk (1978 and later works) have also been

annotated in Praat: whereas syntactic units such as clauses or sentences would be relevant for written texts, we considered that Intonation Phrases are quite appropriate for the chunking of speech recordings since prosody (including pauses, different degrees of stress and boundaries) gives some clue on information structure. If we consider the following example from the corpus:

/ y avait un espèce d'écran géant donc un matériel d'enfer /

The utterance could have two possible interpretations due to the structure of spoken French and the possible placement of the conjunction *donc* ("so") which can be placed before, in the middle or after the clause in its syntactic domain: (a) so there was some sort of huge screen, high tech resources, or (b) there was some sort of huge screen so they had high tech resources. Now if we consider prosody, the ambiguity is not present anymore and the two Intonation Phrases are in fact:

/ y avait un espèce d'écran géant / donc un matériel d'enfer /

This is not determined by the presence of a pause as there is none in the example but rather by the fact that there is a pitch rise on "géant" and a reset on "donc". If "donc had been part of the first Intonation Phrase, it would still have been low in pitch but the pitch reset would have occurred on "un", so that there would have been a prosodic break between "donc" and "un".

Other studies have previously established a relationship between prosodic units and gestures. For instance, Loehr (2004) has shown that there is a timing relationship between Intermediate Phrases and gesture phrases (described in the next section). His study is in the framework of J. Pierrehumbert's autosegmental theory of intonation and what he terms Intermediate Phrases corresponds to Intonation Phrases in Selkirk's metrical theory so the prosodic units we are considering are the same. Below is represented the metrical analysis of the Intonation Phrase "tu signes le papier" ("you sign the paper", the example is also given in section 2.3).

(						x)	Intonation Phrase
(	x	)	(		x)	Accentual Phrases
(	x	)	(		x)	ω
	(x	)		(	x)	Σ
x	x	x	x	x	x	σ
tu	signes		le	papier		

In this study, the distinction between Major and Minor accentual phrases (Kratzer and Selkirk, 2007) was not relevant since only Intonation Phrases were annotated, but it is important to understand exactly what elements they comprise. The stresses in Accentual Phrases correspond to pitch accents in the autosegmental theory, whereas stresses at Intonation Phrase level correspond to phrase tones in the autosegmental theory. Selkirk analyses a further level, the sentence level, which is not relevant here. Stresses at this level would correspond to edge tones.

For all these reasons, IPs seemed to be quite an appropriate unit in a comparison between speech and gesture, and their timing has been directly compared to the timing of gesture phrases as described in the next paragraph, whereas gesture strokes have been linked to lexical affiliates. All speech transcriptions and annotations made in Praat were then imported in Anvil which was used for the annotation of gestural phenomena.

2.2 Gesture annotations

Although the general OTIM project has started the annotation of various gestures, movements and postures that include all body parts, the author of the present paper has been more particularly concerned with the annotation of hand gestures. So far, 1477 gestures have been annotated on 75 minutes of speech (the ultimate aim being to annotate all the gestures during 3 hours of corpus). Each hand gesture was described in terms of symmetry (single-handed vs. two-handed gesture, symmetric vs. asymmetric hand configuration). We then annotated each gesture's phases (Kendon, 1980): preparation – stroke – hold – retraction – recoil, as well as the gesture's phrase (Kendon, op. cit.), that is the gesture in its whole, to which we assigned a dimension (McNeill, 2005) or a function regarding co-occurring speech (we retained the semiotic types proposed by Kipp, 2003). Each gesture was also described in terms of hand shape, gesture trajectory, space, velocity and amplitude, although these descriptions which were useful in determining lexical affiliates, were not used *per se* in the present study. The gesture onset corresponds to the first frame in which the hand(s) moves from its rest position whereas the offset corresponds to the first frame in which the hand returns to its rest position when the gesture is produced in isolation. When the gesture is produced in between two other gestures without any return to rest position, its onset corresponds to the first frame in which the hand changes trajectory from the previous gesture (initiates the preparation or stroke of the gesture). Its offset corresponds to the last frame before the hand changes trajectory for the preparation or stroke of the next gesture. One has to keep in mind that due to the granularity of the videos (24 frames per second), the onset and offset of hand gestures are defined less precisely than the onset and offset of speech shown in Figure 1 below.

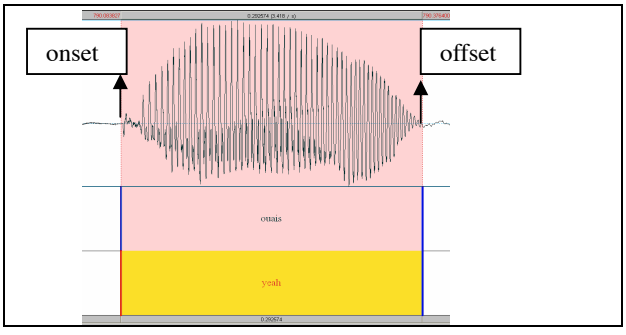


Figure 1: Speech onset and offset.

Among the annotations, we retained only the iconic hand gestures for the study, which was based on a number of 244 gestures out of a total of 286 iconics (42 were discarded, either because it was not possible to determine a lexical affiliate in speech due to the absence of a verbal affiliate or due to the fact that it was not possible to determine precisely a word in speech which would have a close meaning to the one conveyed by the gesture; some of the gestures were also interrupted and not taken into consideration.

2.3 Lexical affiliate

In order to establish a relationship between gesture and speech in terms of timing, it is necessary that the link between them be of an explicit nature. Schegloff (1984) described lexical affiliates as "the word or words deemed to correspond most closely to a gesture in meaning." In the case of iconics, it appears that in 85.3 % of the occurrences, it is possible to determine a lexical affiliate that actually corresponds to a word in the co-occurring speech, which is quite a high rate. This close correspondence between lexical affiliates and what Kipp calls 'redundant iconics' (2003:153) is shown in Figures 2 and 3.

tu signes le papier

you sign the paper

Figure 2: Iconic gesture corresponding to the lexical affiliate "sign" in terms of hand configuration and movement.

Figure 3: Iconic gesture and Anvil annotation corresponding to the lexical affiliate "long".

This close correspondence between gesture feature and meaning in speech is not so explicit with other gesture dimensions such as metaphorics, for instance, which may be used to add a modality to the entire IP as presented in Figure 4:

on n'en avait pas reparlé

we hadn't spoken about it again

Figure 4: Metaphoric gesture adding a modality to the IP with no precise lexical affiliate.

The explicit affiliation is the reason why we chose to study the timing relationships between iconics and co-occurring speech which was also the choice made by Chui (2005), whereas Schegloff (1984) for similar reasons, chose deictics (the number of deictics in our corpus was too small with only 137 occurrences to motivate such a choice, this depending much on the type of corpus), and Leonard & Cummins (2009) chose beats. Other studies had a larger understanding of lexical affiliation and did not restrict their observations to a particular dimension (Loehr, 2004).

3. Results

The first observation that needs to be made concerning the timing of gestures and co-occurring speech is that gesture units (Gstroke at lexical level and Gphrase at phrase level) are generally longer than corresponding verbal units (word and IP), as shown in Figure 5 below, and that the difference between the mean duration of phrasal units is smaller than the difference of the mean duration at word level.

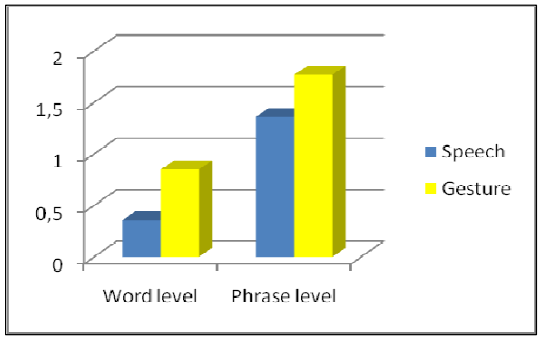


Figure 5: Mean duration (in seconds) of lexical units (affiliate/Gstroke) and phrasal ones (IP/Gphrase).

In terms of timing relationships (*cf.* percentages given in Table 1 below), at word level, when comparing the onset and offset of gesture stroke and lexical affiliate, we observe that a large majority of strokes (72 %) start before the onset of the lexical affiliates and an even greater proportion of strokes (87 %) end after the offset of lexical affiliates. A paired T-Test showed that the anticipation of strokes on lexical affiliates is highly significant: the mean difference (M=-0.454, SD=0.692, N=244) is significantly greater than zero (t(243)=-10.2, two-tail p=1.07e-20). A 95 % C.I. about stroke/affiliate onset is (-0.54, -0.36). However, although a higher proportion of gesture strokes ended after the offset of the

lexical affiliate, the mean difference (M=0.03, SD=0.76, N=244) is not significant (t(243)=0.80, two-tail p=0.42). A 95 % C.I. about stroke/affiliate offset is (-0.05, 0.13). These statistics show that the difference in timing of the onset and offset of gesture and speech are not only due to the fact that gesture strokes are generally longer than lexical affiliates but also that there is a marked preference for anticipation in the gesture production. It is also quite important to say that in accordance with McNeill's remarks on the question of co-occurrence of gesture and speech (2005), gesture strokes and lexical affiliates are generally produced in overlap in our corpus with only 22 % of strokes being completed before the production of the corresponding speech affiliate.

At phrase level (GPhrase vs. IP), the tendency is exactly similar with an anticipation of Gphrase of the same order as the one at word level (70 % of Gphrases start before the onset of IPs). A paired T-Test showed that the anticipation of GPhrases on IPs is significant: the mean difference (M=-0.19, SD=0.79, N=244) is significantly greater than zero (t(243)=-3.8, two-tail p=0.0001). A 95 % C.I. about GPhrase/IP onset is (-0.29, -0.09). Although the percentages are not clearly cut for the offset (61 % of gesture offset occurring after IP offset), the paired T-Test showed a mean difference (M=-0.22, SD=0.86, N=244) significantly greater than zero (t(243)=-3.96, two-tail p=9.7e-05). A 95 % C.I. about GPhrase/IP offset is (-0.32, -0.11). In all the occurrences, as opposed to the production of lexical affiliates and gesture strokes, an overlap between the production of Gphrases and IPs was observed. There was no occurrence of a Gphrase completed before its corresponding IP.

Lastly, comparing the production of gesture phrases and lexical affiliates, we only found 21 cases (8.6 %) where the gesture phrase was completed before the production of the lexical affiliate (although it was overlapping the IP containing the affiliate). Most of the cases contained some verbal hesitation as in the following example:

le village / il fait une espèce de / il est sur une espèce de colline [The village makes some sort of / is on some sort of hill.]

Where the speaker produces two identical iconic semi-spherical gestures representing a 'hill'. What is apparent here is that the idea of a hill had already formed in the speaker's mind but due to the false start, the first gesture is not synchronized with the lexical affiliate 'hill'.

Concerning the comparison of gesture and speech production of Gphrases vs. affiliates, we note that in 95 % of the cases, the Gphrase onset starts before the onset of the affiliate. The paired T-Test showed a mean difference (M=-0.82, SD=0.76, N=244) significantly greater than zero (t(243)=-16.90, two-tail p=6.50e-43). A 95 % C.I. about GPhrase/affiliate onset is (-0.92, -0.72). A high proportion (75 %) of Gphrase offsets occur after the offset of affiliates. Once again, the paired t-test provided evidence that the mean difference (M=0.595, SD=0.92, N=244) is significantly greater than zero (t(243)=10.05, two-tail p=4.21e-20). A 95 % C.I. about GPhrase/affiliate offset is (0.47, 0.71).

	% of gestures starting		% of gestures ending	
	Before speech	**After speech**	**Before speech**	**After speech**
Gstroke/ Affiliate	72	28	12	87
Gphrase/IP	70	30	39	61
Gphrase /Affiliate	95	5	25	75

Table 1: Percentage of gestures starting/ending before or after speech.

The results concerning gesture-speech timing relationships may be summarized in the following figure:

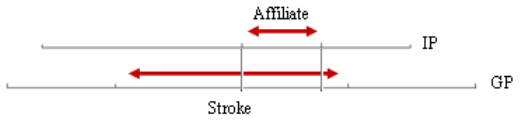


Figure 6: Timing relationships between gesture and Intonation Phrase, and between gesture stroke and lexical affiliate.

4. Discussion and conclusion

In this paper, we presented the results of one of the few studies on gesture-speech synchrony in the case of iconics. The choice of the iconic dimension was justified by the explicit relationship which exists between redundant iconics and lexical affiliates (namely words).

The results in this study – obtained from 244 iconics produced by 6 speakers during 75 minutes of spontaneous French – show that the timing relationships between gesture and speech are much alike in French and in English, as opposed to Chinese. Indeed, in Chinese, Chui (2005:878) found a higher proportion of gestures synchronized with speech than gestures anticipating speech (60.1 % vs. 35.6 %). In English, on the contrary, Schegloff (1984), who worked on deictics, observed that gesture strokes are generally produced in anticipation to the lexical affiliate. In a recent study, Leonard & Cummins (2009) also find an anticipation of gesture in English. Their work was more precisely based on the temporal alignment of beats' phases with lexical affiliates. They showed – on a very restricted corpus though – that the onset of the gesture stroke anticipated on the vowel onset in the corresponding affiliate. They also found, like in the present study, that the gesture offset occurred after speech. Although we did not have any movement detection device during the recording of the corpus[1] (and would therefore not reach the same degree of precision as Leonard & Cummins), the corpus has also been transcribed into phonemes so we should be able to go into finer detail in the future. More refinement will also be needed concerning the relationship between

[1] This type of recording would not be quite possible with spontaneous interactions.

gesture stroke and other speech and gesture dimensions. For instance, Rochet-Capellan et al. (2008) showed that in French and Portuguese, deictics' alignment with speech was very much dependent on the number of syllables in the co-occurring speech: although they found general gesture-speech synchrony, they also observed a 'tendency to delay pointing events with the increase of *n-syl*' which could result 'from the interaction between the two systems' (p. 160). Working on deictics as well, Kranstedt et al. (2006) found that the initiation of the gesture generally starts slightly after the co-occurring speech and that the stroke generally ends before the affiliate (p. 145). The difference between these last two studies and our results may very well lie in the fact that they are based on experimental corpora, whereas the present study is based on uncontrolled speech, but it would be interesting in a future study to investigate whether the variability in gesture-speech timing can be explained by different gesture amplitude. In their experimental setting Kranstedt et al. (2006) the participants were pointing to objects on a table, some of which were quite near the participants, others being quite distant. Our complex annotation on the CID which codes gesture amplitude would make such an enquiry possible.

What we added to the studies on timing relationships was the fact that not only gesture strokes (i.e. the relevant part of the gesture) and lexical affiliates could be compared, but that we can also compare the timing relationships between gesture and intonation phrases, since the lexical affiliate is in the same type of relationship with the entire IP as the stroke is with the gesture phrase, which means that a correspondence can be established between stroke and lexical affiliate as between gesture phrase and IP. At phrase level, the timing relationships are of the same order as those at word level. This corroborates what Loehr (2004) found for English, although his results showed higher gesture / speech synchrony (but he mentions variability). This difference can be explained by the fact that Loehr considered all gesture types.

At last, this study shows one of the many interesting aspects of the annotation of multimodal corpora, since only this type of annotation allows a comparison between units from different modalities, such as co-verbal gestures and speech: some of the studies quoted in this paper have been produced without any systematic annotation of either gestures or speech units. They certainly helped in formulating hypotheses on timing relationships, but it is extremely difficult to obtain precise results in terms of temporal alignment, even when one watches a video frame by frame, whereas greater precision can be attained when using annotation software like Praat for speech and Anvil for gesture phenomena. The results in such studies can be used to improve the generation of animated agents.

5. Acknowledgements

This research is supported by the French National Research Agency (Project number: ANR BLAN08-2_349062) and is based on annotations made by various team members besides the author of the current paper. The OTIM project is referenced on the following webpage: http://aune.lpl.univ-aix.fr/~otim/. The author thanks the reviewers for their valuable comments and suggestions on a previous version of this paper.

6. References

Beattie, G. and R. Aboudan (1994). Gestures, pauses and speech - an experimental investigation of the effects of changing social-context on their precise temporal relationships. *Semiotica*, 99, pp. 3--4.

Bertrand, R., Blache, P., et al. (2008). Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. *Traitement Automatique des Langues*, 49(3), pp. 105--133.

Blache, P., Bertrand, R., et al. (2009). Creating and Exploiting Multimodal Annotated Corpora: The ToMA Project. In M. Kipp (ed.), *Multimodal Corpora*. Berlin, Heidelberg: Springer-Verlag, pp. 38--53.

Boersma, P. & Weenink, D. (2009). *Praat: doing phonetics by computer (Version 5.1.05)* [Computer program]. Retrieved May 1, 2009, from http://www.praat.org/

Campbell, N. (2009). Tools and Resources for Visualising Conversational-Speech Interaction. In M. Kipp et al. (Eds), *Multimodal Corpora. From Models of Natural Interaction to Systems and Applications*. Berlin, Heidelberg: Springer-Verlag, pp. 176--188.

Chui, K. (2005). Temporal Patterning of Speech and Iconic Gestures in Conversational Discourse. *Journal of Pragmatics*, 37, pp. 871--887.

Ferré, G. (2002). Les pauses démarcatives déplacées en anglais spontané : marquage prosodique et kinésique. *Lidil*, 26, *Gestualité et syntaxe*, pp. 155--169.

Kendon, A. (1980). Gesture and speech: two aspects of the process of utterance. In M.R. Key (ed.), *Nonverbal Communication and Language*, The Hague: Mouton, pp. 207--227.

Kida, T. & Faraco, M. (2008). Prédication gestuelle. *Faits de Langues*, 31-32 (La prédication), pp. 217--226.

Kipp, M. (2001). Anvil - A Generic Annotation Tool for Multimodal Dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1367--1370.

Kipp, M. (2003). Gesture Generation by Imitation. From Human Behavior to Computer Character Animation. PhD Thesis, Saarbrucken: Saarland University.

Kipp, M., Neff, M., et al. (2007). An annotation Scheme for Conversational Gestures: How to Economically Capture Timing and Form. In *Proceedings of Language Resources and Evaluation*, 41, pp. 325--339.

Kranstedt A., A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth. (2006). Deictic object reference in task-oriented dialogue. In G. Rickheit and I. Wachsmuth (eds.), *Situated Communication*, Berlin: Mouton de Gruiter, pp. 155–208.

Kratzer, A. & selkirk, E. (2007). Phase theory and prosodic spellout: The case of verbs. *The Linguistic Review,* 24**,** pp. 95--135.

Leonard, T. and Cummins, F. (2009). Temporal Alignment of Gesture and Speech. In *Proceedings of Gespin*, Poznan, Pologne (24-26 septembre). [CD-Rom].

Loehr, D. (2004). *Gesture and Intonation*. PhD Thesis. Georgetown University.

McNeill, D. (1992). *Hand and Mind : What Gestures Reveal about Thought*. Chicago and London: The University of Chicago Press.

McNeill, D. (2005). *Gesture and Thought*. Chicago and London: The University of Chicago Press.

Nobe, S. (2000). Where do *most* spontaneous representational gestures actually occur with respect to speech? In D. McNeill (Ed.), *Language and Gesture*. Cambridge: CUP, pp. 186--198.

Rochet-Capellan, A., Vilain, C., Dohen, M., Laboissière, R. & Schwartz, J.-L. (2008). Does the Number of Syllables Affect the Finger Pointing Movement in a Pointing-naming Task? *8th International Seminar on Speech Production (ISSP 2008)*. Strasbourg, pp. 257--260.

Schegloff, E. A. (1984). On Some Gestures' Relation to Talk. In J. M. Atkinson and J. Heritage (Eds.), *Structures of Social Action*. Cambridge: CUP, pp. 266--298.

Selkirk, E. (1978). On Prosodic Structure and its Relation to Syntactic Structure. In T. Fretheim (Ed.), *Nordic Prosody II*. Trondheim: Tapir, pp 111--140.

The Bielefeld Speech and Gesture Alignment Corpus (SaGA)

Andy Lücking*, Kirsten Bergman*, Florian Hahn*, Stefan Kopp*, Hannes Rieser*†

*CRC 673 "Alignment in Communication", B1 "Speech-Gesture Alignment"
†CRC 673 "Alignment in Communication", X1 "Multimodal Alignment Corpora"
Bielefeld University
{Andy.Luecking|Kirsten.Bergmann|Fhahn2|Stefan.Kopp|Hannes.Rieser}@uni-bielefeld.de

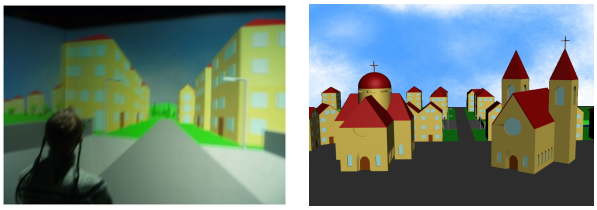
Abstract

People communicate multimodally. Most prominently, they co-produce speech and gesture. How do they do that? Studying the interplay of both modalities has to be informed by empirically observed communication behavior. We present a corpus built of speech and gesture data gained in a controlled study. We describe 1) the setting underlying the data; 2) annotation of the data; 3) reliability evalution methods and results; and 4) applications of the corpus in the research domain of speech and gesture alignment.

1. Introduction

In face to face conversation, interlocutors co-produce language and gestures. The term 'gesture' refers to gesticulations according to Kendon's continuum (Kendon, 1980), that are spontaneous co-verbal hand and arm movements which are meaningful and contribute to the conversational participants' contributions. Both, words and gesture, are temporarily and semantically coupled so that they cohere into bimodal information units (McNeill, 1992). To put it in psycholinguistic terms: speech and gestures of a speaker are *aligned* (Pickering and Garrod, 2004). For the timespan of a dialogue they enter into crossmodal signs called *bimodal* or *multimodal ensembles* (Lücking et al., 2008). However, to date there is no systematic account for the division of labour between verbal and non-verbal means for their cooperative consitution of a common meaning.

We address this challenging topic in an interdisciplinary way viewing it from a linguistic and a computer science perspective. Theoretical linguistic reconstructions, on the one hand, allow for a formally explicit as well as a precise modelling of the interface between speech and gesture. The implementation of theoretical models with computational means, on the other hand, enables us to simulate multimodal communicative behavior in virtual agents or robots. Both research lines necessitate a rich empirical basis in the form of a detailed and systematically annotated multimodal corpus. In Section 2 we present the Bielefeld Speech and Gesture Alignment (SaGA) corpus. We describe the primary experimental data as well as the secondary annotation data. Corpus evaluation in terms of interrater reliability is presented in Section 3. In order to compare the concordance of gesture performance transcriptions we distinguish two kinds of data types and apply chance-corrected agreement measures as well as a method developed in Bergmann and Kopp (2009a) that is based on the translations of annotation predicates into angle measures. Applications from linguistics and computer science that exemplify how the SaGA corpus is utilized in investigating and simulating the alignment of speech and gesture are given in Section 4.



(a) VR presentation phase (b) Part of stimulus: church square with two churches

Figure 1: Experimental setting

2. The SaGA Corpus

The primary data of the SaGA corpus are made up of 25 dialogs of interlocutors which engage in a spatial communication task combining direction-giving and sight description. There is extensive evidence that "speakers gesture more when they talk about spatial topics than when they talk about abstract or verbal ones" (Alibali, 2005, p. 313). This scenario is, therefore, well-suited for systematically studying aspects of natural speech and gesture utterances used to communicate information about the shape of objects and the spatial relations between them. The stimulus is a model of a town presented in a Virtual Reality (VR) environment (see Figure 1(a)). The VR scenario affords better determination and experimental control of the content of multimodal messages. Additionally, it secures that all participants receive the same stimulus. Upon finishing a "bus ride" through the VR town along five landmarks (see for instance the church square with two churches in Figure 1(b)), a router explained the route as well as the wayside landmarks to an unknown and naïve follower.

2.1. Primary Data

Our primary data consists of 25 direction-giving dialogs. Audio- and videotapes were taken of each dialog. For the videotape, three synchronized camera views were recorded (see Fig. 2), as well as body movement data and eye-tracking data from the router. In total, the SaGA corpus consists of 280 minutes video material containing 4961 iconic/deictic gestures, approximately 1000 discourse ges-

Figure 2: Experimental dialogue situation from three camera views, capturing the router (left), the follower (right) and the dialog scenario as a whole (middle).

tures and 39,435 words. To our knowledge, this is the largest and most comprehensive collection of naturalistic, yet controlled, systematically annotated (see below) speech-gesture data currently available.[1] Our multimodal dialogue data are stored, retrieved, and transformed within the Ariadne system (Gleim et al., 2007) which is used as an *Alignment Corpus Management System*.

2.2. Secondary Data

The data has been completely and systematically annotated[2] based on an annotation grid that has been developed according to theoretical considerations and refined in pilot annotation sessions.

All gestures have been segmented in order to specify the *stroke* phase (Kendon, 1980). The gestures (i.e., strokes) are then typed for belonging to a certain kind, namely *deictic*, *iconic* or *discourse*. 'Iconic', a term coined by McNeill (1992), who alludes to a Peircean trichotomy (Peirce, 1867), is an "umbrella term" (cf. Eco (1976)) that covers a variety of different signifying methods. (Müller, 1998), drawing on the work of (Wundt, 1911), sets up a more fine-grained classification of gestures according to the distinction of four *techniques of representation* on the ground of what the hands *do*. According to our domain of application we adopt or modify the sets of representation techniques as posited by (Müller, 1998; Kendon, 2004; Streeck, 2008). The classification of gestures within SaGA now distinguishes the following eight representation techniques:

(1) *Indexing*: pointing to a position within gesture space;
(2) *Placing*: an object is placed or set down within gesture space;
(3) *Shaping*: an object's shape is contoured or sculptured in the air;
(4) *Drawing*: the hands trace the outline of an object's shape;
(5) *Posturing*: the hands form a static configuration to stand as a model for the object itself.
(6) *Sizing*: indicating distances or sizes;
(7) *Counting*: iconic representation of a tally sheet;
(8) *Hedging*: a depiction of uncertainty (typically by a wiggling or shrugging movement).

In addition, each gesture has been coded for its morphology consisting of handshape, wrist position, palm, back of hand orientation. Movement within any of these dimensions is coded in terms of movement features.

- To code *handshape* we use a modified ASL (American Sign Language) lexicon.
- *Palm orientation* is devoted in terms of the direction of an axis orthogonal to the palm, whereby the following six speaker-centric half-axes were used (Herskovits, 1986): forward, backward, left, right, up and down. Up to three of these basic values are combined to encode diagonal or mixed directions, e.g. 'up/right' or 'up/right/forward'. In order to capture palm movements it is possible to build a temporal sequence of these values by means of the ">"-concatenator. 'up>down', for instance, denotes an upwards-downwards movement sequence.
- The orientation of *back of hand* is treated like palm orientation.
- We use *wrist position* for anchoring a gesture within regions of gesture space like "right of body, at the height of shoulder". In addition, the extension of a gesture is specified *via* its distance to the gesturer's body.
- For dynamic gestures the *movement direction* is annotated in terms of the six cardinal directions in space. As already described for palm and back of hand orientation, combinations and sequences of the categories are used to describe directions in between the six basic values as well as temporal sequences.
- To further classify the type of movement trajectory, we distinguish between linear and curved movements. Assume, for instance, the sequence of orientations 'up>right>down>left'. If it is performed linearly, the resulting trajectory is a square whereas it would be a circle if the same sequence would be performed in curved fashion.

We also transcribed interlocutors' speech on the level of words. The dialogs of the corpus are enriched with further information about the overall discourse context. For this purpose, the utterance is broken down into clauses, each of which holding to represent a proposition. Each clause then is annotated by its associated communicative goal. Denis (1997) developed several categories of communicative goals that can be distinguished in route directions. We revised and refined these for our purposes into four categories: (1) Naming a landmark; (2) Landmark property description; (3) Landmark construction description; or (4) Landmark position description.

Following Halliday (1967) we distinguish the thematization structuring of clauses in terms of *theme* and *rheme*. Additionally, the information foci *given*, and *new* are annotated and, borrowing the terminology of (Stone et al., 2003), classified according to the information states 'private' and 'shared'

The gestures of a subset of seven dialogs have also been annotated semantically. Gestures used in object descriptions have been coded for the descriptions referent and some of the referent's spatio-geometrical properties. These object features are drawn from an imagistic representation built for the VR stimulus of the study. Note that this kind of information is hardly unequivocally available for field data.

[1] There is a more multifarious collection of routes, though, hosted at the McNeill Lab (`http://mcneilllab.uchicago.edu/`) which comprises about 13 direction dialogs of different languages (English, Chinese, Huichol) – see also (McCullough, 2005) (McCullough, p.c.).

[2] We used Praat (www.praat.org) for speech transcription and Elan (www.lat-mpi.eu/tools/elan/) for gesture annotation.

3. Reliability Assessments

The annotation data has been evaluated in terms of interrater reliability. Here, a qualitative distinction has to be made, namely the distinction between Type I vs. Type II ratings (Gwet, 2001). Type I measurements are those where the human interpretation effort leading to a rating is well-understood and the outcome easily interpretable. To the contrary, this is not the case for measurements of Type II. Note that Type I ratings usually make up data on an interval or ratio scale, wheras Type II ratings are strongly associated with nominal scales. Accordingly, this difference has to be accounted for in evaluations of respective annotations: Type II ratings have to be adjusted for chance-based agreements (Cohen, 1960), whereas "chance" has no interpretation in Type I ratings. However, in the context of the latter but not the former one can speak of annotation errors. The gesture annotation comprises both types of annotation data, Type I and Type II. The classification of gestures in terms of representation techniques, reference ojects and dialogue context information is interpretive and therefore of Type II. The respective annotation labels are categories on a nominal scale. Descriptions of gesture morphology make up data of Type I. With one exception (hand shape, see below), the labels for annotating a gesture performance are ordered on an ordinal scale. Accordingly, we employ different methods in order to evaluate annotations of representation techniques and context information on the one hand, and annotations of gesture morphology on the other hand. As a chance-corrected coefficient determining the level of agreement to be found in Type II data, we calculate the first order agreement coefficient AC1 developed by Gwet (2001). In order to assess the extent of association between annotations of the Type I gesture morphology, we employ an approach based on angle measures previously used by Bergmann and Kopp (2009a).

3.1. Type II Data.

In the run-up of the reliability study we set a reasonable agreement level of 70% with an α-error of 0.05 and a β-error of 0.85 for Type II annotations. The appropriate sample size of 477 gestures has been drawn from gesture annotations. The Type I morphology sample has been classified by four, the Type II technique sample by three annotators. The resulting first-order agreement coefficient AC1 for gestures' representation technique rating is 0.784. Its confidence interval is (0.758, 0.81). The proportion of agreement on gestures' representation techniques, given that the agreement is not due to chance, is significantly greater than 75%. In particular, this result complies with our reliability level initially demanded. The degree of reliability of the annotations of reference objects and context information was calculated for one dyad taken from the subset annotated for this information. The agreement coefficient AC1 for the classification of reference objects was 0.91, for information structure 0.95, for information state 0.86, and for communicative goal 0.88. All values are collected in Table 1. In sum, the highly interpretive Type II data show a reasonable degree of interrater reliability.

Technique	Referent	InfoStruc	InfoState	Goal
0.784	0.91	0.95	0.86	0.88

Table 1: Overview of Type II data reliability evaluation. Values denote AC1 coefficients.

3.2. Type I Data.

The annotations that make up the secondary Type I data of the SaGA corpus transcribe the movement of a gesture within gesture space – cf. the afore-mentioned annotation description. The gesture space is a three-dimensional region which is spanned over the saggital, transversal, and frontal planes of a speaker. The respective directions thus have a clear spatial interpretation. Nevertheless, annotators may map an observed movement onto different category labels or simply err. However, the disagreement between, say, "movement to the right" and "movement to the right and slightly down", is less than that between "movement to the right" and "movement to the left". Comparing just for sameness of annotation labels would not capture the degree of spatial difference between them. In other words: treating movement annotations as nominal data will miss their ordinal scale information[3]. We address this problem by translating the annotation labels into angular measures which can be analyzed in terms of numeric differences. The smallest angular deviation is 2.36° for the movement direction of hand shapes, the biggest one is 46.16° for back of hand orientation. On average, the angular difference for gesture morphology as a whole is 27° (with average standard deviation SD = 45). Given that the annotation categories resolve gesture space into "slices" of 45° each, the average difference comes close to the theoretically undecidable mean value of 22.5° (45°/2). Table 2 provides an overview of the angular deviations between annotators.

3.3. Hand Shapes.

Evaluating the annotation of hand shapes requires a special treatment, since the categories developed to classify the hand shape observed comprise both Type I and Type II shares. In the first instance, there is a set of basic shapes derived from the ASL lexicon. These Type I labels are then enhanced by Type II modifiers such as "loose" or "spread". The strategy we pursue is to map all modified hand shapes onto their basic type and treat them as Type I data. As a result, we found that the four annotators agree on 83% (AC1 = 0.9, to give the Type II statistics for comparison) of the hand shapes within the reliability sample of gestures.

In sum, the evaluation of the secondary data of the SaGA corpus reveals a satisfactory degree of reliability. Chance-corrected agreement on Type II data surpasses the self-set threshold of 70%. Observed interrater agreement on Type I data results in angular values which, by and large, denote rather harmless dissent between annotators. Hence, the SaGA corpus provides a reproducible data base which can be exploited for empirically driven research.

[3] Since the movement annotation categories are coarse-grained in the sense that they map a range of positions within gesture space onto just one category, they are ordinal rather than interval or ratio scaled.

BoH orient	BoH dir	Palm orient	Palm dir	HandShape dir	Wrist dir	HandShape
20.66° (2.47)	46.14° (13.64)	19.14° (1.92)	36.86° (20.33)	2.36° (1.11)	37.08° (6.5)	83% (AC1 = 0.9)

Table 2: Overview of Type I data reliability evalution. Values denote mean angular deviation between annotations. The respective standard deviation is given in parenthesis. "BoH" stands for "Back of Hand"; "orient" and "dir" abbreviate "orientation" and "direction of movement", respectively. For the sake of completeness the Table also lists the percentage of agreed Hand Shapes – for details, please consult the text.

4. Applications

So far, the SaGA corpus is put to use in two application domains. First, the gesture annotation is used to build an interpretive domain ontology, that is, an underspecified semantic representation of gesture morphology arranged in a typological grid. Second, the annotation data are used to trigger Baysian networks of gesture production as depending from semantic and discourse context factors. Both applications are shortly illustrated subsequently.

4.1. The Typological Grid Methodology

Considering SaGA, the question is: Are the gestures observed, lines, rectangles, the three-dimensional entities arising from them, idiosyncratic tokens or are they systematically used in one datum by two agents and throughout the whole SaGA corpus by many or even all agents? In order to investigate both these typological questions we set up a typological grid (Rieser, 2010) for one datum (SaGA video film 5) in the following way: gestures build a space consisting of hierarchies of simple and more complex morphological entities. The most basic properties we have are the individual annotation predicates like hand shape or palm orientation. For example, for a horizontal line we need the predicates hand shape, wrist movement and palm orientation. The annotation predicates' values are atoms of the gesture space, called features and represented in attribute-value matrices (AVMs). Only unified do these single bits of information describe a horizontal line, as represented by the following AVM (*RH* abbreviates "Right Hand", *FC* abbreviates "Feature Cluster"):

$$\begin{bmatrix} \textit{R-Line-RH} & \\ \text{R-FC-RH-1a} & \begin{bmatrix} \textit{R-FC-RH-1a-cat} & \\ \text{HandShape} & \textit{G} \\ \text{PalmOrient} & \textit{PDN} \\ \text{BoHOrient} & \textit{BAB} \end{bmatrix} \\ \text{R-FC-RH-2a} & \begin{bmatrix} \textit{WristMovement-RH-1a-cat} & \\ \text{PathofWrist} & \textit{Line} \\ \text{WristLocMovDirection} & \textit{MR or ML} \end{bmatrix} \end{bmatrix}$$

The single features form the most basic stratum of the gestural space and the kernel of our observational language. We also set up 0-dimensional entities originating from indexing which are considered to have no spatial extension. Lines come in different shapes and directions, straight, bent, horizontal, vertical and so on. They form the one-dimensional layer below the features and the theoretically motivated cluster layers. Similar to the line distinction, we have two-dimensional entities, rectangles, squares and so on, followed by three-dimensional entities

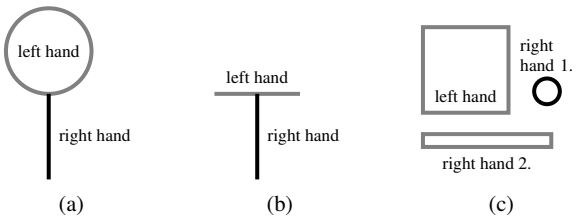


Figure 3: Illustrations of $n+m$-dimensional composite entities ($0 \leq n, m \leq 3$). Reproduced after (Hahn and Rieser, submitted).

such as cuboids, spheres or prisms. An interesting typological fact is that we get composites of n-dimensional entities, the most functionally conspicuous ones being lines touching circles orthogonally from the outside, horizontal and orthogonal lines meeting or two objects held and related to a previously introduced one – see Figure 3 for illustrations. Thus, the typological grid provides a compositional, semantic interpretation for gestures. The methodology applied here secures that even the most complex semantic features of gestures are strictly tied to annotation predicates. It is the systematic and fine-grained annotation of SaGA that makes this empirical backing of gesture meaning feasible. This in turn exposes the typological grid, and hence the reconstruction of gesture meaning, to Popperian falsifiability, a feature that sets this methodology apart from qualitative or interpretive inspections and exemplar-based analyses.

The questions we have to investigate with respect to the grid are: How many features do we use, how many gestures of which dimensions do exist, how many composites of which dimensional parts are there and so on. Statistically based answers to these questions tell us which simple and complex gestural forms Router and Follower exploit. The following results emerged for the grid data (see (Hahn and Rieser, submitted)): Generally speaking, the Router concentrates on depicting routes, regions and locations as well as objects as (part of) landmarks. Composites consisting of $n \geq 2$ gestures provide the possibility to "hold" the landmarks and specify the route to them: at the same time both, landmark and route are relationally placed in Router's gesture space. Interestingly, the Follower sets up his interactive map using one-dimensional gestures most of his time. In other words, he concentrates on representing routes. For both, Router and Follower, the right hand is dominating when gesturing. The Router uses far more two-handed composites than the follower. He populates gesture space with more objects than the Follower does. Since gesture space functions as depictional model, his gesture space is more informative than the Follower's. A series of interest-

ing results emerged with respect to the "atoms" of the gesture hierarchy, the features and how they enter into clusters: The five features, HandShape, BoHOrientation, PalmOrientation, WristPosition, and WristMovementDirection are most frequently used by both Router and Follower in their left and and right hands, respectively. The annotationally motivated grouping of the features HandShape, BoHOrientation and PalmOrientation into feature cluster at the outset of the typological work thus gets statistical support. At the same time the large number of WristPosition features and WristLocationMovementDirection features motivates the set up of clusters for WristPosition and WristMovement. Both Router and Follower predominantly use their right hands. This can be seen from the greater number of feature clusters there.

4.2. Autonomous generation of speech and gesture

The SaGA corpus is also used as an empirical basis to model speech and gesture production. We have proposed an architecture that simulates the interplay between the two modes of expressiveness on two levels (Bergmann and Kopp, 2009b). First, two kinds of knowledge representations – propositional and imagistic – are utilized to capture the modality-specific contents and processes of content planning (i.e., what to convey). Second, specific planners are integrated to carry out the formulation of concrete verbal and gestural behavior (i.e., how to convey it). Of particular importance in this framework is the question how to generate gestural forms from an abstract representation. According to empirical results based on the SaGA corpus, iconic gesture generation on the one hand generalizes across individuals to a certain degree and these commonalities may pertain primarily to gesture's iconicity. On the other hand, inter-subjective differences must also be taken into consideration by an account of why people gesture the way they actually do (Bergmann and Kopp, 2010). Our research methodology to investigate this puzzle of iconic gesture production is based on computational modelling: we have proposed GNetIc (*Gesture Net for Iconic Gestures*), a probabilistic network to model decision-making in the generation of iconic gestures (Bergmann and Kopp, 2009a). Individual as well as general networks are learned from annotated corpora by means of automated machine learning techniques and supplemented with rule- based decision making. Three different types of factors are included in the network to influence the resulting gestures: (1) visuo-spatial referent features, (2) linguistic and discourse context, and (3) the previously performed gesture. A prototype of the generation model is employed in an architecture for integrated speech and gesture generation. In this prototype implementation a virtual agent explains the same virtual reality buildings that we already used in the previously described empirical study. Being equipped with proper knowledge sources, i.e., communicative plans, lexicon, grammar, propositional and imagistic knowledge about the world, the agent randomly picks a landmark and a certain spatial perspective towards it, and then creates his explanations autonomously. Currently, the system has the ability to simulate five different speakers by switching between the respective decision networks built as described above. See Figure 4 for two simulation examples.

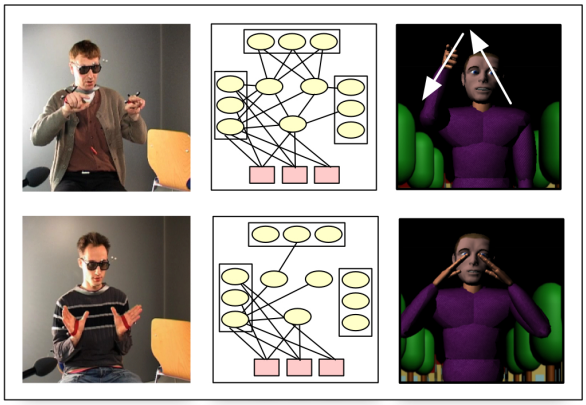


Figure 4: Two example networks (middle column) learned from individual speakers' data (left column) resulting in speaker-specific gesture production (right column). These gestures are simulated for the same referent (a tapered roof) in the same initial situation.

Analyzing the modelling results enables us to gain novel insights into the production process of iconic gestures: the resulting networks learned for individual speakers differ in their structure and in their conditional probability distributions, revealing that individual differences are not only present in the overt gestures, but also in the production process they originate from (as an example see the two differing networks in Figure 4 each of which learned from a particular speaker's data). Whereas gesture production in some individuals is, e.g., predominantly influenced by visuo- spatial referent features, other speakers mostly comply with the discourse context. So there seems to be a set of different gesture generation strategies from which individuals typically apply a particular subset.

In a comparison of learning algorithms for the network structure it turned out that at best 71.3% of the probabilistically modelled generation choices in individual networks could be predicted correctly. The accuracy achieved with general networks is 69.1%. Notably, all accuracy values clearly outperform the chance level baseline of 30%. The results show, by trend, that individual networks perform better than networks learned from non-speaker specific data (Bergmann and Kopp, to appear).

For the rule-based choices of in the model we calculated the angle between the predicted and the empirically observed orientation vector (as in the reliability study). Considering this, the mean deviation for palm orientation of 54.6° (SD = 16.1°) and the mean deviation for back of hand orientation of 37.4° (SD = 8.4°). As concerns the gesture's movement features, the movement type (linear or curved) could be predicted with 76.4% accuracy (SD=13.6). For the movement direction we distinguish between motions through the sagittal, transversal and frontal planes. Each segment in the generated movement description is tested for co-occurrence with the annotated value, resulting in a accuracy measure between 0 (no agreement) and 1 (total agreement). The mean similarity for movement direction

.75 (SD = .09). These are quite satisfying results with deviations which lie well within the natural fuzziness of communicative gestures.

To evaluate GNetIc-generated gestures in terms of their impact on the interaction between humans and machines, we are currently setting up a study to analyze if (1) semantic information uptake from gestures, and (2) the perceived interaction quality (expressiveness, naturalness etc.), is influenced by the agent's gesturing behavior. Generated gestures whose features do not fully coincide with our original data may still serve their purpose to communicate adequate spatial features of their referents – even in a speaker-specific way.

The conclusion to be taken is that the GNetIc simulation approach beside allowing an adequate simulation of speaker-specific gestures, is an valuable means to shed light onto the open research questions of (1) how iconic gestures are shaped and (2) which sources individual differences in gesturing may originate from.

5. Conclusion

The SaGA corpus is a large collection of naturalistic, yet content-controlled multimodal data. In order to make sure that its secondary data fulfill the scientific requirement of reproducibility, the data have been systematically annotated and evaluated in terms of interrater agreement methods. That ensured, the SaGA corpus is used in order to explore empirically the interplay of speech and gesture in giving directions and describing objects.

Achnowledgement

This work has been supported by the German Research Foundation (DFG) and has been carried out in the CRC 673 "Alignment in Communication".

6. References

M.W. Alibali. 2005. Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. *Spatial Cognition and Computation*, 5:307–331.

Kirsten Bergmann and Stefan Kopp. 2009a. GNetIc – Using Bayesian Decision Networks for iconic gesture generation. In Z. Ruttkay, M. Kipp, A. Nijholt, and H. Vilhjalmsson, editors, *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, pages 76–89, Berlin/Heidelberg. Springer.

Kirsten Bergmann and Stefan Kopp. 2009b. Increasing expressiveness for virtual agents – Autonomous generation of speech and gesture in spatail description tasks. In K. Decker, J. Sichman, C. Sierra, and C. Castelfranchi, editors, *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems*, pages 361–368, Budapest, Hungary.

Kirsten Bergmann and Stefan Kopp. 2010. Systematicity and idiosyncrasy in iconic gesture use: Empirical analysis and computational modeling. In S. Kopp and I. Wachsmuth, editors, *Gesture in Embodied Communication and Human-Computer Interaction*, pages 182–194. Springer, Berlin/Heidelberg.

Kirsten Bergmann and Stefan Kopp. to appear. Modeling the production of co-verbal iconic gestures by learning bayesian decision networks. *Applied Artificial Intelligence*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Michel Denis. 1997. The description of routes: A cognitive approach to the production of spatial discourse. *Current Psychology of Cognition*, 16:409–458.

Umberto Eco. 1976. *A Theory of Semiotics*. Indiana University Press, Bloomington.

Rüdiger Gleim, Alexander Mehler, and Hans-Jürgen Eikmeyer. 2007. Representing and maintaining large corpora. In *Proceedings of the Corpus Linguistics 2007 Conference, Birmingham (UK)*.

Kilem Gwet. 2001. *Handbook of Inter-Rater Reliability*. STATAXIS Publishing Company, Gaithersburg, MD.

Florian Hahn and Hannes Rieser. submitted. Corpus-based gesture typology for explaining speech-gesture alignment in mm dialogue. Submitted to SEMDial.

M.A.K. Halliday. 1967. Notes on transitivity and theme in english (part 2). *Journal of Linguistics*, 3:199–247.

Annette Herskovits. 1986. *Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English*. Cambridge University Press.

Adam Kendon. 1980. Gesticulation and speech: Two aspects of the process of utterance. In Mary Ritchie Key, editor, *The Relationship of Verbal and Nonverbal Communication*, volume 25 of *Contributions to the Sociology of Language*, pages 207–227. Mouton Publishers, The Hague.

Adam Kendon. 2004. *Gesture – Visible Action as Utterance*. Cambridge University Press.

Andy Lücking, Alexander Mehler, and Peter Menke. 2008. Taking fingerprints of speech-and-gesture ensembles: Approching empirical evidence of intrapersonal alignmnent in multimodal communication. In *LonDial 2008: The 12th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL)*, pages 157–164, King's College London, June 2–4.

Karl-Erik McCullough. 2005. *Using Gestures during Speech: Self-Generating Indexical Fields*. Ph.D. thesis, The University of Chicago, Chicago, Illinois.

David McNeill. 1992. *Hand and Mind – What Gestures Reveal about Thought*. Chicago University Press, Chicago.

Cornelia Müller. 1998. *Redebegleitende Gesten. Kulturgeschichte – Theorie – Sprachvergleich*, volume 1 of *Körper – Kultur – Kommunikation*. Berlin Verlag, Berlin.

Charles Sanders Peirce. 1867. On a new list of categories. In *Proceedings of the American Academy of Arts and Sciences Series*, volume 7, pages 287–298.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Hannes Rieser. 2010. On factoring out a gesture typology from the bielefeld speech-and-gesture-alignment corpus

(saga). In Stefan Kopp and Ipke Wachsmuth, editors, *Proceedings of GW 2009*.

Matthew Stone, Christine Doran, Bonnie Webber, Tonia Bleam, and Martha Palmer. 2003. Microplanning with Communicative Intentions: The SPUD System. *Comput. Intelligence*, 19(4):311–381.

Jürgen Streeck. 2008. Depicting by gesture. *Gesture*, 8(3):285–301.

Wilhelm Wundt. 1911. *Völkerpsychologie. Eine Untersuchung der Entwicklungsgesetze von Sprache, Mythus und Sitte*, volume Erster Band: Die Sprache. Erster Teil. Wilhelm Engelmann, Leipzig.

A Multimodal Corpus Recorded in a Health Smart Home

A. Fleury[1], M. Vacher[2], F. Portet[2], P. Chahuara[2], N. Noury[3,4]

[1] Univ Lille Nord de France, F-59000 Lille, France and EMDouai, IA, F-59500 Douai, France
[2] LIG Lab., Team GETALP, UMR CNRS/UJF/INPG/UPMF 5217, F-38041 Grenoble, France
[3] INL-INSA Lyon Lab. Team MMB, UMR CNRS/ECL/INSA/UCBL 5270, F-69621 Villeurbanne, France
[4] TIMC-IMAG Lab. Team AFIRM, UMR CNRS/UJF/INPG 5525, F-38710 La Tronche, France
e-mails:Anthony.Fleury@mines-douai.fr,Michel.Vacher@imag.fr,Francois.Portet@imag.fr,
Pedro.Chahuara@imag.fr, Norbert.Noury@insa-lyon.fr

Abstract

Health Smart Homes are nowadays a very explored research area due to the needs for home automation and telemedicine to support people with loss of autonomy and also due to the evolution of the technology that has resulted in cheap and efficient sensors. However, many studies do not include tests in real setting, because data collection is this domain is very expensive and challenging and because of the few available data sets which are anyway difficult to reuse. In this article, we present a dataset acquired in real conditions during an experiment involving 15 participants who were performing several instances of seven activities in a fully equipped Health Smart Home.

1. Introduction

The availability of cheap and efficient sensors has alleviated some issues in the development of Health Smart Homes designed to improve daily living conditions and independence for the population with loss of autonomy. Health Smart Home are nowadays a very active research area (Chan et al., 2008) and one of the greatest challenges they aim to address is to allow disabled and the growing number of elderly people to live independently as long as possible, before moving to a care institution, that could then cater for only the most severely dependent people (they are nowadays overflowed by patients, due to the demographic evolutions). Independent living also reduces the cost to society of supporting people who have lost some autonomy.

Three major goals are targeted. The first is to assess how a person copes with her loss of autonomy by continuously monitoring her activities through sensors measurements (Kröse et al., 2008; Fleury et al., 2010). The second is to ease the daily living by assistance through home automation (Wang et al., 2008), to compensate one's disabilities (either physical or mental). Examples include automatic light control and events reminder. The third is to ensure security by detecting distress situations, for instance fall that is a prevalent fear for elderly persons (Noury et al., 2007).

Furthermore, Smart Homes may assist the geriatricians to complete the autonomy evaluation of the person using scales such as the index of Activities of Daily Living (ADL) (Katz and Akpom, 1976). Such scale enables to evaluate the autonomy by considering the different activities and for each, determining if the person can perform it without assistance or with partial or complete assistance. Questionnaires are filled in by the geriatricians during interviews with the person or the family (this can thus be subjective). To fill in such grid automatically and to permit early diagnosis of autonomy loss, a first step is to automatically recognize the ADLs performed by the person in his home.

Such goal, to be achieved, must be validated on real data. Data from smart homes are rarely available and always in different formats. Indeed, no standard has been defined to exchange this kind of data. Moreover, many projects have not still acquired real data and work on simulated ones. For example, Cappelletti et al. (2008) described an ADL corpus that they planned to acquire in a smart home equipped with microphones and video cameras in order to record ADLs performed by 50 different subjects. Until now and to the best of our knowledge, the authors have not yet published the results of this experiment. Philipose et al. (2004) acquired data using RFID tags and a glove wore by the subject in a home where all the objects were tagged. The authors tried to infer the performed activity (from a set of fourteen activities) through the interaction with objects by data mining techniques. Data were recorded during experiments involving 14 subjects during 45 minutes. Kröse et al. (2008) acquired data using numerous sensors (switch, environmental, etc.) and, despite that only two activities were concerned, the validation was made on two elderly people which is still quite rare. Berenguer et al. (2008) acquired data of the general powerline in 18 flats where elderly people were leaving to detect periods of activity and meal taking. Finally, some multimodal datasets have been made available by the MIT with House_N (Intille, 2002), the Georgia Tech with Aware Home (Kidd et al., 1999) and within the GER'HOME project (Zouba et al., 2009). However, none of these includes multisource audio channels. Regarding the previously cited references, only few of them are related to multimodality.

In this paper, we present the acquisition setting of a multimodal corpus recorded in the Health Smart Home of the TIMC-IMAG laboratory. This smart home is introduced in Section 2.. The activities we focused on have been defined according to international standards. This is detailed in Section 3.. Section 4. describes the acquisition setting and the corpus recording on 15 participants. The annotation of this corpus is presented in Section 5.. Finally the acquired corpus, which includes presence, audio and human postural transitions, is detailed in Section 6..

2. Health Smart Home Description

In 1999, the TIMC-IMAG laboratory built up an Health Smart Home in the faculty of Medicine of Grenoble. This

real flat of 47m^2 with all the comfort required is composed of a bedroom, a living-room, a hall, a kitchen (with cupboards, a fridge...), a bathroom with a shower and a cabinet. An additional technical room contains all the materials required for data recording. The flat is depicted in Figure 1.

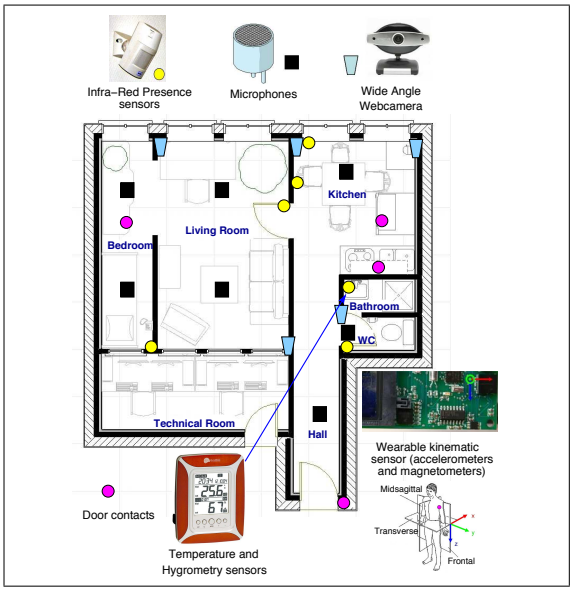


Figure 1: Map and location of the sensors inside the Health Smart Home of the TIMC-IMAG Laboratory in the Faculty of Medicine of Grenoble.

The basis of the flat is a controller collecting information from wireless PID (Presence Infrared Detectors), wireless weight scale, oxymeter and tensiometer. The exchange protocol is the Controller Area Network (CAN) bus protocol. An SQL database containing all these data is maintained. Since 2000, the TIMC-IMAG laboratory has been working with the LIG laboratory to add audio sensing technology in their health smart home. The microphones installed are omni-directional and are set on the ceiling and directed vertically to the floor. Furthermore, several webcams have been placed in the home for the purpose of marking up the person's activity and to monitor the use of some furniture (here the fridge, the cupboard and the chest of drawers). The real-time recording of these sensors is optimally distributed on four computers in the technical room:

1. The first one is devoted to the sound and speech analysis. It is an Intel Xeon 3.4 GHz based computer with 4 GB of RAM and a GNU/Linux OS. It is equipped with a National Instrument acquisition board (National Instrument PCI-6034E) to record simultaneously the seven microphone channels of the flat.

2. The second is dedicated to the capture of three of the USB2 web-cameras and is also receiving the data from the CAN bus of the flat. This one and the next one are Intel Pentium IV 3 GHz based computers with 4 GB of RAM and a MS Windows XP OS.

3. The third one is collecting the data from the two other USB2 web-cameras and from the systems which collects the temperature and hygrometry parameters in the bathroom.

4. The last one is in charge of the door contacts of the kitchen and the bedroom. It is an Intel Centrino 1.8 GHz with 2 GB of RAM with a MS Windows XP OS.

The audio channels are processed by the AuditHIS audio system (developed in C language) which is running in real time (Vacher et al., 2010). The general organization of the system is displayed on Figure 2. AuditHIS is set up through a dedicated module, while other modules run as independent threads and are synchronized by a scheduler. The 'Acquisition and First Analysis' module is in charge of data acquisition on the 8 analog channels simultaneously, at a sampling rate of 16 kHz. Each time the energy on a channel goes beyond an adaptive threshold, an audio event is detected by the 'Audio Detection' module. A record of each audio event is kept and stored on the computer. For each event, Signal-to-Noise Ratio (SNR), room, date and duration are stored in an XML file.

ACTIM6D, our home made kinematic sensor, is a circuit board equipped with a three axis accelerometer (MMA7260Q, Freescale) and a three axis magnetometer (HMC1053, Honeywell). It is kept tight on the subject and creates a new referential in which we analyze the movements of the person. The position and the orientation of this kinematic sensor is shown on Figure 1. This sensor provides an output text file containing the timestamps of the changes of posture and the beginning/end times of each recognized walking sequences (Fleury et al., 2009).

The PIDs (Atral DP8111) sense a change of temperature in their reception cone. They are mainly used for alarm systems and lighting control. Each movement in a determined zone generates a detection that is transmitted through the CAN bus to the technical room PC. Six PIDs have been placed inside the flat. There is one in the kitchen that monitors the space around the table and the cooking place; one in the living room to monitor the sofa; one in the bedroom to monitor the bed; one in the bathroom; one in the toilets and one to monitor the entrance hall. Analyzing the time series (and its evolution) of detections for the location sensor can give relevant information for determining the mobility (transitions between sensors) and the agitation (number of successive detections by a same sensor) of the person (Le Bellego et al., 2006).

Three door contacts have been placed inside furniture (cupboard, fridge and convenient). They are simulated by video cameras. Each frame is thresholded to detect the status 'open' or 'closed'. The output for these sensors is the time of occurrence of the transitions.

The last sensor delivers information on temperature and hygrometry. To be informative, this sensor is placed inside the bathroom to detect a shower. During this activity, the temperature will rise (as the person takes a hot shower), and the hygrometry will also increase. The sensor used is a commercial product (La Crosse Technology, WS810). It measures both information every five minutes.

Finally, video recording was added for two purposes. The first is to create an index of the different activities performed and the second is to simulate new sensors or create

a gold standard for one of the sensor in the dataset. These cameras are USB-2 webcams (Creative Live Cam Voice) with large angle of reception (89°). For an optimal use of the USB bandwidth, the acquisition were distributed on the two computers; in addition, low resolution (320x256) and low frame rate (15 fps) were used. One camera on each computer (bedroom and corridor) also records the sounds that are emitted in the room. All the records are encoded on-the-fly in DivX 5 and the audio channels are recorded in mono at 16 kHz and encoded in MP3. VirtualDub is used to record these videos and a filter of this software is responsible for the time-stamping of all the frames.

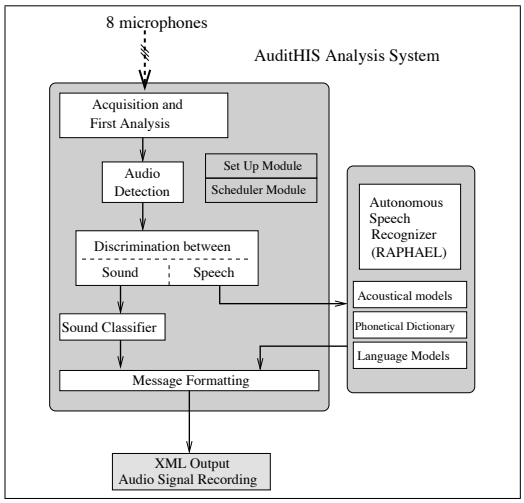


Figure 2: The AuditHIS and RAPHAEL systems

3. Definition of the Activities

To provide assistance to the dweller, a smart home must be able to recognize his/her activities. Activities can be seen as the atomic elements of a person's day and permit to recognize the context in which the person currently is or any deviance from his/her usual daily routine. Recognizing activities could also allow the immediate detection of unknown situations. For example, in Stahl et al. Stahl et al. (2008), the authors used the Leontiev's activity theory to build a hierarchy of activities which themselves can be broken down into the sequence of actions that constitute each activity. This hierarchy is then used for fine grain user-activity analysis to design better user interfaces. However, in our experiment, our focus was primary on the assessment of the person autonomy. Thus, the activities were defined with respect to the tools that geriatricians use to assess autonomy. A traditional tool is a questionnaire based on the index of independence in Activities of Daily Living (ADL) (Katz and Akpom, 1976) and on Instrumental Activities of Daily Living (iADL) (Lawton and Brody, 1969), which evaluates the person's ability to realize different activities of daily living (e.g., doing a meal, washing, going to the toilets...) either alone, or with a little or total assistance. For example, the AGGIR grid (Autonomie Gérontologie Groupes Iso-Ressources) is used by the French health system. In this grid, 17 activities including ten discriminative (e.g., talking coherently, orientating himself, dressing, going to the toilets...) and seven illustrative (e.g., transports, money management...) are graded with an A (the task can be achieved alone, completely and correctly), a B (the task cannot be totally performed without assistance or not completely or not correctly) or a C (the task cannot be achieved). Using these grades, a score is computed and, according to the scale, a geriatrician can deduce the person's level of autonomy to evaluate the need for medical or financial support. One application of smart home could be to compute automatically a score related to the geriatricians' one in order to detect a loss of autonomy as early as possible and to find compensatory solutions to avoid further increase in dependence. Moreover, such a score computed automatically could be evaluated more often and be an indication to the geriatrician on the evolution of the person (currently the evaluation is done only once a year).

ADL	iADL
Feeding	Using the phone
Dressing	Handling money
Going to the toilets	Doing laundry
Continence	Preparing and managing food
Hygiene	Handling medication
Locomotion	Doing shopping
	Using public/private transports
	Maintain accommodation

Table 1: Example of Activities of Daily Living (ADL) and Instrumental ADLs

ADL and iADL scales are described table 1. While ADL is related to daily human needs, iADL is focused on the activities that involve handling of tools (e.g., preparing food) or higher level organisation (e.g., handling money). Moreover, iADLs has been designed to assess more complex ADLs which are necessary for living in society. iADLs are also more specific than ADLs, these latter being only focused on the degree of assistance a person needs. For iADL, the formulation implies more levels (for instance, for the transport, "I can use transport on my own" versus "I can only use taxi on my own but not buses", etc.). Consequently, iADLs are generally lost before ADL-related functions and thus permit to detect incipient decline. Although these scales have been extensively used for 40 years, these are rather unstructured. A rich structured definition of activities can be found in the WHO International Classification of Functioning, Disability and Health (ICF)[1] which among body functions and diseases contains a full taxonomy to represent the human experience in relation to activity and the level of interaction with personal and environmental factors. However, due to our focus on autonomy loss we defined the activities of interest strictly based on the ADL and iADL scales. However, these activities can be easily mapped to the WHO ICF concepts. Seven activities were defined: (1) Sleeping; (2) Resting: watching TV, listening to the radio, reading a magazine...; (3) Dressing and undressing; (4) Feeding: realizing and having a meal; (5) Eliminating: going to the toilets; (6) Hygiene activity: washing hands, teeth

[1] www.who.int/icidh/

...; and (7) Communicating: using the phone. They have been chosen because they cover a large part of the common daily routine and because they were likely to be challenging the sensors network.

4. Experiment

An experiment was designed to acquire data in the Health Smart Home. Fifteen healthy participants (including 6 women) were asked to perform 7 activities, at least once, without condition on the time spent. Four participants were not native French speakers. The average age was 32 ± 9 years (24-43, min-max) and the experiment lasted from 23 minutes 11s to 1h 35 minutes 44s. A visit, before the experiment, ensured that the participants would find all the items necessary to the seven ADLs. Participants were free to choose the order of execution of the ADLs to avoid repetitive patterns. Data was indexed afterwards using video.

One particular point that was raised to the attention of the participants regarding the toilet was that the camera was recording the top of the door and the ceiling to respect their privacy. A code was established with them: if this door is partially closed and the person in the bathroom/toilets, it indicates that she is performing the elimination activity, otherwise she is performing the hygiene activity. This is a non-intrusive way to indicate the intimate activities being performed. Participants were shown what was acquired by the different cameras before signing the agreement form. Globally, the participants were relax and did not hesitate in making themselves comfortable in the flat by actions we did not anticipated (e.g. opening the window, adjust the blind, etc.). This has led to unexpected situations with for instance a participant that had a phone call during the experiment and answered it naturally. Moreover, the recording was not cancelled or moved because of external conditions. For instance, the recording of a participant was performed during a thunderstorm which artificially increased the number of detected sounds. That is why, despite the fact that the experiment could have been conducted in a more rigorous way, we believe that this piece of freedom has led to a more natural corpus.

Regarding the sensors, this flat represents a very hostile environment similar to the one that can be encountered in real home. This is particularly true for the audio information. The sound and speech recognition system presented in (Vacher et al., 2010) was tested in laboratory and gave an average Signal to Noise Ratio of 27dB in-lab. In the Health Smart Home, SNR fells to 12dB. Moreover, we had no control on the sounds that are measured from the exterior of the flat, and a lot of reverberation was introduced by the 2 important glazed areas opposite to each other in the living room. For more details about the experiment, the reader is refereed to (Fleury et al., 2010).

5. Annotation Schema

Different features have been marked up using Advene[2] developed at the LIRIS laboratory. Different softwares have been tested, such as Observer XT, ELAN or the well-known ANVIL, but Advene is the only one that have demonstrated a correct handling of large videos. Advene can be used with different formats and is able to read ANVIL-like XML schema definition. Moreover, a nice feature of Advene is to be able to markup explicit links between annotations of different tracks (e.g., marking up causal links or associations). A snapshot of Advene with our video data is presented on Figure 3.

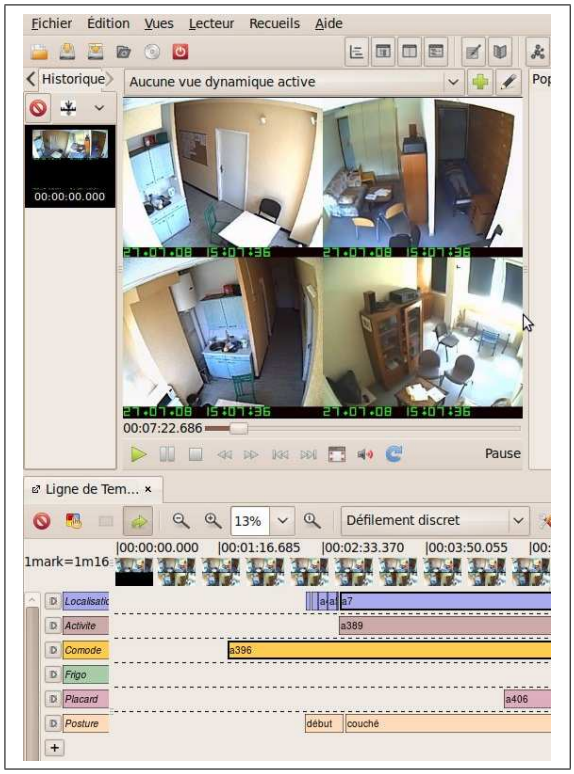


Figure 3: Snapshot of Advene

On this figure, we can see the organisation of the flat and also the videos collected. The top left is the kitchen, the bottom left is the kitchen and the corridor, the top right is the living-room on the left and the bedroom on the right and finally the bottom right represents another angle of the living room. One of the cameras is not shown on this mosaic: the bathroom/toilets. As previously said, even if a camera is set at this place, it records only the position of the door and not the activity of the person.

Advene allows to organize annotations elements — such as type of annotations, annotations, and relationships — under schemes defined by the user. In this work, a unique schema is used because of the homogeneity among the annotations types. The following annotation types have been set: location of the person, activity, chest of drawer door, cupboard door, fridge door, posture. As shown on figure 3, each annotation type represents a source of information. For the doors state, the content of the annotations is a simple text flag, '1', that indicates an interval of time where the door is open. For the other annotations a pair 'attribute=value' has been used instead, it provides more flexibility to define other attributes for the same annotation type. For instance, if we consider several people participating in the same ob-

[2] http://liris.cnrs.fr/advene/

Annotation Type	Description	Attribute	Content Type
Localization	The room in which the person is	room	text
Activity	ADL currently performed	name	text
Drawer door	If door is open or not	flag	text
Cupboard door	If door is open or not	flag	text
Fridge door	If door is open or not	flag	text
Posture	If the person is standing, lying or sitting	position	text

Table 2: Annotation Types

servation, when annotating an activity, we can include the identity of the person performing it in addition to the ADL description. If necessary, doors state can be easily translated into the 'attribute=value' pair since their content is simple text. Table 2 sums up the annotation types.

To markup the numerous sounds collected in the smart home, none of the current annotation software have shown advantages. Indeed, the duration of each sounds is too small and the number of audio channels too high (seven) to be properly managed by a resource consuming annotator. We thus developed our own annotator in Python that plays each sounds while displaying the video in the context of this sounds and proposing to keep the AuditHIS annotation or select another one in a list. About 2500 sounds and speech have been annotated in this way.

6. Acquired Corpus

6.1. PID

More than 1700 firings have been recorded. Surprisingly, the sensitivity of the PID was not as good as expected. The sensitivity to detect a change of rooms in 10 records is 80%. To succinctly recall the functioning of PIDs, they detect perturbations of a background temperature (estimated via infrared radiations) thus a person walking between the PID sensor and an area at a different temperature (such as a wall or the floor) will trigger the sensor. The problem of missing detections could be explained by the fact that the experiments have been done in summer. Thus, the difference between the temperature of the wall in the flat and the one of the participants' body would not have been sufficient to provoke an infra-red energy change in the sensor. This would be especially true when the movement is rapid. However, this problem reflects that no source is 100% reliable and that PIDs, though largely used in smart home, should be supplemented by other location related sensors (e.g., audio).

6.2. Doors Contacts

During the experiment, 136 states for the Fridge (9 per participant), 143 for the cupboard (9.5 per participant), and 91 for the chest of drawers (6 per participant) were recorded. This data is particularly interesting to track furniture usage when preparing a meal or dressing.

6.3. Audio

Every audio event was processed on the fly by AuditHIS and stored on the hard disk. For each one, an XML file was generated, containing the results of the process and the important information. These events do not include the ones discarded because of their low SNR (less than 5 dB, threshold chosen empirically). The events were then filtered to remove duplicate instances (same event recorded on different microphones).

Category	Sound Nb.	Mean SNR (dB)	Mean length (ms)	Total length (s)
Human sounds (except speech)	36	12.05	100.8	3.35
Speech	669	11.2	435	291.0
Object handling	1302	11.9	58.6	76.3
Outdoor sounds	45	9	174.4	7.85
Device sounds	72	8.03	208.5	15.1
Water sounds	36	10.1	1756.1	63.2
Other sounds	395	9.5	93.9	37.1
Overall sounds (except speech)	1886	11.2	107.8	203.3
Overall speech	669	11.2	435	291.0
Overall	2555	11.2	193.5	494.3

Table 3: Everyday Life Sounds and Speech

During the experiment, 1886 individual sounds and 669 sentences were collected. These periods have been manually annotated after the experiment. The most important characteristics of this corpus are summarized in Table 3. The detailed audio corpus is presented Table 4.

The total duration of the audio corpus, including sounds and speech, is 8 min 23 s. This may be seen as short, but daily living sounds last 0.19s on average. Moreover, the person is alone at home, therefore she rarely speaks (only the phone). Similarly, few sounds are emitted excepted during particular activities or when the person is moving in the flat.

The mean SNR of each class is between 5 and 15 dB, far less than the in-lab one. This confirms that the health smart home audio data acquired was noisy. Also, the sounds were very diverse, much more than expected in this experimental conditions were participants, though free to perform activities as they wanted, had recommendations to follow.

The speech part of the corpus was recorded in noisy conditions (SNR=11.2dB) with microphones set far from the speaker (between 2 and 4 meters) and was made of phone conversations. Some sentences in French such as *"Allo"*,

Category	Sound Classe	Sound Nb.	Mean SNR (dB)	Mean length (ms)	Total length (s)
Human sounds:		**36**	**12.05**	**100.8**	**3.35**
	Cough	8	14.6	79	0.6
	Fart	1	13	74	0.01
	Gargling	1	18	304	0.3
	Hand Snapping	1	9	68	0.01
	Mouth	2	10	41	0.01
	Sigh	12	11	69	0.8
	Song	1	5	692	0.7
	Throat Roughing	1	6	16	0.02
	Whistle	5	7.2	126	0.6
	Wiping	4	19.5	76	0.3
Object handling:		**1302**	**11.9**	**58.6**	**76.3**
	Bag Frisking	2	11.5	86	0.1
	Bed/Sofa	16	10	15	0.2
	Chair Handling	44	10.5	81	3
	Chair	3	9	5	0.01
	Cloth Shaking	5	11	34	0.1
	Creaking	3	8.7	57	0.1
	Dishes Handling	68	8.8	70	4.7
	Door Lock&Shut	278	16.3	93	25
	Drawer Handling	133	12.6	54	7
	Foot Step	76	9	62	4
	Frisking	2	7.5	79	0.1
	Lock/Latch	162	15.6	80	12.9
	Mattress	2	9	6	0.01
	Object Falling	73	11.5	60	4.4
	Objects shocking	420	9	27.6	11.6
	Paper noise	1	8	26	0.03
	Paper/Table	1	5	15	0.01
	Paper	1	5	31	0.03
	Pillow	1	5	2	0
	Rubbing	2	6	10	0.02
	Rumbling	1	10	120	0.1
	Soft Shock	1	7	5	0
	Velcro	7	6.7	38	0.2
Outdoor sounds:		**45**	**9**	**174.4**	**7.85**
	Exterior	24	10	32	0.77
	Helicopter	5	10	807	4.4
	Rain	3	6	114	0.3
	Thunder	13	7.5	208	2.7
Device sounds:		**72**	**8.03**	**208.5**	**15.1**
	Bip	2	8	43	0.08
	Phone ringing	69	8	217	15
	TV	1	10	40	0.04
Water sounds:		**36**	**10.1**	**1756.1**	**63.2**
	Hand Washing	1	5	212	0.2
	Sink Drain	2	14	106	0.2
	Toilet Flushing	20	12	2833	56.6
	Water Flow	13	7	472	6.1
Other sounds:		**395**	**9.5**	**93.9**	**37.1**
	Mixed Sound	164	11	191	31.3
	unknown	231	8.5	25	5.8
Overall sounds except speech		**1886**	**11.2**	**107.8**	**203.3**

Table 4: Every Day Life Sound Corpus

"Comment ça va" or *"A demain"* are extract of usual phone conversation. No emotional expression was asked from the participants.

According to their origin and nature, sounds have been gathered into sounds of daily living classes. A first class is constituted of all the ones generated by the human body. Most of them are of low interest (e.g., clearing throat, gargling). However, whistling and song can be related to the mood while cough and throat roughing may be related to health. The most populated class of sound is the one related to the object and furniture handling (e.g., door shutting, drawer handling, rummaging through a bag, etc.). The distribution is highly unbalanced and it is unclear how these sounds can be related to health status or distress situation. However, they contribute to the recognition of activities of daily living which are essential to monitor the person's activity. Related to this class, though different, were sounds provoked by devices, such as the phone.

The most surprising class was the sounds coming from the exterior to the flat but within the building (elevator, noise in the corridor, etc.) and exterior to the building(helicopter, rain, etc. outside). This flat has poor noise insulation (as it can be the case for many homes) and we did not prevent participants any action. Thus, some of them opened the window, which was particularly annoying (the helicopter spot of the hospital is at short distance). Furthermore, one of the recordings was realized during rain and thunder which artificially increased the number of sounds.

It is common, in daily living, for a person, to generate more than one sound at a time. Consequently, a large number of mixed sounds were recorded (e.g. mixing of foot step, door closing and locker). This is probably due to the youth of the participants and may be less frequent with aged persons. Unclassifiable sounds were also numerous and mainly due to situations in which video were not enough to mark up, without doubts, the noise occurring on the channel. Even for a human, the context in which a sound occurs is often essential to its classification (Niessen et al., 2008).

Despite the length of the experience, the number of recorded sounds is low and highly unbalanced for most classes. Thus, the record of a sufficient number of sounds needed for statistical analysis method will be a hard task. Moreover, to acquire more generic models, sounds must be collected into different environments. Another problem is that it is hard to annotate sounds with high certainty and to know the required level of detail. The corpus contains many sounds that can be seen as super class of others (Objects shocking, Exterior...). The source of the sound is also difficult to recognize, but it may be of great interest to classify sounds according to their inner characteristics: periodicity, fundamental frequency, impulsive or wide spectrum...

6.4. Video

75 video records have been collected (5 per participants, one per room). These videos have been used to mark up the different activities of daily living and are now used to create a gold standard for some sensors. In our future experiment we plan to avoid use of any video processing to make smart home system respectful of privacy.

6.5. Wearable Sensor

ACTIM6D created one file for each subject. Two of the subjects have corrupted data that made it unusable. For the others, it has been synchronized with the video data using a required movement and then processed by the algorithm (detection of postural transitions and walking periods).
The processed data showed that, contrary to the 70% of correct detection and classification of postural transitions on scenarios, the results in daily living are far from being satisfying. An ongoing work is to annotate the postural transitions in the videos, to create a new gold standard and being able to improve parts of the detection algorithm responsible of the non-detections.

7. Conclusion

Health Smart Homes, despite their recent developments, have led to few experimentations in daily living conditions. This paper presents a multimodal dataset, acquired on 15 subjects in daily living conditions, in a fully equipped and complete health smart home. It has been annotated using video cameras by, firstly, determining the beginning and end of each activities of daily living, for pattern recognition and classification purposes and by, secondly, enhancing it by adding annotations to estimate sensors accuracy. We are currently studying the opportunity of releasing part of the corpus to allow comparison of methods from other projects that works on ADL with our classification algorithm on the same dataset. Future work includes acquisition of other and larger datasets, keeping a good methodology and organization so that these datasets can be used for different tests and algorithms.

8. Acknowledgments

The authors would like to thank the participants and C. Villemazet and H. Glasson for their technical support.

9. References

M. Berenguer, M. Giordani, F. Giraud-By, and N. Noury. 2008. Automatic detection of activities of daily living from detecting and classifying electrical events on the residential power line. In *HealthCom'08, 10th IEEE Int. Conf. on e-Health Networking, Applications & Service*.

A. Cappelletti, B. Lepri, N. Mana, F. Pianesi, and M. Zancanaro. 2008. A multimodal data collection of daily activities in a real instrumented apartment. In *Proc. of the Workshop Multimodal Corpora: From Models of Natural Interaction to Systems and Applications - LREC'08*, pages 20–26, Marrakech, Morocco, 27 May 2008.

M. Chan, D. Estève, C. Escriba, and E. Campo. 2008. A review of smart homes- present state and future challenges. *Computer Methods and Programs in Biomedicine*, 91(1):55–81, Jul.

A. Fleury, N. Noury, and M. Vacher. 2009. A wavelet-based pattern recognition algorithm to classify postural transition in humans. In *17th European Signal Processing Conference (EUSIPCO 2009)*, pages 2047 – 2051, Glasgow, Scotland, Aug. 24–28.

A. Fleury, M. Vacher, and N. Noury. 2010. SVM-based multi-modal classification of activities of daily living in health smart homes: Sensors, algorithms and first experimental results. *IEEE Transactions on Information Technologies in Biomedicine*, 14(2):274–283.

S. S. Intille. 2002. Designing a home of the future. *IEEE Pervasive Computing*, 1(2):76–82.

S. Katz and C.A. Akpom. 1976. A measure of primary sociobiological functions. *International Journal of Health Services*, 6(3):493–508.

C. D. Kidd, R. Orr, G. D. Abowd, C. G. Atkeson, I. A. Essa, B. MacIntyre, E. D. Mynatt, T. Starner, and W. Newstetter. 1999. The aware home: A living laboratory for ubiquitous computing research. In *Proc. of the Second Intl. Workshop on Cooperative Buildings, Integrating Information, Organization, and Architecture*, pages 191–198. Springer-Verlag.

B. Kröse, T. van Kasteren, C. Gibson, and T. van den Dool. 2008. Care: Context awareness in residences for elderly. In *Proc. of the Conf. of the International Society for Gerontechnology*, Pisa, Tuscany, Italy, June 4–7.

M P Lawton and E M Brody. 1969. Assessment of older people: self-maintaining and instrumental activities of daily living. *Gerontologist*, 9(3):179–186.

G. Le Bellego, N. Noury, G. Virone, M. Mousseau, and J. Demongeot. 2006. A model for the measurement of patient activity in a hospital suite. *IEEE Transactions on Information Technologies in Biomedicine*, 10(1):92 – 99.

Maria Niessen, Leendert van Maanen, and Tjeerd Andringa. 2008. Disambiguating sounds through context. In *Semantic Computing, 2008 IEEE International Conference on*, pages 88 – 95, 4-7 Aug.

N. Noury, A. Fleury, P. Rumeau, A.K. Bourke, G. O Laighin, V. Rialle, and J.E. Lundy. 2007. Fall detection - principles and methods. In *Proc. 29th Annual Intl. Conf. of the IEEE-EMBS 2007*, pages 1663–1666, 22-26 Aug.

M. Philipose, K. P. Fishkin, M. Perkowitz, D. J. Patterson, D. Fox, H. Kautz, and D. Hahnel. 2004. Inferring activities from interactions with objects. *IEEE Pervasing Computing*, 3(4):50 – 57, Oct.

C. Stahl, D. Heckmann, M. Schneider, and A. Kröner. 2008. An activity-based approach to the design of user assistance in intelligent environments. In *Capturing Ambient Assisted Living Needs, International Workshop at AmI 2008 Conference*, 19 Nov.

M. Vacher, A. Fleury, F. Portet, J.-F. Serignat, and N. Noury, 2010. *New Developments in Biomedical Engineering*, chapter Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living, pages 645 – 673. Intech Book, Feb. ISBN: 978-953-7619-57-2.

J.-C. Wang, H.-P. Lee, J.-F. Wang, and C.-B. Lin. 2008. Robust environmental sound recognition for home automation. *IEEE Trans. on Automation Science and Engineering*, 5(1):25–31, Jan.

N. Zouba, F. Bremond, M. Thonnat, A. Anfonso, E. Pascual, P. Mallea, V. Mailland, and O. Guerin. 2009. A computer system to monitor older adults at home: preliminary results. *International Journal of Gerontechnology*, 8(3), Jul.

A multi-software integration platform and support for multimedia transcripts of language

Christophe Parisse* and Aliyah Morgenstern**
*Modyco, Inserm, CNRS/Paris Ouest Nanterre La Défense University
**Prismes, Paris III Sorbonne Nouvelle University
200 av de la République, 92001 Nanterre cedex, FRANCE
E-mail: cparisse@u-paris10.fr, aliyah.morgenstern@gmail.com

Abstract

Using and sharing multimedia corpora is a vital feature for research about language, but the number of different and often not easily compatible tools available makes this difficult to do. As the aims of the COLAJE project are to use multimodal linguistic data about language development in oral and sign languages, it was necessary to create a system (VICLO) that allowed sharing and using data coming from at least three different sources Clan (CHILDES), Elan (MPI) and Praat (U. of Amsterdam). For this reason, a multi-purpose storage format based on the TEI was created, which allowed us to store information coming from all (these) origins, and include every type of specific information. When part of the information is processed by a specific software, the changes are integrated later in the system without loosing information specific to other software. Thus it is possible to store information shared and not shared between the different corpus editing tools. This common base allowed us to implement complementary features such as fine-grained participant and metadata information, common visualisation and data-retrieval tools. VICLO is based on XML technology and all data can be displayed using all purpose web browsers.

1. Introduction

Using and sharing multimedia data and transcripts is a vital feature for research and applications on language, especially those based on conversation analysis as well as pragmatic and semantic analyses. Recent advances in the use of video media, speed of computers and price of video-recording material have made it much easier to gather, describe and process corpora of language that include video and direct correspondence between transcription and video, what is usually called linking between transcription and video. The editing and linking process can be realised easily today thanks to a wide choice of freely available, multiplatform, and robust software such as CLAN (childes.psy.cmu.edu/clan/), ELAN (www.lat-mpi.eu/tools/elan/), PRAAT (www.fon.hum.uva.nl/praat/) and many others (Anvil, Exmaralda, Transcriber, Transana, etc. – see http://icar.univ-lyon2.fr/projets/corinte/confection/alignement.htm for a more complete description of the tools). These tools are often open-source software, so it is reasonable to assume that the most commonly used ones will be maintained in the future by a large community of users and developers. This certainly helps people to invest into creating new video-linked corpora and databases.

These advances are highly useful for sign language and language acquisition, as in these domains using visual media along linguistic data is mandatory for different reasons. Sign language is obviously a visual medium of communication; for language acquisition it is virtually impossible to generate serious work about the semantic and pragmatic aspects of language interaction without visual support. The latter is also often mandatory to simply understand what very young children are saying because comprehension is very poor outside the (visual) context. The same remarks would also apply to other fields of linguistic analyses, such as adult interaction.

A common feature between these two domains is that in both cases it is very difficult to design a single piece of software that would cover all the uses and needs of people doing research or using this material for education purposes. Another common feature is that corpus creation (recording, transcription, linking, editing) is very costly in term of human hours of work (but not in terms of material or software). Unfortunately these two common features clash (one with the other): the high costs would suggest that any data ever produced should be used and reused as much as possible whereas differences between applications make it difficult to reuse data that was produced initially using one piece of software only. For example, software such as Clan and Transcriber allow the coding of situational information (but they code it differently) and this information does not exist (yet) in Elan and Praat, so it would be lost during the conversion process. Another example is interdependence between tier levels: Elan offers a much more powerful package than other applications so this information will also be lost during conversion. A final and serious problem is the conversion between Single Timeline Multiple Tiers (STMT) organisation of data (used by Elan, Praat, Exmaralda, etc.) and Ordered Hierarchy of Content Objects (OHCO) data (used by Clan). When data created using OHCO software are converted into STMT software, elements which are not coded for time alignment (linking) in the first case have to be modified to be handled correctly in the second case. Backward conversion may not reproduce the original hierarchy.

For all these reasons, using more than one application for the same data is difficult. At first glance, as most type of applications have importing and exporting features, it

would seem that this is not a real problem. However, conversion is always performed on a common core basis. Only the features which are shared between two applications are converted. Other data are lost, so the use of multiple software is often a case of one-way conversion from one reference software (the tool the data was build with) towards another tool that has interesting complementary features but that is used only for some specific one-time feature. A good example of this procedure is conversion from Clan towards Praat. Either the "one utterance only conversion procedure" is used and the goal is only to analyse more finely this utterance with Praat, but any modification done within Praat cannot be converted back to Clan; either "the whole file procedure" is used, but then information about participants and sequences is lost so a conversion back to Clan will result in a very different data from the original one; so conversion back and forth between Clan and Praat is unlikely to happen..

2. A multi-software integration platform

We are facing a paradox: the difference between editing tools makes it difficult to use multiple tools; but this very difference is what makes it interesting to share data between tools as they have complementary features and qualities. This could also be considered an economic issue due to the cost and labor involved in corpus creation.

2.1. Goal

Our goal is to propose a solution to these limits by using a common repository which would not be based on core features of the data designed for all types of applications to be used, but on encompassing features. This means that the common format used contains recipients for all types of data for all tools, and that it is used as a pivot and common repository. This makes the preservation of specific software information possible. Data that is specific to a tool A and unused by others is kept in the repository so that it can be reintegrated when the rest of the data has been edited and modified by a tool B, and conversely. Such an integrative system offers advantages that go beyond data sharing. It makes it easier to create complementary features such as metadata and fine-grained descriptions of target participants' behaviour because this data will benefit all corpora. It will allow us to integrate metadata from different origins, including for example OLAC (www.language-archives.org), Dublin Core (dublincore.org/documents/dcmi-terms/) or ISLE IMDI (www.mpi.nl/IMDI/). It makes it also possible to make new interrogations and to display features that could be used on data created by different tools.

The goal of the COLAJE project (financed by the ANR, France) is to create a functional platform, VICLO (French translation of Visualisation and Interrogation of Oral Language Corpora), that includes such features and is compatible with Clan, Elan and Praat and allows easy integration with other tools such as editing tools and computer linguistic tools. Compatibility with Clan includes compatibility with the new CHILDES-XML format and the Talkbank project (www.talkbank.org). The VICLO platform is demonstrated on the COLAJE website (see www.modyco.fr/corpus/colaje/viclo/).

As the purpose of the project is to create a platform that is easy to use and to maintain, the technical solutions use only open-source and easy access software. Ready made data as visualised by the final user do not need any software installation since these data can be browsed through a web navigator such as Firefox, Safari or Chrome. Processing (converting and preparing for display) is implemented in XSLT as much as possible (any transformation uses XML data as a starting point) and in Perl for conversion starting from non-XML data. The format used for the repository is XML and is based on the TEI XML format.

3. Implementation issues

3.1. Common format

The choice of the Text Encoding Initiative (TEI) as a basis for the container format for all data is only natural as TEI is based on a reliable base (XML) and is a multi-purpose storage format for language corpora. Unfortunately, it had to be quite thoroughly extended for three reasons. First, conventions for the storage of oral language data are only general guidelines and many elements related to specific metadata and tier structured layers are not optimal. There are also issues about future implementation of structured data, decomposition into words and sub-lexical units, and description of syntactic information. Second, the multiple possibilities of various applications such as Clan, Elan and Praat were not included in the design of the TEI, although sometimes it is possible to redirect the initial purpose of some parts of the TEI (see below). Third, the TEI was obviously not designed to store data specific and software specific information, which is necessary to maintain the integrity of the original information in a software specific fashion.

3.2. General purpose additions to the TEI

Four additions were made to the TEI format, following the general structure of TEI data: participant information, tier information (especially the structural organisation of the tiers), specific vocabularies for the coding of specific tiers, and fine-grained information about participants and description of the recording session. These four types of information are stored in the description profile of the TEI header (see Table 1 above). This rich information is not part of the language corpus itself, but is of vital importance for scientific purposes because it provides information about the people involved, the coding features and the organisation of the data. This is some extended type of metadata, as demonstrated by the *textDesc* feature which is specific to the COLAJE project.

The participant information constitutes the main entry into the participant, tier and vocabulary data. Participants bear no relation one to another in the structure of the

corpus (even though they often have kinship relationships!). Participant information contains elements that are directly specific to the person involved in the corpus and is usually independent from the corpus collection purposes – age, sex, socio-economic status, etc.). Participant information is linked to tier information which contains the various levels of description of the language data: orthography, phonetics, gestures, prosody, gaze, situation, actions, etc. This information is open ended as there is no limit to what future research purposes may be. Constraints on the structure of a tier are possible through the use of vocabularies, a feature that is Elan specific (note that Clan has a similar feature but this was never translated into data constraint representations). Specific structure for specific tiers such as the orthographic and phonetic tiers is not yet implemented but may be included in the future when software such as CLAN-XML and Phon will offer this feature in their internal data, or when planed integration with lexicometric or language processing tools will be performed (see future improvements).

```
<teiHeader> … <profileDesc> ...(TEI tags)
<participantStmt> …
   <!-- example generated from Elan -->
   <participant name="wit" longname="With or
   without gaze" language="en" type="With or
   without gaze" />
   <!-- example automatically generated from
   Clan -->
   <participant name='chi' longname='Antoine'
   type='participant' role='target_child'
   age='2;04.03' birth='10-APR-2006' sex='male'
   desc='description' />
...
<tierStmt> ...
   <!-- example generated from Elan -->
   <tierDesc type="With or without gaze"
   longname="With or without gaze"
   parent="participant" vocid="With or without
   gaze" xgraphic_references="false"
   align="true"/>
...
<vocabularyStmt>
   <vocabulary name="With or without gaze">
   <vocabularyDesc/>
   <token xml:id="Without gaze"> <tokenDesc/>
...
<textDesc>
  <sg name="saillant_features">
   <g name="motor">
    <v name="sitting_position" val="yes"></v>
    <v name="crawling" val="yes"></v>
```

Table 1 : Examples of TEI for Oral Corpus extensions

3.3. Coding of language data

Clan, Elan and Praat make different uses of the word *tier* because they have different underlying structures, OHCO for Clan and STMT for Elan and Praat. The TEI includes a mix of the two information structures but has no tier concept (neither in the sense described in 3.2 nor in the STMT sense). It includes a 'u' concept which is an entry into a text part, which may or not correspond to the sentence or the turn. This concept is included along the concept of timeline which allows for representation of STMT formats. Time anchors allow to implement complex linking information using the TEI. They will be used to implement sequences of utterances. We chose to keep the concept of timeline as it was the easiest way to preserve data organisation for most oral language tools. This means that all elements in Clan need to be mapped onto a timeline point or aligned (using anchors) with other elements (this allows to time reference all elements when transformations from Clan to Elan are done). The main problem is that the mapping process is somewhat arbitrary because information about overlapping is not finely detailed in Clan, when it is described. Backward transformation is possible; however it is not yet implemented because, as Clan works quite well with overlapping timelines, this is not really an issue.

```
<u wh='chi' xml:id='id2' start='23.92'
end='24.357'>
<tier type='ortho'>papi Michel .</tier>
<tier type='pho'>papi miʃel</tier>
<tier type='sit'>GDF is sitting down on the
sofa</tier>
</u>
```

Table 2 : Example of TEI for Oral Corpus coding of text data

The 'u' format was kept because this was the TEI format but the 'u' for 'utterance' should be in fact changed to 'e' for 'entry' because nothing in the data format specifies that 'u' is an utterance. It can be a piece of any size of language that is produced by one speaker only. Tiers are all included in the main entry, but it is possible for individual tiers to have specific time linking, inside the duration of the main entry, which corresponds to the notion of constraint stereotype in Elan. The orthographic line is not included in the main entry, but as a separate tier, which frees the representation from a strict text oriented classical representation and makes the coding of multimodal non linguistic data possible.

3.4. Information about specific software

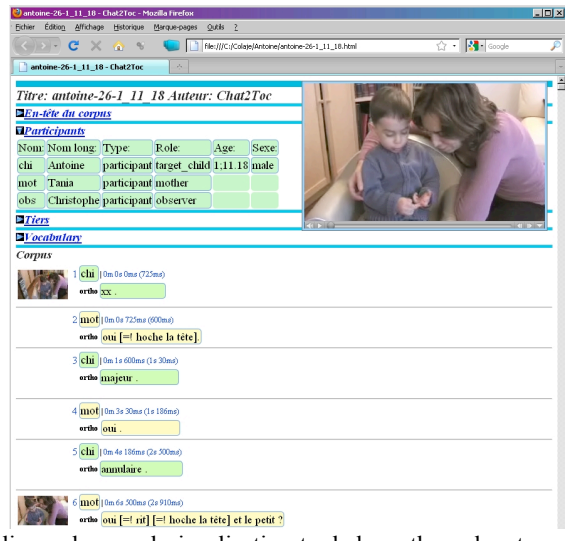
To keep specific software information in the data, and put it back when a file is changed by another type of software, it is necessary to send the changes made in a file. This is true for multi-purpose information such as the *textDesc* (see above) or more specific information about linking or tier structure. It is thus necessary to keep track of which software generated the data first, and which software it is used with later. Information about original filename, time of conversion, type and name of tools used, conversion of media software (for example from video to audio in the case of conversion from Elan to Praat) is kept in the multipurpose *notesStmt* of the TEI header. Information related to external audio and video data files is kept in the *recordingStmt* in the *sourcesDesc* part of the TEI header.

3.5. Editing text description data

Descriptions of the texts do not follow a fixed format common to all files, as ~~it~~ is usually the case for data formats, technical implementation details or even usual metadata which are normalised so as to make information available anywhere on the web. Descriptions consist~~s~~ for example of information about the age of acquisition of cognitive milestones for young children (when ~~did~~ they began to walk, for example) which is useful information for a researcher working on child language acquisition because child mobility may have an impact on their pragmatic contact with other adults. But this will be of no interest for people working on later acquisition of complex syntax. So the material to be edited and inserted in the corpus is prone to change. To this purpose, a specific PHP application was developed so that it was possible to set up the description and material to be edited using a single configuration file. Structure of XML data is the same for all possible descriptors as the specific information is stored only in the values of XML parameters and nodes, not the names of the nodes and parameters. A specific interrogation system for this data is under development. This interrogation will be automatically guided by the nature of the data stored in the descriptors.

3.6. Visualisation

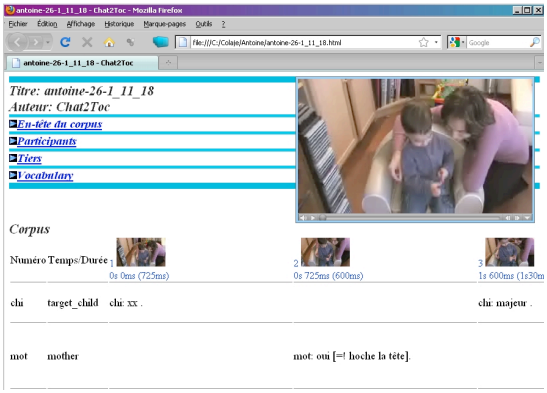
Specific tools for visualisation of the data had to be developed in order to be able to display the specific features such as the text description (textDesc) which could not be displayed by any native tool because it is original data created in VICLO. However, having at our



disposal several visualisation tools has other advantages. First, the data can be displayed without learning or installing a new editing tool. Cross-tools visualisation is made easier. Second, it is possible to create a large number of new display formats and respond more easily to specific requests because browser only software is easier to achieve and not constrained by editing needs. Finally, visualisation tools will became handy when one of the basic goal of the project, data interrogation will become a reality.

Figure 1: Text presentation of data

Visualisation is implemented through web browsers and XSL transformation. It includes support of video and picture display, but does not offer the possibility of continuous playback. Such an option may be possible in the future when XML software such SMIL will be more advanced. However, real-time display of linked information is functional, so the absence of continuous playback is not so much a limitation. The direct use of a web browser (not a web server, distant or local) does not raise major issues of speed efficiency, outside a long time for the initial loading of information. On the other hand, it allows fast development and good reliability as the data generated is automatically validated thanks to the use of XML. Several format are already proposed, such as text format (Clan like, see Figure 1) and partition format (Elan like, see figure 2). Other formats are under study and our goal is to make it possible for each one to set up the system so that it meets every one's specific



needs.

Figure 2: Partition presentation of data

4. Future improvements and goals

There are a lot of possible improvements for the VICLO platform. First, it is still in its developmental, prototype phase and new technical problems may arise. Inclusion of other widely used software such as Transcriber, Anvil and Exmaralda should be undertaken. It is clear that not all specific features from all tools will be maintained at 100% in our data model. In addition, the use of tools outside the systems prevents us from guarantying full integrity of the data, although it is always possible to record changes and go back to previous versions. Second, the development of TEI extensions calls for the creation of a TEI SIG for Oral language which was one of the conclusions of the CatCod conference (Orleans, France, December 2008). Third, the major goal of VICLO is not to create a model of repository, format and tools for corpus data, but to generate new scientific results thanks to the use of new software and approaches. To this effect, one of the goals in the near future is to help researchers use their data efficiently due to better

visualisation software and improved data manipulation and mining software. For example, data could be displayed in a huge variety of formats (including utterance based formats, turn based formats, etc.). Another interesting feature is interface with data manipulation software, such as spreadsheet software (OpenCalc), lexicometric software (Lexico, Le Trameur), statistical software (R) and natural language processing tools (NLTK).

Finally, it should be stressed that the goal of VICLO is not to limit the standards and formats used by the research community but on the contrary to be opened to the rich features offered by multiplatform applications. In this sense, although VICLO is not part of the CLARIN Project (www.clarin.eu), it could perfectly fit into this project, especially during the future construction phase.

For the same reason, there is no actual plan to limit or to delve into the semantics of the transcriptions and annotations. Although we admit that a good interoperability is impossible without such common semantic grounds, we think that the semantic levels should be controlled by the existing tools such as Clan, Elan, Praat, etc. and that semantic compatibility should, at least at the beginning of our project, be assured by the user themselves. Also, our plan is not to create a new multipurpose annotation standard or format, as proposed by Bird and Liberman (2000). The TEI format is, right now, rich enough to allow coding for all the formats we have been working with. Our problem is not into creating a more powerful system, but rather in dealing with the limited features of each application (each application having it's own limits and their own strengths) so that these limits will not impede the strengths of the other applications. In this sense, having a more powerful descriptive tool is not necessary at this moment.

5. Acknowledgements

The COLAJE project is funded the ANR (France), which includes three supporting laboratories, MoDyCo CNRS-Paris Nanterre University, Prismes-Sorbonne Nouvelle Paris 3 University, and LILT CNRS-Lille 3 University.

6. References

Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. Speech Commun., 33(1-2), 23-60.

Dublin Core: http://dublincore.org/documents/dcmi-terms/

ELAN: Language Archiving Technology http://www.lat-mpi.eu/tools/elan/

ISLE IMDI: www.mpi.nl/IMDI/

MacWhinney, B. (1991). The CHILDES project - Computational tools for analyzing talk. Hillsdale, NJ: Lawrence Erlbaum Associates.

OLAC: http://www.language-archives.org

Paul Boersma & David Weenink (2009): Praat: doing phonetics by computer (Version 5.1.05) [Computer program]. Retrieved May 1, 2009, from http://www.praat.org/

PHON: http://phon.ling.mun.ca/phontrac/wiki/

TalkBank: http://talkbank.org

TEI: http://www.tei-c.org

The LDOS-PerAff-1 Corpus of Face Video Clips with Affective and Personality Metadata

Marko Tkalčič, Jurij Tasič and Andrej Košir

University of Ljubljana Faculty of electrical engineering
Tržaška 25, Ljubljana, Slovenia
marko.tkalcic@fe.uni-lj.si, jurij.tasic@fe.uni-lj.si, andrej.kosir@fe.uni-lj.si

Abstract

This paper presents a corpus of video clips of users responding to emotional stimuli. The corpus is unique for two reasons. First, the emotions are annotated in the valence-arousal-dominance space instead of the usual coarse basic emotions. Second, the subjects are annotated with their personality parameters which offers a new ground for further investigations on personality and emotions. The corpus has been compiled for the needs of our research on recommender system. The paper provides information about the corpus acquisition procedure, corpus basic statistics with few examples and a short description of the research work where the corpus has been used in the past.

1. Introduction

Emotion recognition is an important area of affective computing. Several modalities, or combinations of them, can be used to detect user's emotions in human computer interaction and other areas. Some of these modalities use the video stream of the observed person as source (face expression, posture detection, hand gestures detection). Methods for the detection of emotions are being developed (Zeng et al., 2009). These methods use different corpora of video clips annotated with emotional metadata (Kanade et al., 2000; Pantic et al., 2005). Generally, such corpora are missing contextual information that might be useful in the development of emotion detection algorithms.

We present a corpus of video clips of users responding to emotion elicitation visual stimuli with additional annotations. These annotations include the induced emotive state, end users' personality parameters, demographic data and explicit ratings of the visual stimuli. The corpus presented here has been compiled for the needs of our research work of affective and personality based user modeling in recommender systems (Tkalčič et al., 2009a; Tkalčič et al., 2009b; Tkalčič et al., 2010).

The presented corpus has some unique properties that are not present in other corpora to the best of the authors' knowledge. First, the emotions are annotated in the valence-arousal-dominance (VAD) space which is less coarse then the usual basic emotions space. Second, the subjects are annotated with their personality in the big-five personality space. We believe that the presented corpus can be further exploited. As personality plays an important role in our emotional mechanisms (John and Srivastava, 1999) we believe it is an important contextual information in a corpus used for the development and validation of emotion detection methods. For example, extroverted people are believed to be more expressive than introverted people so the success rate of automatic methods for emotion detection could vary depending on user's personality differences. We encourage the usage of the proposed corpus for research work on emotion detection.

The remaining of the paper includes the acquisition procedure, the description of the corpus with some examples, the details about the distribution of the corpus, experiments where the corpus has been used, the discussion with future work guidelines and few concluding sentences.

2. Acquisition procedure

We used the emotion induction (sometimes referred also as emotion elicitation) experimental approach (Bradley and Lang, 2007). We used a subset of 70 images from the IAPS set of standardized visual stimuli to induce emotive responses in the subjects (Lang et al., 2005; Bradley and Lang, 2007). We assessed the personality of the subjects using the IPIP 50 questionnaire (Goldberg et al., 2006; http://ipip.ori.org/newQform50b5.htm, last accessed February 2010). We had 52 participants involved in the acquisition procedure.

2.1. Requirements for the data corpus

Our research work on recommender systems required a dataset of usage history of users interacting with an image consumption application. We needed the following data fields

- image: represented the item to be consumed as well as the emotion elicitation visual stimulus
- the induced emotive state as a triple in the VAD space
- personality parameters of the big five personality model as described by (Goldberg, 1998)
- explicit ratings on a five level Likert scale
- video clips of the users' responses to the visual stimuli

The general quality measure for a sample of data is that it should be a good representation of the larger set of data it attempts to represent. In our case there were two critical dimensions where the corpus should reflect the wider dataset:

- (i) a wide spectrum of emotive states
- (ii) a wide spectrum of participants' personalities.

We addressed both criteria by carefully selecting the visual stimuli and by analyzing the subjects' personalities.

2.2. Emotion induction procedure

We used a subset of the IAPS database of images for inducing emotive responses. The subset was chosen carefully to cover equally the value-arousal plane (see Fig. 1). The users' goal was to select images for their computer's wallpaper. The users were instructed to watch each image from the subset and give an explicit rating from 1 to 5. During their interaction with the application, that was build in Matlab (see Fig. 2 for a snapshot of the application's GUI), the users were recorded with a web cam. The web cam was positioned on top of the monitor so the participants' gaze was below the camera position. The basic unit of the dataset was thus a video clip of a user responding to a single visual stimulus from the IAPS database.

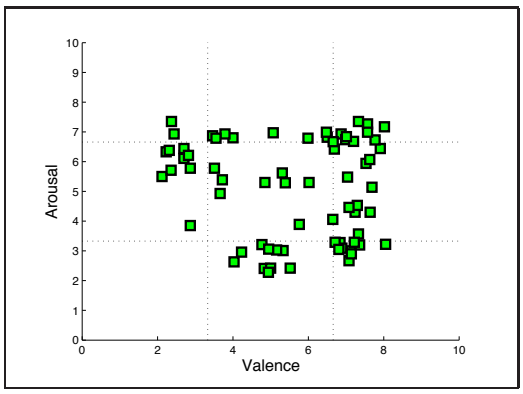


Figure 1: Distribution of the induced emotions of the visual stimuli in the valence-arousal plane.

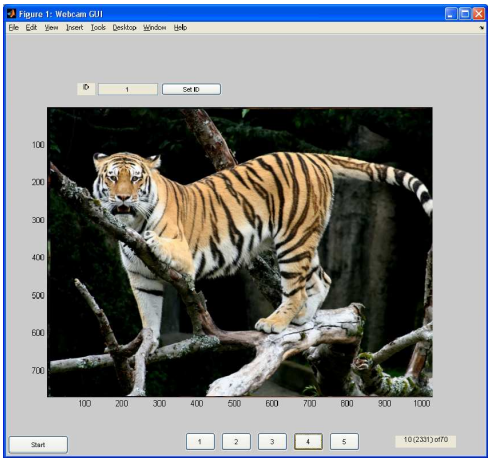


Figure 2: User interface of the application of the emotion induction experiment.

2.3. Personality and its assessment

Personality accounts for the individual differences in their emotional, interpersonal, experiential, attitudinal and motivational styles (John and Srivastava, 1999). The E factor tells the degree of engagement with the external world (in case of high values) or the lack of it (low values). The sub-factors of E are friendliness, gregariousness, assertiveness, activity level, excitement-seeking and cheerfulness. Extrovert people (high score on the E factor) tend to react with enthusiasm and often have positive emotions while introverted people (low score on the E factor) tend to be quiet, low-key and disengaged in social interactions. The N factor refers to the tendency of experiencing negative feelings. People with high N values are emotionally reactive. They tend to respond emotionally to relatively neutral stimuli. They are often in a bad mood which strongly affects their thinking and decision making. Low N scorers are calm, emotionally stable and free from persistent bad mood. The sub-factors are anxiety, anger, depression, self-consciousness, immoderation and vulnerability. The distinction between imaginative, creative people and down-to-earth, conventional people is described by the O factor. High O scorers are typically individualistic, non conforming and are very aware of their feelings. They can easily think in abstraction. People with low O values tend to have common interests. They prefer simple and straightforward thinking over complex, ambiguous and subtle. The sub-factors are imagination, artistic interest, emotionality, adventurousness, intellect and liberalism. The C factor concerns the way in which we control, regulate and direct our impulses. People with high C values tend to be prudent while those with low values tend to be impulsive. The sub-factors are self-efficacy, orderliness, dutifulness, achievement-striving, self-discipline and cautiousness. The sub-domains of the A factor are trust, morality, altruism, cooperation, modesty and sympathy. The A factor reflects individual differences in concern with cooperation and social harmony.

We used the IPIP 50 questionnaire to assess the personality big five factors of the participants. The questionnaire consisted of 50 items, 10 per each big five personality factor.

2.4. Participants

We had 52 students of a secondary school who participated in the experiment. The average age was 18.3 years (SD = 0.56). There were 15 males and 37 females.

3. Corpus

The corpus consists of 3640 video clips of 52 participants responding to 70 different visual stimuli. The video files are segmented by user and by visual stimulus. The annotations are stored in text based files. The participants cover a heterogeneous area in the space of the big five factors.

3.1. File formats

The video files are encoded with the xvid codec and have the resolution of 320 x 240 pixels at the frame rate of 15fps. The filename notation is `USERID_ITEMID.AVI` . For example, the filename `41_1534.avi` represents the video clip of the user with the id 41 responding to the visual stimulus image with the id 1534 from the IAPS database. There is a total of 258 Mbytes of video clips with the duration of 5 hours and 55 minutes.

The annotations are stored in three different formats: excel, semicolon delimited text and ARFF weka format. The filenames are `LDOS-PerAff-1.xls`, `LDOS-PerAff-1.csv` and `LDOS-PerAff-1.arff`.

3.2. Corpus properties

Each video clip is annotated with a line in the annotation file. Tab. 1 shows an extract from the annotations files. The annotations files have the following columns: user_id, image_id, image_tag, genre, watching_time, wt_mean, valence_mean, valence_stdev, arousal_mean, arousal_stdev, dominance_mean, dominance_stdev, big5_1, big5_2, big5_3, big5_4, big5_5 gender, age, explicit_rating, binary_rating. Descriptively, they contain the ID of the observed user, the ID of the observed image used to induce emotion, the recorded watching time, the average watching time of the observed item, the image tag, the image genre, the first two statistical moments of the valence, arousal and dominance values of the induced emotive state of the observed image, the big five parameters of the observed user, the gender, age and explicit rating given by the user to the observed image.

The participants showed heterogeneity in the distribution of the big five personality parameters as can be seen in Fig. 3. Unfortunately it is not possible to assess whether the personality distribution of the participants reflects a wider group of users because such norms do not exist (Goldberg et al., 2006).

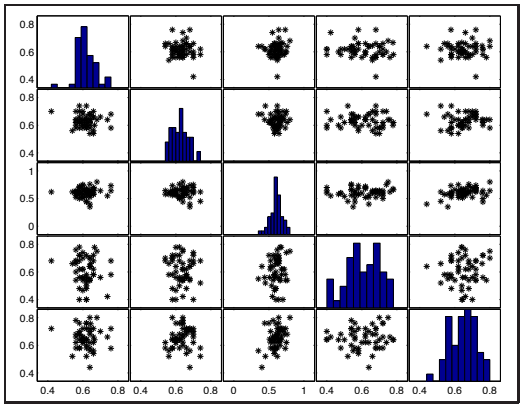


Figure 3: Distribution of the participants' personalities as pairs of big five factors scatter plots and histograms of single factors.

We chose the subset of visual stimuli from the IAPS database randomly with the constraint to cover equally a wide range of induced emotional values in the valence-arousal plane. The distribution of the induced emotions of the visual stimuli used is depicted in Fig. 1. This distribution is more suitable for the development of emotion detection methods that cover the whole valence-arousal plane (all the basic emotions).

3.3. Examples

Tab. 1 shows an extract from the annotation files. Fig. 4 and Fig. 5 shows snapshots from the video files of two subjects. The subject in Fig. 4 offers very little dynamics in face expressivity while the subject in Fig. 5 shows extreme face dynamics.

4. Experimental design and results

In this section we present three experiments that we are conducting on the described corpus. Two of these experiments have been completed (the affective content based recommender and the personality based collaborative recommender) while we are still working on the emotion detection experiment.

4.1. Affective content based recommender system

Part of the presented corpus was used to develop and evaluate an affective user and item modeling approach in a content based recommender (CBR) system. The CBR scenario included users interacting with an application where they could observe images (items) and giving explicit ratings on a five scale Likert scale to each item. The user' goal was to rate images as candidates for the personal computer's wallpaper. We modeled the items with the first two statistical moments of the induced emotive response over a large set of users which was provided by the IAPS database. This represented the affective item model. We also modeled the items with the genre and watching time tags which represented the generic item modeling approach. In the evaluation we compared the performances of both models, the affective and generic, in a CBR scenario. We used several machine learning algorithms to predict the users' ratings. We evaluated the performance of the CBR using confusion matrix based measures precision, recall and F measure. We performed a statistical analysis which showed that the usage of the affective modeling approach yielded significantly better CBR performance than the usage of generic modeling. Parts of the results were published in (Tkalčič et al., 2009a) and (Tkalčič et al., 2010).

4.2. Personality based collaborative recommender system

Another part of the presented corpus was used in an experiment where we evaluated a novel personality based user similarity measure for collabrative filtering. A well known drawback of collaborative filtering methods is the new user problem. It occurs when a new user joins the system and the system has little or no knowledge on the user's preferences. As the user similarity measures rely on explicit ratings given by the users, when there are little ratings available, the algorithm for finding similar users tend to give bad choices. Consecutively, the predicted ratings have low correlation with real ratings. In order to alleviate the new user problem we introduced an initial questionnaire to assess the personality of each user. We chose the IPIP questionnaire with 50 questions (ipi, 2009) which yielded five parameters for each user in the big-five personality model space (Johnson, 2009). We constructed a user similarity measure as an Eclidian distance in the big five personality space. We compared the proposed user similarity measure with a generic rating based user similarity measure in a collaborative recommender system scenario. Again, we calculated the confusion matrix based scalar measures precision,

user_id	image_id	image_tag	genre	watching_time	wt_mean	valence_mean	valence_stdev	arousal_mean	arousal_stdev	dominance_mean	dominance_stdev	big5_1	big5_2	big5_3	big5_4	big5_5	gender	age	explicit_rating
10	6910	Bomber	action	2614	5307	5.31	2.28	5.62	2.46	5.1	2.46	3.2	2.7	2.9	3.5	2.9	0	18	4
10	9331	Assault	action	2240	3214	2.03	1.35	6.04	2.35	0	0	3.2	2.7	2.9	3.5	2.9	0	18	3
10	7052	HairDryer	still	2223	2665	4.93	0.81	2.75	1.8	5.82	1.93	3.2	2.7	2.9	3.5	2.9	0	18	1
10	1280	Rat	animal	1906	3093	3.66	1.75	4.93	2.01	5.05	2.2	3.2	2.7	2.9	3.5	2.9	0	18	1
10	2394	Medicalworker	people	1943	2993	5.76	1.74	3.89	2.26	0	0	3.2	2.7	2.9	3.5	2.9	0	18	3

Table 1: Extract from the annotation table of the corpus.

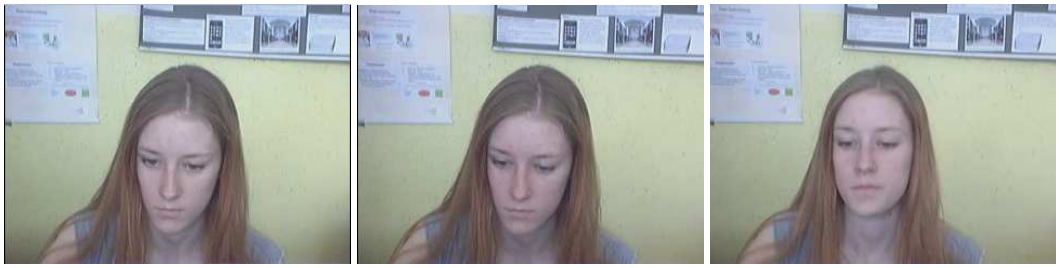


Figure 4: Snapshots of a subject with very low dynamics of face expressivity.

Figure 5: Snapshots of a subject with high dynamics of face expressivity.

recall and F measure. Statistical analysis showed that the proposed personality based user similarity measure yielded significantly better results than the rating based user similarity measure which makes it more suitable not only to alleviate the new user problem but also to use when the new user phase dies away. The results of this research were published in (Tkalčič et al., 2009b).

4.3. Emotion detection from video clips

Part of our ongoing research work is the detection of emotion from face videoclips. Our goal is to develop a method for detecting emotions in users with two novel properties: (i) the inclusion of personality parameters as features and (ii) detection in the valence-arousal-dominance (VAD) space (instead of the coarse space of basic emotions). We intend to use the emotion detection method for automatic tagging of users and items for affective profile building.

The current design of the experiment includes the extraction of the users' faces using the Viola-Jones algorithm (Viola and Jones, 2004) and fine registration using the active appearance model (AAM) tracker developed by (Saragih and Gocke, 2009). We plan to extract low level features using Gabor wavelets and applying the Hidden Markov Model (HMM) to reduce the number of variable features (due to the variable length of video clips) to a fixed number. We will combine personality parameters and low level features in a machine learning (ML) algorithm. We will evaluate several ML algorithms.

5. Copyright and privacy issues

5.1. Distribution of the corpus

Our exclusive interest is the promotion of research. Thus the distribution of the corpus is free for use in academic, not-for-profit research at recognized educational institutions. The researcher who wishes to receive the corpus must fill in and submit the EULA (end user license agreement) form available at `http://slavnik.fe.uni-lj.si/PerAff/` according to the instructions on the form. Within 30 days after receiving your form we will send you a username and password for downloading the corpus. The researchers are expected not to publish or distribute the material in any form.

6. Discussion and future work

The corpus presented in this paper was created to support our work on affective and personality based recommender systems (Tkalčič et al., 2009a; Tkalčič et al., 2009b; Tkalčič et al., 2010). We evaluated the impact that emotive parameters and personality has on user ratings. We proposed a novel approach for modeling users with VAD emotive parameters in the context of an image recommender system. Furthermore we developed a novel user similarity measure based on the big-five personality traits. The personality based user similarity measure yielded statistically equivalent performance of a collaborative recommender system than the usual rating based user similarity measure while withstanding the new user problem.

Beside affective and personality user modeling for recommender systems the corpus can be used for a variety of research work we are unable to foresee right now. Anyway we provide a list of interesting topics where the presented corpus could be used

1. a comparison of efficiency of different expression detection algorithms
2. development of a non intrusive personality detection method based on face video clips of induced emotions
3. relation between expression detection and personality

The latter is, in the authors' opinion, one of the most interesting issues. How can user's personality help emotion detection techniques? For example, knowing that a subject has a certain personality profile could help the emotion detection algorithm to fine tune its internal parameters and thus achieve greater accuracy. The subjects in Fig. 5 and Fig. 4 have different personalities. They differ also in the dynamics of face expressions. It would be interesting to see whether there is a combination of personality parameters that is correlated with face expression dynamics. This would surely be helpful for the emotion detection algorithm to adapt its internal thresholds.
The authors have already undertaken on the emotion detection with the inclusion of personality, as described in Sec. 4.3.. At the time of writing of this paper we have not get so far to be able to include any results.

7. Conclusion

The affective computing community can make good use of the proposed corpus with its personality annotations which makes it unique. It features almost six hours of 52 subjects face video recordings with 70 different induced emotive states. The video sequences are annotated with big-five personality parameters of subjects and metadata related to the items.
In this paper we provide the background for the construction of the corpus. We describe the acquisition procedure and give basic dataset statistics. We also give the description of the recommender systems application where the presented corpus was validated. Instruction for accessing the corpus are given. We suggest a list of topics where the specifics of the LDOS-PerAff-1 corpus could be used.

Acknowledgement

The authors would like to thank the teachers and students from the Gimnazija Poljane school in Ljubljana for their participation. We are also thankful to Matevž Kunaver, Tomaž Požrl and other members of the LDOS group who have helped in the implementation of the acquisition procedure. This work has been partially funded by the European Commission within the 6th framework of the IST under grant number FP6-27312. All statements in this work reflect the personal ideas and opinions of the authors and not necessarily the opinions of the European Commission.

8. References

Margaret M. Bradley and Lang, 2007. *The International Affective Picture System (IAPS) in the Study of Emotion and Attention*, chapter 2. Series in Affective Science. Oxford University Press, 198 Madison Avenue.

Lewis R. Goldberg, John A. Johnson, Herbert W. Eber, Robert Hogan, Michael C. Ashton, C. Robert Cloninger, and Harrison G. Gough. 2006. The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40:84–96.

Lewis R. Goldberg. 1998. What is beyond the big five? *Journal of Personality*, 66(4):495–524.

http://ipip.ori.org/newQform50b5.htm. last accessed: February 2010. Sample ipip big5 questionnaire. web site.

2009. International personality item pool: A scientific collaboratory for the development of advanced measures of personality traits and other individual differences. `http://ipip.ori.org/`, June.

Oliver P. John and Sanjay Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. In Lawrence A. Pervin and Oliver P. John, editors, *Handbook of Personality: Theory and Research*, pages 102–138. Guilford Press, New York, second edition.

John A. Johnson. 2009. Descriptions used in ipip-neo narrative report. `http://www.personal.psu.edu/faculty/j/5/j5j/IPIPNEOdescriptions.html`, June.

T. Kanade, J.F. Cohn, and Y. Tian. 2000. Comprehensive database for facial expression analysis. In *Proceedings of the fourth IEEE International conference on Automatic Face and Gesture Recognition*, page 46.

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

M. Pantic, M.F. Valstar, R. Rademaker, and L. Maat. 2005. Web-based database for facial expression analysis. In *Proc. IEEE Int'l Conf. Multmedia and Expo (ICME'05), Amsterdam, The Nederlands*, July.

J. Saragih and R. Gocke. 2009. Learning AAM fitting through simulation. *Pattern Recognition*, 42(11):2628–2636.

M. Tkalčič, J. Tasič, and A. Košir. 2009a. Emotive and Personality Parameters in Multimedia Recommender Systems. *ACII 2009 Affective Computing and Intelligent Interaction*, page 33.

Marko Tkalčič, Matevž Kunaver, Jurij Tasič, and Andrej Košir. 2009b. Personality based user similarity measure for a collaborative recommender system. In C. Peter, E. Crane, L. Axelrod, H. Agius, S. Afzal, and M. Balaam, editors, *Proceedings of the 5th Workshop on Emotion in Human-Computer Interaction - Real world challenges*, pages 30–37. Fraunhofer Verlag, September.

Marko Tkalčič, Jurij Tasič, and Andrej Košir. 2010. The need for affective metadata in content-based recommender systems for images. In Mark Maybury, editor, *Multimedia Information Extraction*, chapter 19. MIT Press.

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

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

Corpus-based analysis of users' emotional strategies to convince virtual characters

Magalie Ochs[1] and Rui Prada[2]

[1]CNRS-LTCI, Télécom ParisTech, Paris, France, magalie.ochs@telecom-paristech.fr
[2]INESC-ID and IST-Technical University of Lisbon, Portugal, rui.prada@gaips.inesc-id.pt

Abstract

Most of the research in Affective Computing focuses on the emotions *felt* by the users during human-machine interaction. In this paper, we explore the users' emotions, not necessarily felt, but deliberately expressed to try to achieve a specific goal: to convince a virtual character. A video game in the virtual environment Second Life has been developed to collect data on the users' emotional strategies to convince in human-machine negotiation. The analysis of the resulting corpus highlights different emotional strategies of the users depending on their personality.

1. Introduction

During interpersonal interaction, people may express emotions different from their felt emotions to follow some sociocultural norms or to try to achieve specific goals (Ekman and Friesen, 1975). Recent research in Human and Social Sciences has highlighted the *emotional gaming* phenomena (Andrade and Ho, 2008). To *game emotion* means to strategically modify the expression of a current felt emotion to try to influence someone else's behavior. For instance, people sometimes use specific expressions of emotion to convince someone else in a negotiation (Andrade and Ho, 2008; VanKleef, 2007; Liand and Roloff, 2006).

In this paper, we focus on the *emotional gaming* of users for the purpose of influencing virtual characters' opinion during a negotiation in a virtual environment. To collect information on the users' emotional strategies in human-machine negotiation, we have developed a video game in the virtual environment *Second Life*. Several users have interacted with different emotional virtual characters with the goal to convince them using emotions. The analysis of the resulting corpus has enabled us to highlight users' emotional strategies during negotiation and some correlations with their personality.

The paper is structured as follows. In the next section, existing works in Human and Social Sciences related to the emotional strategies during interpersonal negotiation are presented. In Section 3, we introduce the video game, called *the virtual negotiation place*, developed in Second Life to collect information on users' emotional strategies to convince virtual characters. The method to collect the corpus and the results of the analysis are presented and discussed Section 4. We conclude Section 5.

2. Theoretical Background: Emotional Strategies in Interpersonal Negotiation

Recent research in Human and Social Sciences has shown that one's expression of emotion may influence other's decision in a negotiation process (Andrade and Ho, 2008; VanKleef, 2007; Liand and Roloff, 2006). During interpersonal interaction, people sometimes game emotions (i.e. express emotions not necessarily felt) to try to change the course of a negotiation. Several studies have highlighted that both *happy* and *anger* emotion expression have beneficial effects on negotiation. On one hand, people may strategically choose to express *happiness* and suppress sadness and anger to others to elicit liking from them (Clark et al., 1996). Indeed, as shown in (Knutson, 1996), people are perceived likable when they express joy. In the context of a negotiation, positive emotion can signal cooperativeness and trustworthiness and may elicit cooperation, trust, and concession from others (Liand and Roloff, 2006). On the other hand, *anger* expression of emotion impresses the other party as aggressive and competitive. People who express anger are perceived as more dominant but less likable (Knutson, 1996). But, people with low power are strongly affected by their opponent's emotions (anger emotion), whereas those with high power are unaffected (Vankleef et al., 2006). Finally, people with low power concede more to an angry persuader that to a happy one (Vankleef et al., 2006). Moreover, as highlighted in (Liand and Roloff, 2006), a congruence between what the receiver expects and what the persuader expresses can lead to a successful negotiation. People generally expect that their emotional expression evokes complementary and similar emotional responses in others (Keltner and Kring, 1998; Morris and Keltner, 2000). For instance, anger should evoke fear or guilt (low-power emotions (Liand and Roloff, 2006)), distress should evoke empathy, etc.

Based one the research in Human and Social Sciences presented, we consider three emotional strategies during a negotiation: (1) the expression of joy, (2) the expression of anger, and (3) the expression of congruent emotion. In order to identify how users use emotions during a negotiation with virtual characters in a virtual environment, we have developed a *virtual negotiation space* with different emotional virtual characters in the environment *Second Life*.

3. Virtual Negotiation Space

The virtual negotiation space has been created in the 3D on line virtual world *Second Life* (Linden-Lab, 2003). Second Life is a free networked multi-user world-like environment in which users are represented as avatars that can communicate with others and interact with objects in the virtual environment. The virtual negotiation space has been created as a game environment. The user, through his avatar

Figure 1: Virtual characters in the virtual negotiation space. From the left to the right: the emotional persuasive virtual character, the random emotional virtual character, and the non-emotional virtual character.

in Second Life, has to convince virtual characters to give him the boxes they have. At the beginning of the game, the user meet a first virtual character which explains the goal of the game. To convince the virtual characters, the user has to chat, through the chat channel, with the characters. The virtual character explicitly advices the user to use emotion to try to persuade the other virtual characters. To express emotion, the user directly types the type of the emotion at the end of the sentence. He can use three types of emotion: *anger*, *sadness*, and *happy*.

In order to analyze the users' emotional strategies depending on the interaction with different emotional virtual characters, the virtual negotiation space has been populated with three different virtual characters: (1) an *emotional persuasive* virtual character, (2) a *random emotional* virtual character, and (3) a *non-emotional* virtual character (Figure 2). Each of these virtual characters has been programmed to response to user's messages. They are endowed with a sentences database containing predefined responses depending on the character' opinion (for instance, "I do not want to give this box" or "I should not keep this box"). The virtual characters are not able to analyze the user's message but response automatically given their opinion. The virtual characters do not use specific arguments to convince. The *random emotional* virtual character expresses randomly the emotion of anger, sadness or joy. The *non-emotional* character does not express any emotion during the interaction with the user. Only the emotional persuasive virtual character takes into account the user's emotions to choose the emotion to express. Indeed, the *emotional persuasive* virtual character is endowed with a model of emotional strategies based on the research in Human and Social Sciences presented above (Section 2). The emotional persuasive character expresses anger in response to the user's expression of sadness. If the user expresses anger, the virtual character displays an empathic message (such as "You look sad, I'm sorry for you") with an expression of sadness. In response to the user's expression of joy or neutral emotion, the virtual character expresses joy (Ochs and Predinger, 2010). In Second Life, the emotional virtual characters that we have developed express emotions in two ways: their facial expressions and an object attached to their chest called *EmoHeart*. *EmoHeart* appears when the virtual characters express emotions, and its texture depends on the type of the expressed emotion (Figure 2). *EmoHeart* (Neviarouskaya et al., 2009) provides an additional channel for visualizing emotions in a vivid way while the facial expression of emotion in Second Life may be elusive. To express empathy, the emotional persuasive virtual character uses additionally predefined sentences, such as "You look sad, I'm sorry for you ".

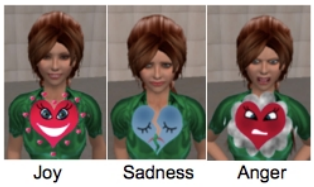


Figure 2: Examples of virtual characters' facial expressions and EmoHeart textures

4. Collection of the users' emotional strategies corpus

4.1. Method

Participants. We have asked 17 subjects (three women, fourteen men) to play the game. The subjects' ages ranged between 21 and 30 years old. They have in average few experience using Second Life (in average 2 on a Likert scale of 7 points), some experience with computer games (in average 5 on a Likert scale of 7 points), and with virtual environments in general (in average 4 on a Likert scale of 7 points). The participants were mainly French (12 on 17) with 2 Brazilian, 2 African, and 1 Malaysian.

Procedure. The participants have played the game in our research institute. We have presented the study to the user as a game test. Given the link between emotions and personality (Revelle and Scherer, 2009; Salovey et al., 2000), we aimed at analyzing the impact of personality on the user's emotional strategies. Consequently, at the beginning of the test, we asked the participants to fill a personality test to assess the big five personality factors (extroversion, agreeableness, conscientiousness, emotional stability, and intellect) and their emotional intelligence (Goldberg et al., 2006). Then, each participant has interacted with the three virtual characters presented above (emotional persuasive, random emotional, and non-emotional): all participants interact with all character types.

The goal was to convince the characters to give their box to the user. For each interaction with a virtual character, four dialog turns occur. A dialog turn corresponds to an exchange of messages from the user to the virtual character and from the virtual character to the user. After the four dialog turns, the virtual character stops the conversation. For each participant, we have predefined if, at the end, the virtual characters agree to give the box or not. The order of the characters in which the user interacts with and the final opinions of virtual characters (pros or cons) have been counterbalanced to avoid an effect of the order of the characters or of the final opinions of them on the results. The subjects have received 1000 Japanese yen at the end of the test for their participation.

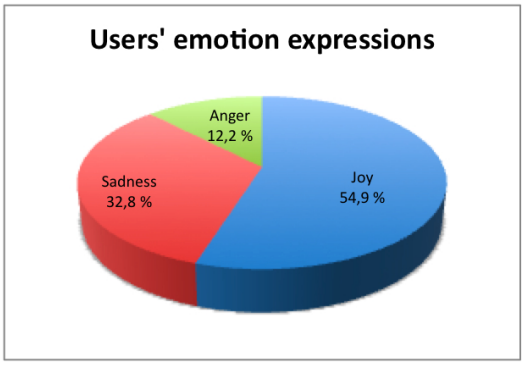


Figure 3: Global repartition of the users' emotion expressions

	Joy	*Sadness*	*Anger*
extroversion	+	-	n.s.
agreeableness	n.s.	n.s.	n.s.
conscientiousness	+	n.s.	-
emotional stability	+	-	n.s.
intellect	+	-	n.s.
emotional intelligence	++	-	- -

Table 1: Correlation between the personality factors and the expression of emotion

Corpus description. To analyze the emotions used by the participants to try to convince the virtual characters, we have recorded in a file the sentences exchanged between the users and the virtual characters. During the recording, the sentences have been automatically annotated by the type of the speaker (user id, persuasive virtual character, random virtual character, or non-emotional virtual character) and the emotion expressed by the speaker (anger, sadness, joy, or neutral). The resulting corpus is composed of 418 annotated sentences (209 of the users and 209 of the virtual agents). It is composed of 172 expressions of joy (112 of the users and 60 of the agents), 98 expressions of sadness (67 of the users and 31 of the agents), 75 expressions of anger (25 of the users and 50 of the agents), and of 73 neutral expressions (5 of the users and 68 of the agents).
Next section presents a deeper analysis and discussion of these results.

4.2. Results

Emotion types expressed by the users during a negotiation with virtual characters. First of all, we have analyzed the types of emotion that the participants have used during their interactions with the virtual characters. In general, as illustrated Figure 3, they have mainly used joy, sometimes sadness, and few times anger. The expressed emotion depends on the participants' personality and emotional intelligence. To highlight the relation between participants' emotion expressions and their personality, we have computed the Pearson correlation coefficient. In the paper, we report the medium ($0.3 < c < 5$) and large ($0.5 \leq c < 1$) correlations. A positive (respectively negative) medium correlation is noted "+" (respectively "-"). A large correlation is noted "++" for positive correlation and "–" for negative correlation (n.s. means no significant correlation) (Table 1).
In Table 1, The medium correlation for the extroversion personality factor shows that the more the user is extroverted the more he uses joy emotion and the less he uses sadness to convince the virtual characters. Similarly, the emotional stability and intellect personality factor of user are positively correlated to the expression of joy and negatively correlated to the expression of sadness. The consciousness factor seems correlated to the expression of joy and anger, leading a user with a high value of consciousness to express more joy and less anger to convince. The agreeableness factor seems to not have an impact on the types of expressed emotion. Finally, the results reveal a large correlation between emotional intelligence and emotion expressions: the more the user is emotionally intelligent, the more he expresses joy and the less he displays anger and sadness.

Users' emotional reactions to virtual characters' emotion expression in a negotiation. We also have analyzed the types of emotions used by the participants in response to virtual character's emotions. In order to identify significant differences, we have performed a T-test to compare the frequency the participants used each emotion type in response to the emotions displayed by the virtual characters.
Concerning the emotion expressed by the participants in response to the *joy expression* of a virtual character, the results reveal significant differences: the participants have expressed significantly more joy than sadness ($p < 0.05$) and anger ($p < 0.01$). Large positive correlations appear between the expression of joy in response to joy and the extroversion personality factor, the emotional stability factor and the emotional intelligence of the user ($c \geq 0.5$): the more the user is extroverted, emotional stable or emotionally intelligent, the more he expresses joy in response to a characters' expression of joy.
Concerning the emotional response to a characters' *expression of sadness*, the participants have expressed significantly more joy than anger ($p < 0.05$), and sadness than anger ($p < 0.01$). However, no significant difference between joy and sadness appears. The expression of joy and sadness in response to a virtual expression of sadness is correlated with the intellect personality factor (large, $c \geq 0.5$): the more the user has a high value for the intellect personality factor, the more he expresses joy and the less he expresses sadness in response to sadness. Medium positive correlation appears between joy expression and extroversion personality factor whereas a medium negative correlation exists between anger expression and user's emotional intelligence.
Concerning the emotion expressed by the participants in response virtual characters' *anger expression*, no significant difference appears between the types of emotion used. However, the results reveal correlations with personality factors: the more an user is extroverted or emotional stable the less he expresses sadness in response to anger (respectively large and medium negative correlation), and the more the user is emotional intelligent the less he displays anger

in response to anger (medium negative correlation).
Finally, the results of the T-test reveal an effect of the emotion type expressed by the characters on the emotional response of the user. In response to joy, participants have used significantly more joy than in response to sadness ($p < 0.01$) or anger ($p < 0.05$); and in response to anger, the participants have significantly preferred to use anger than in response to sadness ($p < 0.05$).

Users' emotional strategies to convince virtual characters. We have analyzed the sequences of emotional expressions of the participants in order to try to highlight their emotional strategy to convince. A T-test has been performed to evaluate the effect of the number of dialog turns occurred on the type of emotion expressed. We have particularly analyzed the influence of participants' emotional intelligence on their strategy considering that the more the user is emotional intelligent, the better is the strategy. Concerning the emotion of *joy*, participants have significantly more expressed this emotion at the first dialog turn, than at the second, third or fourth one ($p < 0.01$). The users' emotional intelligence seems also to have an influence: the more the user is emotionally intelligent the more he expresses joy at the second dialog turn (medium correlation) and at the third one (large correlation). Concerning *sadness*, participants have significantly less expressed sadness at the first dialog turn than at the second ($p < 0.01$), third ($p < 0.05$) or fourth one ($p < 0.01$). Medium negative correlations appear with the emotional intelligence: the more participants are emotional intelligent the less they have expressed sadness at the first or second dialog turns. The *anger* expression have been significantly more used at the end of the dialog (third or fourth dialog turn) than at the beginning (first or second dialog turn) ($p < 0.05$). Moreover, the more the participants are emotional intelligent, the less they expressed anger at the end of the dialog (large correlation for the third dialog turn and medium one for the fourth dialog turn). A medium positive correlation appears for the first dialog turn, showing that the more the participants are emotionally intelligent the more they have displayed anger at the first dialog turn.
In general, the T-test reveals that the participants have significantly more expressed joy at the first dialog turn than sadness or anger ($p < 0.01$). At the second dialog turn, joy is significantly more displayed than anger ($p < 0.01$), and, similarly, sadness is significantly used more than anger ($p < 0.01$).

4.3. Discussion

First of all, the analysis of the corpus shows that the main emotional strategy used by the users (and particularly the emotional intelligent users) to try to convince a virtual character is the expression of *joy*. However, the users' emotional strategy depends on the emotion expressed by the virtual character. The users prefer displaying an emotion of joy in response to the virtual character's expression of joy. In response to sadness, depending on their personality (and more particularly the intellect factor), the user displays either sadness or joy. When the virtual character expresses anger, the user tends to display more anger than when the virtual character expresses an emotion of sadness. However, maybe because of the few number of anger expressions, we cannot conclude that the user displays more anger in response to the virtual character's anger expression than sadness or joy. The personality of the user (his extroversion, emotional stability and emotional intelligence) may provide information on the user's emotional strategy facing virtual character's anger expression.
The analysis of the sequence of expressed emotions reveals that the users generally start with the expression of a positive emotion (joy) at the beginning of the negotiation and express negative emotion (sadness or anger) at the end. On the contrary, it seems that emotional intelligent users prefer to display negative emotion (and in particular anger) at the beginning and to finish by expressing a positive emotion.
In conclusion, the corpus-based analysis of users' emotional strategies during a negotiation with virtual characters highlights the types of emotion used to convince depending on the users' personality factors[1] The next step is to use these results to model the emotional strategies of virtual characters with different personalities.

Acknowledgment. This work was supported by a postdoctoral fellowship of the Japan Society for the Promotion of Science (JSPS).

5. References

B.E Andrade and T. Ho. 2008. Gaming emotions. Technical report, Experimental Social Science Laboratory.

M.S. Clark, S.p. Pataki, and V.H. Carver, 1996. *Knowledge structures in close relationships*, chapter Some toughts and findings on self-representation of emotions in relationships. Lawrence Erlbaum.

Paul Ekman and W.V. Friesen. 1975. *Unmasking the face. A guide to recognizing emotions from facial clues*. Prentice Hall Trade.

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

Dacher Keltner and Ann M. Kring. 1998. Emotion, social function, and psychopathology. *Review of General Psychology*.

B. Knutson. 1996. Facial expressions of emotion influence interpersonal trait inferences. *Journal of Non-verbal Behavior*, 20(165-182).

S. Liand and M.E Roloff, 2006. *From Communication to Presence: Cognition, Emotions and Culture towards the Ultimate Communicative Experience*, chapter Strategic Emotion in Negotiation: Cognition, Emotion, and Culture. IOS Press.

Linden-Lab. 2003. Second life. www.secondlife.com.

Michael W. Morris and Dacher Keltner, 2000. *Research in Organizational Behavior*, chapter How emotions work:

[1] It's difficult to compare the obtained results with the literature on the emotions used during interpersonal negotiation (Section 2) which focuses on the types of emotion without considering the personality of the individual. Moreover, our context, very specific since the users have to type emotions, differs from natural interpersonal interactions.

the social functions of emotional expression in negotiations. Elsevier.

A. Neviarouskaya, H. Prendinger, and Ishizuka M. 2009. Emoheart: Automation of expressive communication of emotions in second life. In *International Conference on Online Communities and Social Computing*.

M. Ochs and H. Predinger. 2010. A virtual character's emotional persuasiveness. In *International Conference on Kansei Engineering and Emotion Research (KEER)*.

W. Revelle and K. R. Scherer, 2009. *The Oxford companion to emotion and the affective sciences*, chapter Personality and emotion. Oxford UniversityPress, Oxford.

P. Salovey, B. Bedell, J. Detweiler, and J. Mayer. 2000. Current directions in emotional intelligence research. In M Lewis and J.M. Haviland-Jones, editors, *Handbook of Emotions*, pages 504–520. Guilford Press.

G.A. Vankleef, C. DeDreu, D. Pietroni, and A. Manstead. 2006. Power and emotion in negotiation: Power moderates the interpersonal effects of anger and happiness on concession making. *European Journal of Social Psychology*.

G.A. VanKleef. 2007. Emotion in conflict and negotiation: Introducing the emotions as social information (easi) model. In *IACM*.

SALEM - Statistical AnaLysis of Elan files in Matlab

Marc Hanheide[1], Manja Lohse[2], Angelika Dierker[2]

1 University of Birmingham, School of Computer Science, B15 2TT, UK
2 Bielefeld University, Technical Faculty, Universitätsstraße 25, 33615 Bielefeld, Germany
E-mail: m.hanheide@cs.bham.ac.uk, mlohse@techfak.uni-bielefeld.de, adierker@techfak.uni-bielefeld.de

Abstract

This document proposes SALEM (Statistical AnaLysis of Elan files in Matlab) as a toolbox for the statistical analysis of data from human-machine interaction that are annotated in Elan. The authors show how SALEM allows to analyze annotations quantitatively in an effective manner. The paper introduces the position of SALEM in the data processing chain, its functionalities, and how it contributes to the evaluation process.

1. Introduction

Interaction studies with humans and intelligent systems are gaining more and more attention with systems exhibiting more advanced abilities. This is true in various areas such as cognitive robotics, assistance, interaction analysis and others more. In the interaction of systems and humans, a corpus must not only encode the behavior of the participating humans, but also the behavior of the intelligent system to facilitate a manifold analysis of more complex interaction scenarios. Both need to be brought together in order to allow a real cross-disciplinary analysis. Merging and temporally aligning the system-focused annotations with manual annotations that describe the humans' behavior results in a comprehensive and rich representation of the interaction situation. With the help of these so-called systemic corpora we can (i) learn and study patterns of deviation or failures in interaction and (ii) identify correlations between system and human behaviors.

Figure 1 shows an exemplary process of creating a systemic corpus comprising the recording of data, integrating, and finally analyzing it. The left side displays the data source. It includes audio and video data, as well as system log files from the intelligent system. In our work with multiple intelligent systems we use logging frameworks such as LOG4J[1] or LOG4CXX5[2] to generate the system log files. These frameworks natively support time-stamped events and can also set the type (usually corresponding to a logging level) and the emitter (e.g., called "logger" in log4j). The content of these events is either unstructured text or structured text (if the developers agree on a coding scheme in advance). It would also be possible to include other data sources. However, since our aim is to display these logging events as (time-stamped) annotations in Elan, all sources need to have an extent in time. In the next step of the processing chain all data are aligned and synchronized in time. After transformation and synchronization all recorded data can be displayed and manually revised,

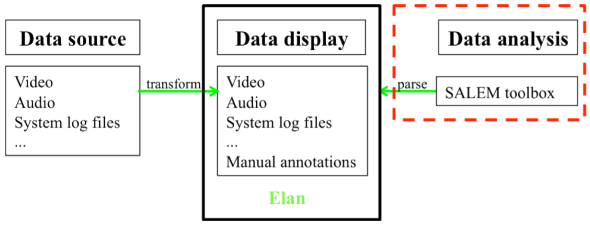


Figure 1: Data processing chain

e.g., in Elan[3] which furthermore allows to add manual annotations (Figure 1 [data display]). Alternatively, other XML-based annotation tools with more or less similar abilities could be used (e.g., Anvil[4], Interact[5]). With respect to manual annotations one important step is to check their correctness for later analysis (e.g., if a coding scheme is used, only codes that are defined in the scheme are allowed; the annotations do not contain misspellings). Once all the data are represented in one format (here in the format of the annotation tool) they can be analyzed automatically. This is where the SALEM (Statistical AnaLysis of Elan files in Matlab) toolbox comes in. In the following, we will introduce the basic idea behind the usage of SALEM, its functionalities, and advantages.

2. Basic idea of SALEM

After conducting interaction studies, every researcher encounters the question of how to analyze data efficiently and effectively. This question is strongly influenced by what and how data are recorded. As has been introduced in Figure 1, we propose to collect the data in annotation tools like Elan, because with the help of these tools, multiple data sources can be integrated with manual annotations. The annotations are structured in multiple layers, so-called 'tiers'. For example, in human-robot interaction (HRI) one tier may contain the speech of the human that has been manually transcribed based on the video. Another tier may contain what the

[1] http://logging.apache.org/log4j/
[2] http://logging.apache.org/log4cxx/
[3] http://www.lat-mpi.eu/tools/elan/
[4] http://www.anvil-software.de/
[5] http://www.mangold-international.com/en/products/interact.html

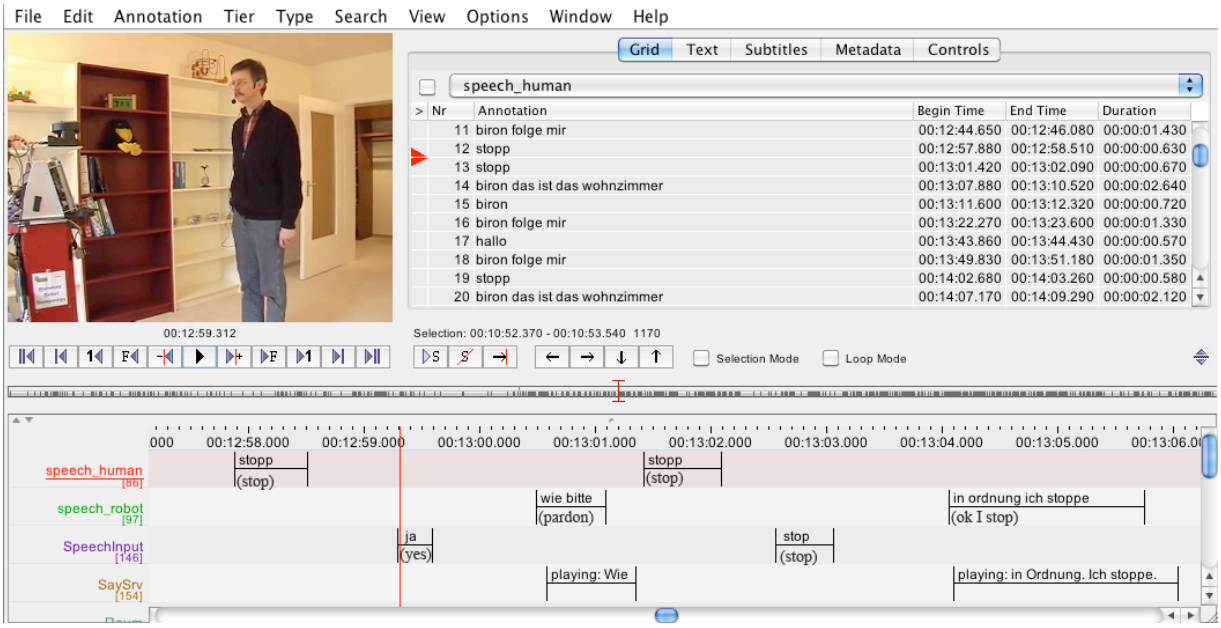


Figure 2: Example of Elan display

robot understood which has been extracted from the log files. This example shows that manual annotation might have to be integrated with data from automatic log files. The integration can easily be achieved with an XML format as used by Elan and other annotation tools such as Anvil. However, annotation tools usually only include limited functionalities for quantitative analysis. Thus, to analyze the data, one can go through the files manually or, alternatively, use the tools' export options to acquire text or XML-based files. Thereafter, these files have to be imported in another software for analysis (for example, Matlab[6], SPSS[7]). This approach is rather laborious because whenever a change is made to the annotation file, it has to be exported and imported to the analysis software again. To work around this problem, the SALEM toolbox parses Elan files directly and offers advanced and in-depth statistical analyses with Matlab. Thus, it closes the cycle of importing automatic log files into the annotation tool, importing the corresponding video/audio streams into the annotation tool, annotating manually, and analyzing the data in an efficient way. In this process, one main advantage of the SALEM toolbox is that it allows comparing annotations of different modalities, structural features of the interaction, or whatever has been annotated in the tiers. For example, the video can be used to manually annotate human speech which can then be compared to the speech that the robot understood because they are represented in one file that can be analyzed, in our case using Matlab (see Figure 2). Analyzing the data with the SALEM toolbox does not only alleviate the analysis process, but also increases the consistency of the analysis. This is because all evaluations are conducted using the statistical functions that are predefined by the toolbox, for example, a T-test will always be calculated in the same way based on the same formula.

3. Functionalities of SALEM

Our proposed automatic statistical analysis of annotation files consists of a number of routines to compute statistics on the temporal distribution of specific annotations, their correlation to one another, and their comparison with regard to duration and dedicated values. A core concept of SALEM to allow a rich analysis is "slicing". It has been introduced to facilitate temporal correlation between annotations of different tiers. The idea is that we can automatically select subsets of annotations for the computation of the statistics based on synchrony or overlaps. Therefore, one tier is chosen as the master tier and the whole set of annotations is sliced according to the existence of annotations in that master tier. We also support slicing based on specific values of the annotation in that master tier. Slicing is based on the analysis of overlaps of the master tier annotations with annotations in all other tiers as illustrated in Figure 3. If, for instance, one tier codes all time intervals in which the robot was speaking, slicing allows to compute statistics only for those annotations that overlap with this 'speech_robot' annotation.

Built upon this general slicing concept, SALEM to date has the following functionalities which were developed based on requirements that arose during the analysis of two corpora of HRI data (Lohse, 2010; Lohse, submitted). Of course these functionalities are currently limited, however, new ones can be integrated in the toolbox if needed.

[6] http://www.mathworks.com/products/matlab/
[7] http://www.spss.com/software/statistics/

Parsing, displaying structure of parsed files, and slicing:

- parsing single Elan files or a set of Elan files at once (which allows for the analysis of the data of single users, groups of users, groups of trials that belong to certain conditions, and all users of an experiment)
- plot all annotations in the tiers
- slice the files with respect to time (specifying one or more beginnings and endings of timeslots)
- slice all annotations of a single tier (for example, if the file is sliced on the basis of the tier 'speech_human', then in all other tiers only the annotations that are overlapping with instances of human speech are taken into account)
- slice the files with respect to one or more values of the annotations in a single tier (for example, slice all annotations of eye gaze that have the value "1" which means that the user is looking at the robot)
- examine one specific annotation in a tier (for example, the 12th annotation in the tier 'gaze direction')

Analyzing:

- descriptive statistics for all data of the parsed files or the slices (for each tier):
  - count of annotations (number of occurrences)
  - minimum duration of the annotations (in seconds)
  - maximum duration of the annotations (in seconds)
  - mean duration of all annotations (in seconds)
  - median of the durations (in seconds)
  - overall duration of all annotations (in seconds)
  - variance and standard deviation of the duration of all annotations
  - beginning of first annotation (in seconds)
  - end of last annotation (in seconds)
  - overall duration of all annotations as a percentage of the time between the beginning of the first annotation and the end of the last annotation
- the descriptive statistics for slices additionally include for all tiers
  - count and percentage of time that the annotations in a tier overlap with the reference tier for four types of overlap: (a) the annotation extends the annotation in the reference tier (begins before the annotation and ends after the annotation in the reference tier), (b) the annotation is included in the annotation in the reference tier (begins after the annotation in the reference tier and ends before the annotation in the reference tier), (c) the annotation begins before the annotation in the reference tier begins and ends before it ends, (d) the annotation begins after the begin of the annotation in the reference tier and ends after the end of the annotation in the reference tier (see Figure 3)
- statistics for the annotated values in a certain tier:
  - duration of all annotations for all values
  - descriptive statistics for all values (see descriptive statistics for all tiers)
  - T-tests for the duration of the annotations and the duration of the overlap
  - predecessor transition matrix for all values in the tier (percentages with which all values preceded all other values; for example, annotations with the value '1' are preceded by annotations with the value '2' in 62% of the cases)
  - successor transition matrix for all values in the tier (percentages with which all values succeeded all other values; for example, annotations with the value '1' are succeeded by annotations with the value '2' in 36% of the cases)

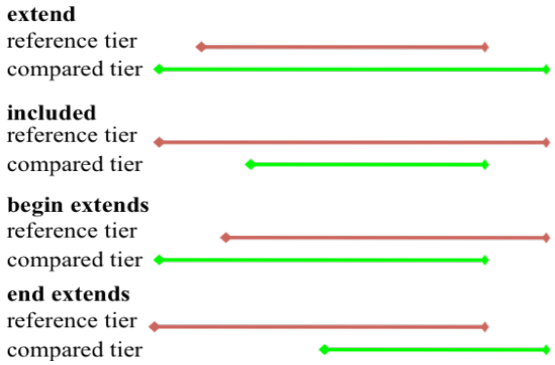


Figure 3: Overlap types

4. Conclusion

In this paper, we presented a toolbox for the statistical analysis of systemic, and thus inherently multi-modal, corpora. It supports a quantitative analysis of interaction corpora by merging automatically generated annotations with manual annotations and subsequently analyzing them using functions provided by the SALEM toolbox. We successfully applied the proposed approach in several user trials with interactive robots. The universal concept of slicing introduced by SALEM allows to define subsets of annotations to compare different cases or situations and consequently facilitates the analysis in a context-aware manner. SALEM also eases the work flow by processing the annotations directly without requiring prior export which is often error-prone. Moreover, the toolbox is usable for people with little knowledge about Matlab. It saves them from learning single commands for the statistical functions and speeds up the analysis. Since the same functions are used for all data, the toolbox supports a consistent analysis. SALEM is extendable if more statistical functions are needed or if it shall be used with other annotations tools which are based on XML.

The SALEM toolbox is available free of charge from: http://aiweb.techfak.uni-bielefeld.de/content/salem-statistical-analysis-elan-annotation-using-matlab

5. Acknowledgments

This work has been supported by the Collaborative Research Centre 673 Alignment in Communication, founded by the German Research Foundation (DFG), and the European Community's Seventh Framework Programme [FP7/2007- 2013] under grant agreement No. 215181, CogX.

6. References

Lohse, M. (2010). Social, functional, and problem-related tasks in HRI - a comparative analysis of body orientation and gaze. *2nd AISB workshop on New Frontiers in Human-Robot Interaction*. Leicester, UK.

Lohse, M. (submitted). Investigating the influence of situations and expectations on user behavior - empirical analyses in human-robot interaction. Doctoral Thesis. Bielefeld University, Technical Faculty, Germany.

Collecting and Annotating Conversational Eye-Gaze Data

Kristiina Jokinen [1]
Doshisha University and University of Helsinki
kristiina.jokinen@helsinki.fi

Seiichi Yamamoto, Masafumi Nishida
Doshisha University, Japan
seyamamo@mail.doshisha.ac.jp
mnishida@mail.doshisha.ac.jp

ABSTRACT
This paper describes our work on the collection and annotation of conversational eye-gaze data. The corpus contains 28 multiparty conversations among participants who are freely chatting with each other on everyday interesting issues, and the corpus includes real-time eye-tracking data of one of the participants. The eye-tracked person is also videoed from the front with a separate camera, and this allows synchronizing the two views of their gaze behaviour. Interactions are balanced so that comparison of communicative behaviour along familiarity, gender, and language between the participants in the different groups is also possible. At the moment, six of the dialogues are annotated, and preliminary analyses on the annotated data are also described.

INTRODUCTION
In about 20 years, technology to track and analyse eye-gaze has evolved to the level where experiments are easy (and cheaper) to arrange, and the resulting data is fairly robust and accurate. Although eye-trackers have long been used as diagnostic tools in medical and cognitive psychology laboratories, their technical evolution has only quite recently allowed a wider range of use for eye-trackers from interface devices to interaction monitoring. Eye-trackers have also become more popular in communication research, and many studies use eye-trackers and include eye-tracking data in order to build models for human-machine interaction, see e.g. Ishii and Nakano (2008) among others. Currently eye-tracker technology is actively developed, and an overview of the technology and the current trends can be found e.g. in Jacob and Karn (2003) and Duchowski (2003).

Our aim in collecting eye-gaze data has been two-fold. First, given the possibility of new technology to collect data on human communicative behaviour, it seems only natural to do so: this allows us to explore the limits of the technology, and simultaneously also to widen the scope of research techniques and methods. Second, since eye-gaze plays a crucial role in fluent communication (Argyle and Cook, 1976), it is important to be able to study the speaker's gaze behaviour and focus of attention using empirical data also from the signal-level point of view. This provides objective data for interaction studies, and thus complements subjective interpretations of the important communicative events. As we wanted to focus on naturally occurring human communicative activity, we collected conversational corpus which contains data with two distinctive characteristics:

1. It contains three-party conversations instead of two-party dialogues,
2. Conversations exhibit free-flowing associative activity rather than task related activity.

The first aspect has consequences on the participants' roles and mutual relations, and also on interaction management. In two-party dialogues, the two participants share the space between them and can thus both observe the other and be aware of the other observing them. The participants take turns between the two of them, and both are responsible for dialogue management and for the success of the interaction (given their respective roles). In multi-party dialogues, however, the conversation takes place within a context which is spatially more complex: interaction space is not only larger but it also contains areas which are not directly shared by all the interlocutors: two participants can converse between them, while the others remain observers. This makes interaction management more complex (Healey and Battersby, 2009): some of the interlocutors may be aware of a particular aspect of interaction while the others are unaware of the same aspect. Consequently, also the models for mutual knowledge and grounding of information are more complex, and it can be assumed that the eye-gaze functions as an important signal in interaction management.

Concerning free-flowing vs. task-based dialogues, it was considered useful that communication models be based on data that aims at engaging participants in an activity which is as unconstraint as possible so as to avoid any other additional requirements or cognitive demands on the interlocutors' behaviour than what was already imposed by the setting as such. Moreover, it can be assumed that any likely differences in the participants' eye-gaze behaviour would come out most clearly if they could have a friendly chat instead of being required to focus their attention on some (artificial) task. It must be emphasised that this kind of free chatting is not claimed to be more natural or intuitive than interaction in task-related situations; quite

[1] The collection and analysis was carried out when the author was the NICT Visiting Scholar at Doshisha University, Japan.

contrary, both are considered natural in what comes to the typical behaviour of the interlocutors, differing only in the activity type and the interlocutors' social roles. However, as the activity that the speakers are engaged in is known to affect the participants' behaviour, it can be assumed that free chatting among peers is one of the most neutral types of interaction in this respect: the constraints arise mainly from the communicative needs as such. The results could then be extended to, and compared with situations where more constraints are imposed due to the task (e.g. participants focus their gaze on an object in the shared environment or have very distinctive social roles such as teacher-student, leader-follower).

In this paper we describe our work on the collection and annotation of conversational eye-gaze data. The paper is structured as follows. We first describe the eye-tracker used in the data collection, and then go into details of the data collection setup. We present the collected data, and briefly refer to the annotation work and the results of preliminary analyses. We finish with discussions of future work.

EYE-TRACKER

Usually eye-trackers are desk-mounted video-based systems which have the camera on the desk besides or under the screen, and which can show the focus of the user's gaze on screen in real time. An additional computer is needed to do the image processing for the eye, although the most advanced systems today integrate the optics for videoing the eyes and the computational processing with the screen. It is also possible to have head-mounted eye-trackers which free the user from sitting in front of a screen.

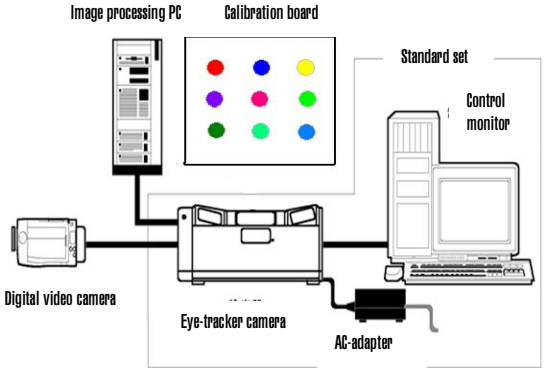


Figure 1: The eye-tracking setup.

In our experiments, we used the NAC EMR-AT VOXER eye-tracker. The system setup is shown in Figure 1, and as is seen in Figure 1 the standard set consists of the eye-tracker camera and a control monitor, and also includes image processing PC, the calibration board, and the digital video camera that records the situation.

The actual camera for tracking the eye is shown in Figure 2. The system operates by sending an infrared light beam to the eyes and measuring the angle of reflections from the cornea by the two camera devices. Figure 2 shows the light emitting device in the middle and the two cameras that record the reflection of the light from the right and left eye are on its left and right, respectively.

Figure 2: NAC EMR-AT VOXER Eye-tracker.

Figure 3 depicts the optimal measures needed for calculating the optics of light reflection. The angle between the table and the eye-tracker camera should be about 28 degrees, and the user's eyes should be at 517mm distance from the eye-tracker camera and about 40 cm higher than the table top on which the camera is placed. In this position, there is about 20 cm margin to move the head forward and backward without disturbing the tracking accuracy.

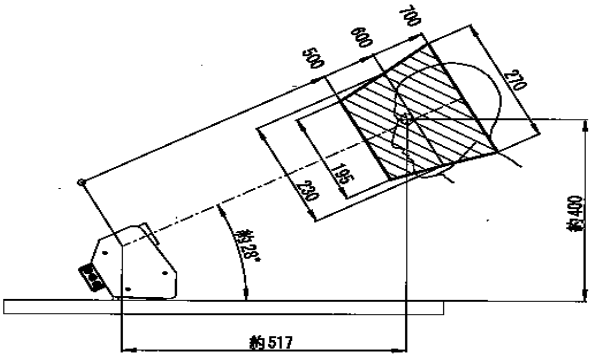


Figure 3: Measures for optical calculations. The numbers refer to millimetres. (Taken from the eye-tracker manual).

The tracker also uses the shape of the eye to locate the gaze. The shape is determined in the setup process and calculated with respect to the white and black pattern recognized in the picture of the eyes. Figure 4 shows the control panel for determining eye shapes. The pictures of both eyes by the two cameras are shown so that the person's right eye is on the left and the left eye on the right. The patterns can be modified by changing the relative amount of black and white parts in them, and thus reach the best fit with the person's overall eye-shape.

The overall setup and operation of the eye-tracker is managed via the control panel that shows on-line view of the operation of the system. Figure 5 presents a snapshot of

the control panel with a person's eye being tracked during calibration. It shows the camera view of where the user is looking at (empty table and the calibration board at the back), the two frontal face views by the two camera devices measuring the light reflection, and a close-up view of the eye which is being tracked (in this case the person's right eye). The reflection points of the beam are also shown as cross points of a horizontal and a vertical line. In the live camera view of the eye in the top right corner of the panel there are two intersecting lines and they correspond to reflections from the pupil and from the cornea of the eye.

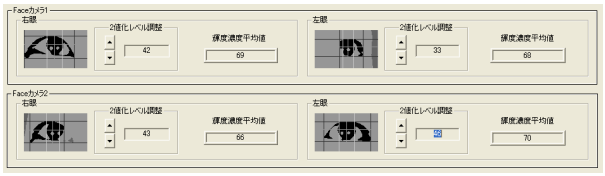


Figure 4: Determining the shape of the eye.

During the calibration phase, the user looks at nine points on the calibration board (cf. Fig. 1), while the system measures the eye's position, shape, and the reflexions of the infrared light. Also head movement is taken into account to compensate movements up, down, and sideways. The sampling rate of the eye-tracker is 60 Hz.

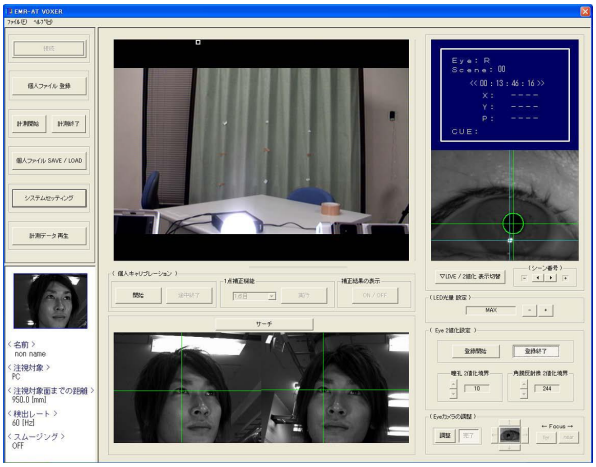


Figure 5: View of the control panel.

DATA COLLECTION

The data was collected during the first author's visit to the Doshisha University in Kyoto, Japan. The collection setup is shown in Figure 6. Three participants sit in a triangle formation, and one of them (the eye-tracked speaker, ES) has their eye movements recorded by the eye-tracker (the rightmost person in Figure 6). The two other participants, the left-hand speaker (LS) and the right-hand speaker (RS) are videotaped by the camera (in the foreground of Figure 6) as reference points to where ES's gaze is focused on.

In typical eye-tracking experiments ES looks at the stimulus on a computer screen which is still and stable with respect to the ES. However, our setup differs from this in that we now have a group of three partners who converse with each other and none of them is necessarily still. Quite contrary, instead of a static computer screen, LS and RS are dynamic targets as they can move their head and whole body backward and forward and also tilt their head or bend their body sideways. The optics of the eye-tracker is rather robust, and allows ES head movements of about 20cm depth and 30 degree angle. Thus the participants can have fairly natural body and head movements that are typical for free conversations. However, if the ES head movements are very large, eye-gaze data cannot be captured. Also, if the participants laugh, as they often do during chatting, the eye-tracker loses some data, since ES's eyes become small and the relevant eye-patterns cannot be found. Of course, when ES blinks, no eye-gaze can be recorded either.

A special problem was caused by the special lamp in front of ES. This was needed in order to help the eye-tracker to distinguish the eye-shape, but it was generally considered rather bad as it shone directly at ES and distracted ES from the two other partners. A better general lighting or spot lights from the ceiling were considered as solutions to the problem in the future collections. In general, the current set up was fine. However, for some participants it did not help in the recognition of their eye pattern: their eyes could not be tracked, apparently due to difficulties in distinguishing between the white and the colour, or to their eye region being too small to be captured by the camera.

Figure 6: Data collection setup

The data collection took place in two phases. In the first phase, which was also a pilot case, six conversations were collected, while in the second phase, 22 conversations were collected with different participants. All conversations are among three interlocutors, and are 10 min long. Before conversation recordings, the participants were told the purpose of the study, and they also signed a consent form that allows video-recordings to be used in research and shown publicly.

As already mentioned, the participants were not involved in any particular task, but were instructed to learn more about their partners and discuss issues that they were interested in. The unfamiliar conversations (see below) were directed especially to elicit conversational information about first encounters and situations where people get introduced to

each other. Consequently, conversations are lively chatting on topics that range from hobbies and weekend plans to studies and travelling. The participants seem to behave naturally despite of being videotaped and the general laboratory conditions. Especially, the restriction of the head movement due to the eye-tracker's technical limitations did not seem to have a big effect on the naturalness of the dialogues. As an explanation to this, it was also suggested that Japanese people in general move little during conversations and thus the eye-tracker constraint did not have a noticeable effect.

In the first, pilot phase, the participants were six Japanese students from the laboratory (5 male, 1 female). They were familiar with each other although did not necessarily know each other very well. In order to get a mixture of participants with minimum contact with each other in the experimental setting, the participants rotated among themselves so that the eye-tracked person was always a new participant in each triad. The rotation is schematically shown in Figure 7.

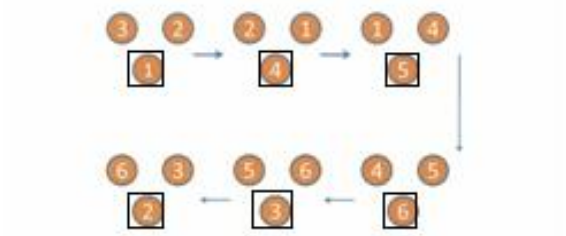


Figure 7: Rotation of the participants in the pilot phase

For the second collection we recruited participants from among the students in other laboratories and friends outside the university. There were 18 different participants with their age range in early 20's, and they were grouped so that we got 14 conversational triads with unfamiliar participants. All subjects were given book cards as tokens of our gratitude

Figure 8: View of the eye-tracked person

The instructions were the same as in the pilot case, but it was emphasised that the participants should introduce themselves properly first.

The collection setup was also similar to the one in the first phase, except that we also had a second camera that recorded the eye-tracked person's face and head movements, see a snapshot in Figure 8. We thus have two types of data concerning the ES behaviour: what the persons themselves focus their gaze on (through the eye-tracker), and how they are seen by the other partners (through the second camera). By synchronizing the two camera views, we will be able to compare the two types, "inside" and "outside" views, of the ES gaze behaviour.

DATA DESCRIPTION

We collected 14 unfamiliar conversations where the interlocutors did not know each other before. Four of the unfamiliar conversations are within female-only groups, four within male-only groups, and six within mixed gender groups (two male students and one female student).

We also collected four familiar conversations among the students of the laboratory, so that together with the six already collected ones the number of familiar conversations would be 10, and the gender balance would equal that of unfamiliar ones (four male-only, six mixed-gender groups). Moreover, four conversations among familiar participants speaking English were also recorded, in order to provide data for comparison between Japanese and English conversations.

Altogether the corpus contains 28 conversations (14 familiar and 14 unfamiliar conversations), balanced with gender, and including eye-tracker information on top of the video data. This amounts to about 280 minutes (4 hours 40mins) of natural conversations, see Table 1.

Type	Group	Number	Description
Familiar pilot	Mix	3	Familiar partners – pilot study
	Male	3	Familiar partners – pilot study
Familiar	Mix	3	Familiar partners
	Male	1	Familiar partners
Familiar English	Mix	4	Familiar English-speaking partners
Unfamiliar	Mix	6	Unfamiliar partners
	Male	4	Unfamiliar partners
	Female	4	Unfamiliar partners
Eye-tracked face		22	Videos of the eye-tracked person (from all conversations except Familiar-pilot)

Table 1 Statistics of the collected corpora.

An example of the video data is shown in the snapshot in Figure 9. The gaze-path shows that ES shifts focus from left to right, and after some fixations on the forehead and left eye of RS, gaze is focussed on the right eye of RS.

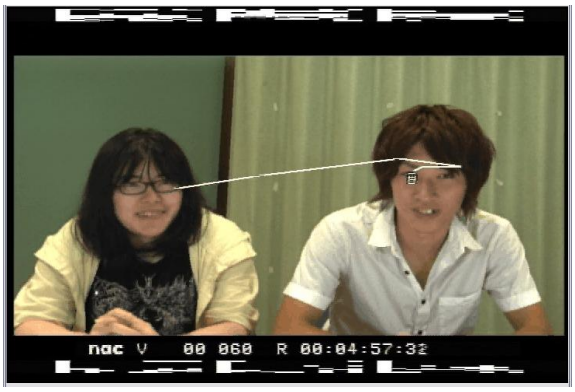


Figure 9. Sample video showing a gaze path from left to right.

ANNOTATION AND ANALYSIS

Five minute clips of each of the six familiar conversations collected in the first phase have also been annotated with dialogue acts, gaze, facial expressions, and turn-taking behaviour, and the annotated data has been used for experimental studies concerning the relation between gaze and turn management. Annotation was done according to the MUMIN annotation scheme (Allwood et al. 2007) which was modified to take the gaze information into account. In addition, the dialogue act annotation was included according to the guidelines developed in the AMI project (www.amiproject.org).

Annotation was done with the Anvil annotation tool (Kipp 2001) by three students who had basic understanding of the task and goals of the exercise, but no previous experience in annotation. The annotator agreement was measured by Cohen's kappa-coefficient, and reached the kappa value of 0.46. This corresponds to a moderate agreement.

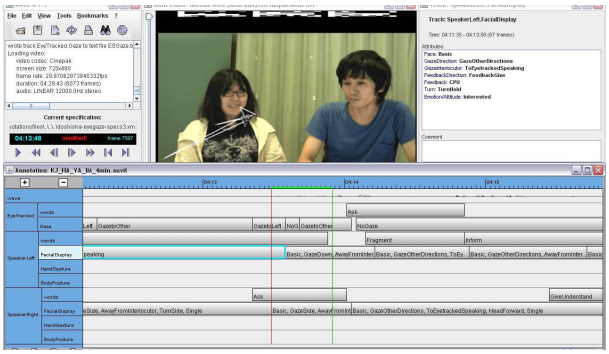


Figure 10: Anvil annotation board with a gaze-path on the speaker on the left.

A view of the annotation board is shown in Figure 10. It consists of separate groups for each of the three participants: eye-tracked speaker (ES), left-side speaker (LS), and right-side speaker (RS). LS and RS have tracks for dialogue acts (words), facial display, hand gesturing, and body posture, while ES has tracks only for dialogue acts and gaze behaviour, for the obvious reason that the video only records the ES' gaze path and voice (the first annotation did not include the frontal video view of ES). However, all annotation levels contain the same features and feature values for communicative functions dealing with feedback and turn management as well as for emotion of the speakers.

We have also conducted several experiments on the annotated data so as to study the relation between eye-gaze and turn-taking. The experiments and results have been reported in our papers (Jokinen et al, 2009; Jokinen et al. 2010). We confirmed earlier observations concerning eye-gaze and its use in interaction management to indicate if the speaker wants to continue talking or is willing to talk. Eye-gaze thus has an important role in smooth communication as it helps the interlocutors to manage their turns in a cooperative manner, and it allows effective interaction management without explicit spoken expressions.

However, we also noticed that in multiparty conversations, the turn management seem to be signalled with head turns rather than eye-gaze, although gaze is important as an initial signal of who could be next speaker. In multiparty conversations, head movement may function as a more visible signal of the speaker's focus of attention and willingness to take turn. In two-party dialogues, eye-gaze may be enough to signal the partner's intention to take the turn or to give the turn, but in multi-party dialogues, the participants may not share the context completely and the partner's focus of attention needs to be expressed in a more visible manner.

CONCLUSIONS

The collected conversational eye-gaze corpus is one of the few corpora integrating eye-tracker information and, to the best of our knowledge, the first systematic attempt to collect eye-gaze data in natural multi-party conversational setting. The corpus contains conversations among familiar and unfamiliar partners, and also gender balance is taken into consideration by female-only triads. This allows comparison of data along these lines. The conversations are conducted in Japanese and there are also four English conversations by native or near-native speakers, which can provide basis for intercultural comparison on dialogue and gaze behaviour.

The corpus still needs to be annotated and transcribed in a more detailed manner, and the annotation scheme is to be revised as well. However, it is considered a useful start for further activities on analysing eye-gaze and communicative behaviour in natural conversational situations, as well as on collecting and coding multimodal data that includes eye-gaze information. For instance task-based dialogues, as discussed above, would provide interesting extensions and

comparison points to the existing corpus. As for the analysis, we will pursue the two-level approach as advocated e.g. in Jokinen (2009): the data is studied from the point of view of speech and visual signals as well as from the point of view of human interpretation. The signal-level analysis provides empirical evidence of the events that have occurred, while the human annotation assigns meaning to those that are observed as communicatively important events. The study of such complex issues as those related to human communication requires that evidence is collected from different perspectives, and complementing the new multimodal technology with human analysis provides novel possibilities for this. Work is already going on concerning the relation between speech signal and the annotated data, and we can also extend this with techniques that recognize face and body movement. Such work will be useful considering the envisaged future applications that can sense and interact with their environment.

CORPUS AVAILABILITY

The corpus is accessible from the authors, and we invite interested colleagues and researchers to contact us for further collaboration and usage of the data for human-human and human-machine communication studies.

ACKNOWLEDGEMENTS

We would like to thank Kazuaki Harada, Ryosuke Imayoshi, and Shota Sasaki for their help with the experiments and data annotation, and our subjects for taking part in the experiments. The first author would also like to thank NiCT for the grant that allowed her to conduct research in Japan.

REFERENCES

1. Allwood, J., L. Cerrato, K. Jokinen, C. Navarretta and P.Paggio 2007. The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing Phenomena. In Martin, J.C., Paggio, P., Kuenlein, P., Stiefelhagen, R. and Pianesi, F. (Eds.) Multimodal Corpora for Modelling Human Multimodal Behaviour. Special issue of the International Journal of Language Resources and Evaluation, 41(3-4), 273-287.
2. Argyle, M., and M. Cook 1976. Gaze and Mutual Gaze. Cambridge University Press, Cambridge
3. Duchowski, A.T. 2003. Eye-tracking Methodology: Theory and Practice. Springer
4. Healey, P.G.T. and S. Battersby 2009. The Interactional Geometry of a Three-way Conversation.
5. Ishii, R. and Y. Nakano. 2008. Estimating User's Conversational Engagement Based on Gaze Behaviors. In H. Prendinger, J. Lester, and M. Ishizuka (Eds.): IVA 2008, LNAI 5208, pp. 200–207, 2008. Springer-Verlag Berlin Heidelberg.
6. Jacob, R.J.K. and K.S. Karn 2003. Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises (Section Commentary). In J. Hyona, R. Radach, and H. Deubel, (Eds.): *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, pp. 573-605, Elsevier Science, Amsterdam.
7. Jokinen, K. 2009. Gaze and Gesture Activity in Communication. In C. Stephanidis (Ed.): Universal Access in Human-Computer Interaction. Proceedings of the 5th International Conference of UAHCI, Held as Part of HCI International, San Diego, CA.
8. Jokinen, K., Nishida, M. and S. Yamamoto 2009. Eye-gaze Experiments for Conversation Monitoring. Proceedings of the IUCS'09 conference, ACM, Tokyo
9. Jokinen, K., Nishida, M. and S. Yamamoto 2010. On Eye-gaze and Turn-taking. Proceedings of the Workshop "Eye-gaze in Intelligent Human-Machine Interaction", International Conference on Intelligent User Interfaces. Hong Kong.
10. Kipp, M. 2001. Anvil – A Generic Annotation Tool for Multi-modal Dialogue. In: Eurospeech. pp. 1367–1370.

SensHome: Towards A Corpus for Everyday Activities in Smart Homes

Jochen Frey, Robert Neßelrath, Christian H. Schulz, Jan Alexandersson

German Research Center for Artificial Intelligence – DFKI GmbH
Stuhlsatzenhausweg 3, Saarbrücken, Germany
{⟨firstname⟩.⟨lastname⟩}@dfki.de

Abstract

We present our planned efforts within the SensHome project in building a corpus of daily activities in smart environments. The recordings consisting of measurable events as provided by the instrumentation along with video and audio recordings. SensHome foresees a three-step development where the instrumentation is verified in a dual-reality setting followed by recording in controlled environments. Finally, the recordings will be done under real circumstances. We extend previous work on activities as formulated by (Leontiev, 1978) and (Stahl, 2009) with the notion of partial orders of ontologically represented events denoted *episodes* that constitute fundamental building block for annotations.

1. Introduction

Worldwide, considerable efforts are invested in research and development for accessible and user-friendly technology for the sake of coping with the demographic change. In Europe, the European Union alone or possibly in combination with domestic initiatives, e.g., the AAL Joint Programme, fund projects that target different aspects on living or being as a person with special needs. A big part of these efforts include providing different users with intuitive and accessible user interfaces for interacting with appliances and services in the smart home.

The SensHome project, see SensHome.dfki.de, is an effort in creating an infrastructure and a methodology for recording, modeling, annotating and analysing activities in Smart Environments where we initially focus on Smart Homes.

SensHome is one of the projects emerged from the i2home project. Here, the main focus was to inject an ecosystem around the Universal Remote Console (URC) standard, see (Zimmermann and Vanderheiden, 2007; ISO, 2008; Rich, 2009). The URC technology provides an approach called *pluggable user interfaces* that allows for interfacing arbitrary networked appliances or services with personalized and perhaps most important accessible and adaptable user interfaces. Both projects are step stones toward the long-term vision of intelligent and pro-active smart home environments where a person can not only control the environment, the environment even pro-actively supports the person in his/her daily life. To this end, we are particularly interested in the recognition of irregularities.

2. The SensHome Project

SensHome aims at finding an optimal instrumentation of the Smart Home implementing a flexible architecture and infrastructure for recognizing and analyzing everyday activities and thus offering proactive help.

Following the GOAL development methodology (Stahl, 2009) we start with pure virtual modeling of the environment to ensure that, firstly, there is a sufficient instrumentation and, secondly, the SensHome corpus is suitable for recording, annotating and analyzing the relevant activities.

The infrastructure will then be installed in the following controlled environments: I) SensHome Smart Suitcase; ii) DFKI's intelligent kitchen; and iii) Bremen Ambient Assisted Living Lab (BAALL).

To build up a corpus with realistic everyday activities the SensHome System will in the final step be used to instrument a real flat.

This work has been carried out in the scope of the SensHome project funded by the Saarland Government. The opinions herein are those of the authors and not necessarily those of the funding agency.

3. Activity Modeling

To start with, we will annotate our corpus with two levels of annotations: *activities* and *episodes*. However, we consider additional annotations, such as gestures, emotions, spoken language etc.

Our modeling of activities emerges from the Activity Theory by (Leontiev, 1978). Activity Theory provides a hierarchical model for dividing human interactions into *activities*, *actions* and *operations*. A single activity consists of actions which can again be decomposed into operations. Activities are motivated by primary goals or human needs, whereas actions and operations are directed towards secondary goals to achieve these needs. The boundary between actions and operations is represented by a user's specific knowledge and experience which can change over time, that is, an action that requires concentration can be internalized by transforming it into an unconscious operation and externalized vice versa. Activity Theory can be seen as a holistic framework for thinking about human activity and modeling user interaction as it is expressed in the use of technology (Kaptelinin and Nardi, 2006).

Previously, activity modeling has been extended in subject, artifact and object. For instance, (Stahl, 2009) extends the work of (Leontiev, 1978) with location and time point thus answering the question where the activity took place etc.

In SensHome, we include a temporal modeling in the sense that we explicit the details of the activities and their parts by adding the notion of *episodes*. Episodes are partial orders of ontological entities ordered by temporal and locational changes. The ontological entities—concepts and/or instances—stem from an ontology which are relevant to some activity as provided by the smart environment. Con-

sequently, depending on the detail of the models, some sub entities of an episode can occur in parallel. Entities can be associated with a subject, an artifact, location as in (Stahl, 2009). In our model, we additionally can associate an entity with a time point or a time span (Allen, 1984) which might be absolute or just related, e.g. before/after/simultaneously. Examples of episodes are "Turn-around; Open-cupboard; take-a-cup; turn-around; go-to-location-*cooking-area*; ..." and "switch-on-water-tap; fill-glass; switch-off-water-tap". The latter example can, however additionally be annotated as "fill-glass-with-water". All of these entities are tuples built up from entities in on ontology or even other episodes. Currently, we are looking into the UbisWorld (Heckmann, 2006) ontology as a starting point for our modeling, see also `http://www.ubisworld.org/`

4. The SensHome Corpus

There are several possibilities to created and annotate the corpus. Among the current approaches, we are currently considering several options.

One possibility is to follow the approach used in the AMI/AMIDA projects: NTX `http://groups.inf.ed.ac.uk/nxt/`, (Kilgour and Carletta, 2006). There, meetings were recorded with cameras and microphones, the recordings were then annotated with different levels, such as transcriptions, dialogue acts, disfluencies etc. Drawback of NXT is that there is no standard tool for annotating but there is a set of previous tools, e.g. `http://www.amiproject.org`. Consequently, new tools will have to be developed.

There are a number of other alternatives: clearly, ANVIL has become an important tool for corpus creation and annotation. Other tools are ELAN, EXMARaLDA and MacVisSTA, see (Kipp, 2010) for a comparison of these tools.

5. The SensHome Technical Infrastructure

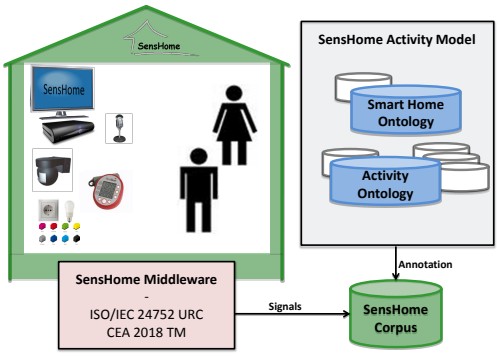


Figure 1: The SensHome Architecture. Activities in the smart home is mediated by the URC middleware through which not only video but also interactions with the environment is captured and piped to the corpus. The Corpus is the annotated with the SensHome Activity Model—a composition of different ontologies.

The SensHome architecture is built upon a new series of industry standards (ISO/IEC 24752 Universal Remote Console & ANSI/CEA 2018 Task Model Description) for interfacing networked appliances by means of a Universal Remote Console (URC) (Zimmermann and Vanderheiden, 2007) and for adding to the UIs, support for interaction, see (Rich, 2009). The implementation thereof is a middleware called universal control hub (UCH) that supports up-to-date prominent communication standards and allows for controlling multiple devices at the same time, see (Zimmermann and Vanderheiden, 2007). The combination of the UCH and the activity management module allows for the implementation of scenarios like leaving home: as a person leaves his house and locks the door, some running appliances should be turned off—TV, hood, oven the heating should depending on the situation be lowered—and the alarm system should be activated.

The SensHome Activity Model is a combination of a number of ontologies. Currently identified ontologies are

Smart Home Ontology In this ontology we basically model the appliances and signals included into the smart home, such as, Lamp, Stove, SwitchOne(lamp), ...

Activity Ontology Here, we model activities of daily living (ADL) and other higher order annotations

Dialogue Acts We will adapt the draft standard "ISO 24617-2 Semantic annotation framework, Part 2: Dialogue acts", e.g., (Bunt et al., 2010).

Temporal Ontology Activities are dependent on temporal information, e.g., (Allen, 1984).

Locational Ontology Modelling location will be based on an extension of the UBISWORLD ontology, see (Heckmann, 2006)

The UCH gateway-based architecture implementing the URC standard managing the communication between controllers and targets: a **Controller**, that is any device for rendering the abstract user interface, e.g., TV, touch screen or the smartphone presented in this paper; a **Target**, which is any networked device or service intended to be controlled or monitored, such as kitchen appliance, home entertainment or security devices; and, finally, a **Resource Server**, a global service for sharing user interfaces and other resources necessary for interacting with the targets. The benefit of this approach is that it is possible to deploy consistent and, particularly, accessible user interfaces which are tailored to particular users.

The URC technology has so far been envisioned as a middleware for interacting with smart environments as in the i2home[1] (Alexandersson, 2008) and VITAL[2] (Zinnikus et al., 2009) projects. Other running projects have started applying this technology for other scenarios, such as health care, smart grid/energy etc. In these projects, the main motivation has been the creation of accessible user interfaces. In SensHome, we will for the first time apply this technology for the creation of a corpus.

[1] `http://www.i2home.org`
[2] `http://www.ist-vital.org`

6. Conclusion and Outlook

We have presented our efforts within the SensHome project in creating a corpus for activities within a Smart Home scenario. The SensHome corpus builds on previous efforts, such as the NETCARITY corpus. There, purely controlled experiments with pre-defined activities are within the corpus. In addition to its annotations with visual features extracted from the video material only, the SensHome corpus will contain higher-order activities like our episodes and explicit temporal annotations, speech and dialogue acts (Bunt et al., 2010).

7. References

Jan Alexandersson. 2008. i2home—Towards a Universal Home Environment for Elderly and Disabled. *KI*, 3/08:66–68.

James F. Allen. 1984. Towards a general theory of action and time. *Artificial Intelligence*, 23:123–154.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu, Claudia Soria, , and David Traum. 2010. Towards an iso standard for dialogue act annotation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation — LREC 2010*, Malta, May.

Dominikus Heckmann. 2006. *Ubiquitous User Modeling*. Akademische Verlagsgesellschaft Aka GmbH, Berlin. ISBN 3-89838-297-4 and ISBN 1-58603-608-4.

ISO. 2008. *ISO/IEC 24752: Information Technology — User Interfaces — Universal remote console — 5 parts*. "International Organization for Standardization".

Victor Kaptelinin and Bonnie A. Nardi. 2006. *Acting with Technology - Activity Theory and Interaction Design*. The MIT Press.

Jonathan Kilgour and Jean Carletta. 2006. The nite xml toolkit: demonstration from five corpora. In *NLPXML '06: Proceedings of the 5th Workshop on NLP and XML*, pages 65–68, Morristown, NJ, USA. Association for Computational Linguistics.

Michael Kipp, 2010. *Multimedia Information Extraction*, chapter Multimedia Annotation, Querying and Analysis in ANVIL, page Chapter 19. MIT Press.

Aleksei N. Leontiev. 1978. *Activity, Consciousness, and Personality*. Englewood Cliffs, N.J.: Prentice Hall. (Original work published in Russian in 1975).

Charles Rich. 2009. Building task-based user interfaces with ansi/cea-2018. *Computer*, 42(8):20–27.

Christoph Stahl. 2009. *Spatial Modeling of Activity and User Assistance in Instrumented Environments*. Ph.D. thesis, Universität des Saarlandes, Department of Computer Science.

Gottfried Zimmermann and Gregg Vanderheiden. 2007. The universal control hub: An open platform for remote user interfaces in the digital home. In Julie A. Jacko, editor, *Human-Computer Interaction*, volume 4551 of *LNCS*, pages 1040–1049. Springer.

Ingo Zinnikus, Klaus Fischer, Jan Alexandersson, and Unai Diaz. 2009. Bringing the elderly into the mainstream of e-society: the vital project. In *IADI e-Society Conference*, Barcelona, Spain.

Look at me!: An emotion learning reinforcement tool for children with severe motor disability

Beatriz López-Mencía[1], David Pardo[1], Néna Roa-Seiler[2], Álvaro Hernández-Trapote[1], Luis A. Hernández[1], Mª Carmen Rodríguez[3]

[1]GAPS, Signal Systems and Radiocommunications Department. Universidad Politécnica de Madrid.
Ciudad Universitaria s/n, 28040 Madrid (Spain)
[2]Centre for Interaction Design. Napier University.
Edinburgh, EH10 5DT. (United Kingdom)
[3]Telefónica I+D,
Emilio Vargas 6, 28043 Madrid (Spain)
E-mail: beatriz@gaps.ssr.upm.es, david@gaps.ssr.upm.es, N.Roa-Seiler@napier.ac.uk, alvaro@gaps.ssr.upm.es, luis@gaps.ssr.upm.es, mcrg@tid.es

Abstract

In this paper we present a freeware emotion learning reinforcement application for children with severe motor disability. The main technological features of this educational support tool are those that provide it with its interactive capabilities, most notably, the embodied conversational agent (ECA) and multimedia elements (webcam, audio and pictures). We also identify the need to develop multimodal corpora to study how best to apply these technologies to aid in the emotion (and, more generally, social) learning processes of these users, and we describe our own data collection and annotation procedure. First real user experiences were a success, as the children were highly motivated to engage with the system.

1. Introduction

People communicate with complex and subtle signals that are put together to form messages following both innate and culturally established interaction rules. The human body can be viewed in this context as a medium to convey and receive messages composed mainly of spoken, gestural and postural elements. Children with permanent and severe movement disorders –cerebral palsy is one such condition– have great difficulty perceiving and performing these interaction procedures, and indeed learning them. A range of therapeutic techniques are commonly employed in these special cases to aid in the development of communication skills. One important area in which children with motor disorders can benefit from educational reinforcement is recognizing and performing emotions.

In the present paper we describe a software tool we have designed as an educational aid for the development of emotional skills. We have aimed to encourage interaction with our educational tool by combining visual and auditory elements in its interface. In particular, the interface features an Embodied Conversational Agent (ECA) that performs facial expressions and body movements, and has been programmed to play out emotions.

An ECA is a realistic virtual human that possesses the ability, to a greater or lesser degree, of engaging in conversation with a human. This implies the ability to understand and generate speech, hand movements and facial expressions. (Cassell, 2000)

It has been shown that in learning applications the use of simple ECAs displaying very basic behaviour can help increase motivation and interest, and lead to improved results over those obtained with systems featuring only text or voice (Moreno et al., 2001; Atkinson, 2002). In fact, the sole presence of the virtual agent can generate a more positive opinion of the system, the interaction with which tends to be seen by users as easier and more fun (Van Mulken et al., 1998). These benefits are motivating efforts to develop educational applications that make greater use of the expressive possibilities of ECAs (Wik & Hjalmarsson, 2009; Roa-Seiler, 2010). Combining ECAs with other kinds of audiovisual information also opens up interesting possibilities. For instance, in a recent study one such system applied to teaching pronunciation and articulation in Swedish worked better than a human tutor (Engwall, 2008). The MYSELF project features a virtual tutor, Linda, that has the ability to recognize certain emotions displayed by students (Anolli et al., 2005).

The benefits of ECAs have also begun to be explored in the context of special education. In Tartaro and Cassell (2006) an ECA is used as a virtual peer to help develop communication and social skills in children with autism. In a related study in which autistic children engaged in a collaborative narrative with both a virtual and a human peer, Andrea Tartaro found that discourse topic management occurred more often with the virtual peer than with the human (Tartaro & Cassell, 2008). The importance of endowing virtual agents with gestural and emotional display skills when used in special education has been stressed by Cosi, who reported improved reading and writing skills in children with speech difficulties using an expressive agent named Lucia (Cosi et al., 2004). The strong supporting role of emotions in the learning process is also explored in Mohamad et al. (2005), who use an emotional interface (an ECA that expresses

emotions) in a therapeutic learning environment for disabled children. Finally, multimedia technology is also being used as learning support for young people with physical disabilities to improve their social skills (Kiung et al., 2008) as well as to strengthen their affective communication skills and their ability to express emotions (Baron-Cohen, 2009; Picard, 2009).

We have offered a sample of the research efforts to apply multimedia and multimodal interaction technologies to special education. Nevertheless, considerably more experimental grounding is required to learn how to use emotions in these systems. A major obstacle in the way of progress in humanlike interaction technologies is the lack of multimodal corpora. This deficiency is particularly acute in the case of applications for children with special educational needs. It is certainly a challenge to develop tools to collect corpora of multimodal emotional behaviour. The HUMAINE Database (Douglas-Cowie, 2007) is a network that seeks to address the problem by providing material and labelling techniques to describe emotional corpora in different scenarios. Likewise we point out some outstanding initiatives in the development of coding schemes for studying multimodal behaviour. In particular MUMIN network (Allwood et al., 2007) is focused in developing multimodal annotation of feedback, turn management and sequencing phenomena information. A more centred approach to annotation of multimodal emotional behaviour is proposed by Jean Claude Martin (Martin et al., 2005).

With respect to existing resources and databases, the Swedish NICE corpus described in Bell et al. (2005) deals with multimodal spontaneous child-computer interaction. Their work is focused on spoken dialogue analysis between children and several embodied conversational agents. On the other hand, the database of Little Children's Interactive Multimedia Project (CHIMP) project is an audio-visual corpus of 50 children conversations with human moderator and a Wizard-of-Oz controlled computer character. This resource is a ground to carry out comparative studies between child-human and child-computer interaction (Black et al., 2009). They annotate the interactional behaviour reaching some interesting results – i.e. the children adaptation to their conversational partner's interaction style (regarding to verbal and nonverbal behaviour).

But in spite of the existence of some general tools such as these, there is a lack of specific resources for the development and evaluation of special education applications with interactive multimedia and multimodal elements capable of handling emotional cues.

Our present contribution lies in this gap. We have developed an educational reinforcement application – featuring interactive multimedia elements– to help children learn facial expressions. To do so we have followed recommendations from the teaching staff at the Infanta Elena Special Education School in Madrid (CPEE Infanta Elena). In this paper we first describe the application (Section 2), focussing mainly on the functional aspect, and, secondly, we describe our multimodal corpus generation scheme, illustrating it with real examples of use (Section 3).

Our tool is freeware and we are making it available for research purposes to the scientific community and special education schools. (See GAPS, 2010)

2. The emotion learning reinforcement tool

In this section we describe the educational software tool we have developed to reinforce the learning of emotional expression by children with severe motor and language disability. After several meetings with the educational staff –teachers and therapists– at CPEE Infanta Elena we derived a set of requirements for our learning tool:

- The animated agent should take the role of a virtual peer –another child, rather than appearing as another teacher. Its role should be to provide feedback by acting out emotions on request and responding to the child's behaviour with positive or negative reinforcement. A Wizard of Oz scheme (with an educator as Wizard) was the method of choice.
- The interface should show a live video image of the user on screen (captured by a webcam). The children showed great interest when they saw themselves on screen and acted as though facing a "virtual mirror". This tool enables educators to recreate the traditional method of teaching emotions using a mirror.
- Positive or negative feedback should be given both as reinforcement (clapping and spoken phrases) and to attract the attention of children.

2.1 Technical description

We used Haptek's software to our create animated figures (Haptek, 2010) –it is freeware flexible enough to realistically recreate a variety of facial expressions. Haptek provides an ActiveX component which allows easy integration in the IDE Visual Basic .NET environment we chose to work with.

This application has two forms, one for the child and the other for the wizard, connected through a communication hub. The hub receives commands from the wizard and resends them to the child's interface where they are interpreted. Both interfaces have been designed using default controls from Visual Basic .NET libraries. The conversational agent uses pre-recorded phrases to provide feedback to the user. A text-to-speech converter could, however, be easily integrated.

2.2 Functional description

The purpose of this application is to reinforce the learning of associations between emotions and their physical display (in the form of facial gestures). The visual interface consists of a frame where the avatar performs an emotion (optionally, a picture representing the emotion in question may appear alongside the virtual agent); a "mirror" frame where children can see their own image; and a control panel with action and reinforcement buttons. Figure 1 shows what the Wizard visual interface looks like.

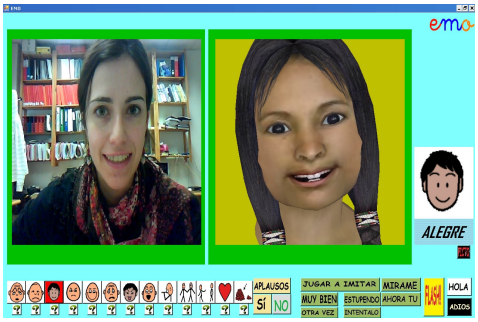


Figure 1: Emo interface

A control panel on the bottom part of the screen shows the available functions (Figure 2):

Welcome and farewell messages (number 1 in Figure 2)**.** The idea is for children to learn to associate these messages with the beginning and the end of the exercise.

Emotions and facial expressions displayed by the ECA. The available emotions and facial expressions were defined by the teaching staff at Infanta Elena. These are "Hungry", "Surprised", "Like", "Dislike", "Angry", "Smiling", "Tired", "Sad", "Happy", "I want more", "I love you", "I want to be alone", "I don't want to be alone" (number 3 in figure 2). The Wizard (teacher) can choose an emotional state by pressing the button that represents it. The ECA then acts out the emotion (the ECA also says aloud the name of the emotion/action that is being performed). (When creating the avatar's behaviour we mimicked –as accurately as possible– the gestural behaviour used in class by the teachers themselves.)

Emotions performed by children. Besides expressing emotions, the virtual agent can also ask the children to perform the same emotion she is acting out. There is a row of pictures that represent emotional states or actions (number 4 in Figure 2). Underneath each picture there is a question mark (?) button. If pressed, the virtual agent asks a question in relation to the corresponding picture. For example, "How do you get angry?"

Reinforcements (number 2 in Figure 2) were included at the request of the teachers and therapists, who highlighted their importance in the learning process. In addition to "yes" and "no" responses from the ECA (visual and auditory stimuli), we included applause and encouragement phrases such as "Well done! Fantastic!", "Very good!", "Great!", "Try again!" There are also reinforcement phrases to try to focus the child's attention on the avatar's face: "Play with me! imitate my face!", "Look", "It is your turn." Optionally, a blinking frame around the ECA can be also added as a visual stimulus to attract the child's attention.

3. The corpus

The first things to consider when collecting data for a corpus are the issues of privacy and consent. We obtained written authorization from the parents to record video and interaction data of the children involved in the tests. Of course, preservation of privacy demands that anonymity be guaranteed, so care must be taken to eliminate all identifying information in the corpus. Further, the parents are required to specify what they allow the recorded information to be used for (strictly internal use, for research purposes only, for scientific publications, etc.).

3.1 General features

Let us now describe the main elements of our multimodal corpus design to collect our test data of children-educator-system interactions. We illustrate the process using interaction data collected in the course of preliminary tests carried out with four children chosen by the therapist and two teachers at Infanta Elena. All of the selected children have cerebral palsy, severe motor problems and are unable to speak. Their ages range from 4 to 14 years, which makes it possible to compare the use of the tool with children of very different ages. Subsequent sessions can be compared later to see if there are observable learning improvements.

Two camera angles were recorded: a frontal medium close up shot of the child's face and a lateral wide shot capturing both the child and the teacher (the wizard). We didn't considered necessary to record the avatar's actions because they could be completely logged paying attention to the audio information and the button events from the wizard interface.

The multimodal corpus is first made up of video (from two camera angles) and audio. Later, metadata is added as annotations to the interaction logs. This we shall see in the following subsection.

3.2 Extracting interaction metadata

The next step is to analyse the recorded interactions and identify relevant events that will help analyse the children's performance with the application. We use ANVIL (Kipp, 2001) to annotate the video recordings. Figure 3 is a screen shot of the annotation tool interface.

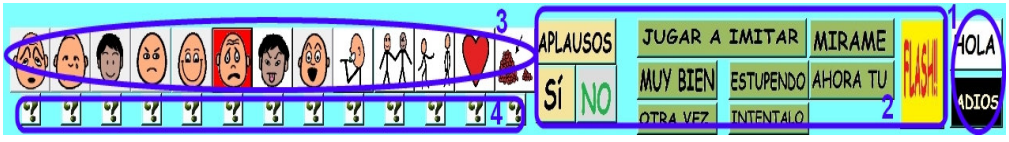


Figure 2: The control panel

Annotation tracks		Description	When it occurs
Duration		Overall duration of child interaction with the application	From avatar greeting to farewell. It machtes up with the pressing of 'Hello' button and 'See you soon' button respectively.
Emotions		Emotions and facial expression performed by the virtual agent during the interaction. The values are: *"Hungry", "Surprised", "Like", "Dislike", "Angry", "Smiling", "Tired", "Sad", "Happy", "I want more", "I love you", "I want to be alone", "I don't want to be alone"*	It happens only the first time a emotion button is pressed.
Feedback	*Reinforcement*	*Replays of emotions*. Redundant emotion during a session.	It happens when it is not the first time that a emotion button is activated during the same session or when the Wizard (the teacher in this case) presses the buttons associated with 'Try again¡' and 'Do it again¡' actions.
		Positive reinforcement. By means of sentences which encourage the child and try to keep the work level.	The buttons associated to these reinforcements are: 'Applause', 'Very good¡' and 'Great¡'
	Calls	*Audio reinforcement*, through the sentence 'Look at me¡'. The ECA tries to capture the attention of the child in order to prepare her/him for working with an emotion.	It occurs when the Wizard presses the button associated to this action.
		Visual reinforcement. It is a blinking frame around the ECA that tries to call the child attention to this part of the screen.	It happens when the Wizard presses 'Flash' button.

Table 1. Annotation structure and description of each track.

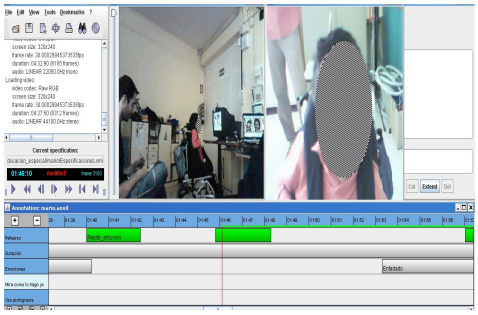


Figure 3: Screenshot of ANVIL video annotation tool.

Firstly, we have annotated only the parameters related to the interaction flow of each session. We have extracted a number of first-level objective items that shows the child-Wizard interaction behaviour. This information contains raw data extracted from the videos files. We hope this information will be helpful for us to evaluate analytically the progress achieved by children throughout different sessions. For example, if the number of times that the Wizard ask to a child to replay a particular emotion is lower during the second session, it could represent that the child has improved this specific emotion learning.

Later, in a second phase of this work, we plan to annotate the children's emotions which are obtained from their facial expressions. For this complex task we will count on the collaboration of the teaching staff and the results delivered from software specialized in emotion recognition through facial expressions.

We present in Table 1 the structure of our annotation scheme and the description of each track which have been extracted from the videos. It is worthy to mention those tracks related with the *Feedback* category. In the Section 2.2 we stressed the importance of the reinforcement for improving the learning experience of children. Due to that we have included two kinds of feedback actions: *Reinforcements* and *Calls for attention*. This information

could show us whenever a child loses the focus of the exercise or the effects that positive reinforcements could have in the achievement of the goals.

All this information has been extracted from the audio recorded in video files. Nowadays we are working on extracting some of the parameters automatically from the events associated to the buttons. Thus the annotation process would be focused to the subjective information, in other words, to the gesture and emotion level.

To obtain the associated metadata we then compute the number of events observed (and annotated) on each of these tracks. Table 2 shows results obtained for two of the children in the preliminary test group. We may observe that one child requires a significantly greater amount of reinforcement than the other, and also that while one child needed his attention to be attracted at the beginning of the interaction, the other child tired of the interaction and needed attention calls toward the end.

Interesting as it may be to look at how the program works for each particular child, it is, nevertheless, by annotating and extracting interaction information that we obtain a reference for a long-term evaluation in which average score tendencies make it possible to associate design and functional elements with specific areas where progress tends to be achieved by the children.

	Child A	**Child B**	**Average (Std.)**
Duration	389	257	323 (93,30)
#Emotions	12	11	11,50 (0,707)
#Calls for attention	1 *(start)*	3 *(end)*	2 (1,414)
#Replays of emotions	8	5	6,5 (2,121)
#Positive reinforcements	17	12	14,50 (3,535)
#Replay + Positive Reinforcement	6	7	6,5 (0,707)
#Total Reinforcements	31	24	27,50 (4,95)

Table 2. Preliminary results for two children

4. Conclusions and ongoing developments

In this paper we have provided a technical and functional outline of an emotion learning reinforcement tool we have developed for use with children who have severe movement disorders. We have also described our corpus design and how we obtain and annotate multimodal interaction data.

Preliminary results are promising. The tool is a success with the teaching staff and the children at CPEE Infanta Elena, and the comments received are overwhelmingly positive. This encourages us to continue our efforts.

An interesting development at this point would be to automate as much of the annotation process as possible, as well as the extraction of the associated metadata. This is something we are working on at present.

However, not all information can be extracted automatically. Experts –the teachers and therapists– are needed to identify, explain and classify much of the activity during the children's interactions with the learning support system. Indeed, an expert's eye is needed to identify expressions displayed by the children, as well as their intensity, and to quantify the progress made by each particular child. The most reliable information would obviously come from each child's own caretakers. Nevertheless, automatic emotion recognition could be used to aid the experts, particularly to maintain consistency through large batches of interaction observation sessions. We are also working to incorporate this feature using Shore emotional recognition software (from FaceDetect Fraunhofer (Shore, 2010)). It will be interesting to compare the data obtained with this recognition tool and the observations made by the experts.

Finally, we are preparing a long-term study on how the the interactive components of our emotion learning tool (ECA and Multimedia) affect the learning process. We are at present waiting to collect all the recordings made throughout the current school year.

5. Acknowledgements

The activities described in this paper were funded by the Spanish Ministry of Science and Technology as part of the TEC2009-14719-C02-02 project. We would also like to thank to the support of the COMPANIONS project IST-34434. We thank the teaching staff at CPEE Infanta Elena for their invaluable help and recommendations. Our gratitude goes also to Virginia León, Luca Bersano, Javier Poza, Jorge López and Jose Luis Naranjo for their work on the tool, and to Vasile Vancea for his contribution to the literature review.

6. References

Allwood J., Cerrato L., Jokinen K., Navarreta C., Paggio P., (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language resources and Evaluation* 41:273-287.

Anolli L., Mantovani F., Balestra M., Agliati A., Realdon O., Zurloni V., Mortillaro M. Vescovo A. and Confalonieri L. (2005). The potencial of affective computing in E-learning: MYSELF Project experience, *INTERACT 05, Workshop on eLearning and Human-Computer Interaction*.

Atkinson, R. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology*, 94(2):416-427.

Bell, L., Boye, J., Gustafson, J., Heldner, M., Lindström, A. & Wirén, M. (2005) The Swedish NICE Corpus – Spoken dialogues between children and embodied characters in a computer game scenario. *Proceedings of Interspeech*, Lisabon, Portugal.

Black, M., Chang, J., Chang, J., and Narayanan, S. (2009). Comparison of child-human and child-computer interactions based on manual annotations. In *Proceedings of the 2nd Workshop on Child, Computer*

and interaction (Cambridge, Massachusetts, November 09). ACM, New York, NY, 1-6.

Baron-Cohen S., Golan O., Ashwin E. (2009). Can emotion recognition be taught to children with autism spectrum conditions?. *Phil. Trans. R. Soc.B* 364, 3567-3574.

Cassell, J. (2000). Embodied conversational interface agents. *Communications of the ACM*, 43(4):70-78.

Cosi, P., Delmonte, R., Biscetti, S., Cole, R.A., Pellom, B., and van Vuren, S. (2004). Italian literacy tutor - tools and technologies for individuals with cognitive disabilities, In *ICALL- 2004*.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., Martin, J.C., Devillers, L., & Batliner, A. (2007). The humaine database: addressing the needs of the affective. In Paiva, A. and Prada, R. and Picard, R. (Ed.), *2nd International Conference on Affective Computing and Intelligent Interaction (ACII'2007), LNCS, vol. 4738* (pp. 488-500). Lisbon, Portugal.

Engwall, O. (2008). Can audio-visual instructions help learners improve their articulation?-an ultrasound study of short term changes. *In Proc. Interspeech*, pages 2631-2634.

GAPS (2010). Available at: http://www.gaps.ssr.upm.es/es/investigacion/tecnologia-soporte-a-la-educacion-especial

Haptek (2010). Available at http://www.haptek.com

Kipp M. (2001) Anvil - A Generic Annotation Tool for Multimodal Dialogue. *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1367-1370.

Martin J-C, Abrilian S, Devillers L (2005) Annotating multimodal behaviors occurring during non basic emotions. *In: 1st international conference on affective computing and intelligent interaction (ACII'2005)*, Beijing, China, Springer, Berlin, 22–24, pp 550–557.

Mohamad, Y., Velasco, C., and Tebarth, H. (2005). Development and evaluation of emotional Interface Agents in Training of learning disabled children, *In Workshop on "The role of emotion in human-computer Interaction"* at 19th British HCI Group Annual Conference.

Moreno, R., Mayer, R., Spires, H., and Lester, J. (2001). The case for social agency in computer-based teaching: do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction, pages* 177-213.

Ng C.Kiung , Liew Y.T., Saripan M. I. , Abas A.F., Noordin N. K. (2008). Flexi E-Learning System: disabled friendly education system. *European Journal of Social Sciences* – Vol 7, nr 2.

Picard R. W. (2009). Future affective technology for autism and emotion communication. *Phil. Trans. R. Soc. B* (2009) 364, 3575-3584.

Roa Seïler N., Benyon D.,(2010). Designing Companions with Kansei. *To be published in proceedings of International Conference on Kansei Engineering and Emotion Research 2010* , Paris, France.

Shore, (2010). Available at: http://www.iis.fraunhofer.de/EN/bf/bv/kognitiv/biom/dd.jsp

Tartaro, A., and Cassell, J. (2006). Authorable Virtual Peers for Autism Spectrum Disorders, *Proceedings of the Workshop on Language-Enabled Educational Technology at ECAI 06*,August pp.28-31.

Tartaro, A. and Cassell, J. (2008). Playing with Virtual Peers: Bootstrapping Contingent Discourse in Children with Autism. in *International Conference of the Learning Sciences*. Utrecht, the Netherlands, ACM Press.

Van Mulken, S., Andre, E., and Muller, J. (1998). The persona e_ect: How substantial is it? *People and Computers*, pages 53-66.

Wik, P. and Hjalmarsson, A. (2009). Embodied conversational agents in computer assisted language learning. *Speech communication*, 51(10):1024-1037.

Capturing multimodal interaction at medical meetings in a hospital setting: Opportunities and Challenges

Bridget Kane, Saturnino Luz, Jing Su

Department of Computer Science
Trinity College
kaneb@tcd.ie, luzs@scss.tcd.ie, sujing@tcd.ie

Abstract

This paper highlights the issues involved in gathering a corpus of data on the multimodal interaction that occurs at a team meeting of medical specialists. Difficulties in capturing the data are described, and the ethical issues are emphasised. Methods to investigate the internal structure of meetings, at the level of discussion topic (patient case discussion) are summarised and the potential benefit that such meeting records promise are reviewed. The hospital setting where the corpora are proposed experience issues in common with any business venture, but in addition demonstrate additional sensitivities because of health service complexities and patient privacy issues.

1. Introduction

This paper discusses the issues involved in the recording, annotation and analysis of conversation among a group of hospital specialists at multidisciplinary medical team meetings (MDTMs). We report on progress so far in the automatic annotation and analysis of these multimodal corpora and the potential usefulness of our approach to information capture in MDTM settings.

A teaching hospital setting has all the issues that one encounters in any workplace setting such as policies, procedures, culture, company business concerns and privacy, with additional requirements because of the fact that the collection is in healthcare.

We see that corpus gathering in this situation is useful and important for three reasons. Firstly, as a research tool to improve our understanding of the role interactions, information sharing and collaboration in this type of work. Challenges encountered in a teaching hospital with regard to the multitude and complexity of specialist roles is normally greater than in most organisations. Through using a hospital forum to study collaborative work, the findings are likely to find application in a variety of other work settings. Secondly, we believe a meeting record in this setting would provide a useful artefact or co-ordinating mechanism for the co-operative work involved in patient care. By demonstrating its utility in this respect we expect that electronic meeting records will be eventually adopted. Finally, a meeting record would facilitate organizational level tasks such as audit and quality review for organizational development and learning. We expect that by integrating meeting records into existing data gathering methods in the hospital will prove more efficient than current practices of independent specialised data collection for a single audit question. Changing the methods for information monitoring and audit review has the potential to reshape organisations and enhance quality improvement initiatives in the long term.

We summarise the main difficulties in achieving meeting corpora in business meeting scenarios, particularly in the health sector, and we also consider issues in conducting this type of research, i.e. different role perspectives, privacy constraints and ethics.

The main goal of this paper is to identify the key issues that workplace researchers need to address, together with the issues specific to the healthcare service. We also discuss how these concerns can be addressed.

2. Background

Routine multidisciplinary medical team meetings are becoming an important event in modern hospital life, particularly in cancer centres, as more and more professional organizations and regulatory bodies are making recommendations for the adoption of MDTMs into patient care pathways (Calman and Hine, 1995). The rationale for their adoption is twofold: i) they provide a useful forum for triple assessment of the patient's clinical findings, thus improving the quality of diagnosis, and ii) as the management of disease becomes more and more complex, multidisciplinary discussion is a useful co-ordinating mechanism for treatment planning and the management of individual patients. Many related work activities outside of the meeting require careful planning and co-ordination for the system of multidisciplinary team patient care to work efficiently (Kane and Luz, 2009; Kane and Luz, 2006a). For example, a cancer patient might require treatment through a few modalities, namely surgery, medical oncology and radiation oncology. The sequence and timing of these interventions might be sequential, concurrent, or in a combination, depending on the tumour type, size and anatomic location. The co-ordination and timing of such treatment strategies can be crucial to a successful outcome for a patient, and requires high levels of co-ordination and interaction among the associated specialities involved.

MDTMs provide a valuable resource for information gathering to inform patient management tasks subsequent to the meeting, and they have potential to be used as an information resource for audit and planning purposes. With the potential value of meeting recordings in mind, we investigated human and technological issues involved in building advanced computing support for collaboration, production

and access of electronic medical records in the context of MDTMs. It is apparent that although recent technological and organisational developments have made digital recording of entire meetings a distinct possibility, the usefulness of this kind of audiovisual database is dependent on how effectively its contents can be accessed, among other factors. We believe that the internal structure within the MDTM can be harnessed so that elements of the discussion, or particular information, can be retrieved from recordings more effectively than linear methods alone will allow.

In conducting our research on meetings in a busy hospital setting, we experience many of the difficulties in common with any complex work setting, such as limited resources, interruptions and rescheduling of tasks due to national and personal holidays, staff illness, etc. Additional issues more prominent in healthcare are encountered, such as medical emergencies, patient concerns for privacy and confidentiality and new developments in technology and treatments. The main issues that we identify here and discuss in this paper relate to respecting the patient's privacy and that of the health professionals collaborating at the meeting.

3. Methods

The work reported here is based on several years of ethnographic observation, supplemented with audiovisual recording, together with questionnaires and interviews conducted with the multidisciplinary team members. Specific exercises that targeted particular research questions are reported elsewhere, such as in (Kane and Luz, 2006b). This paper reports in a more general way on the overall issues concerning the multimodal corpus collection. The meetingroom where the corpus was gathered is shown in Figure 1.

Figure 1 shows the team engaged in discussion. Radiological images from either disk, radiological film or the PACS[1] system and pathology (tissue) samples from a microscope are shown on the main screen display. Images may also be used, from time to time, that were taken at patient procedures that pre-date the meeting, such as video clips taken at surgery or at endoscopy.

3.1. Audio Capture Requirements

The human voice frequency band range is approximately from 80Hz to 1100Hz, and the frequency response range of a selected microphone should span this voice band range, as a minimum. One type of widely used conference microphone offers the frequency response range from 30Hz to 20,000Hz. Given our interest in indexing the recorded meeting data, we adopted the sampling rate of 16kHz, which is commonly used in speech recognition (Lee et al., 1989), in order to convert the microphone's analog output signal into a digital format.

MDTMs are held in a closed meeting room with 10 to 20 participants. Clinical specialists sit beside each other in rows, and face the main monitor. We record voice from clinical specialists for further analysis. The recorded audio needs to offer satisfying audio volume, have a high signal to noise ratio and a moderate frequency response range for human voice.

[1] Picture Archiving and Communication System

We aim to evaluate how a speaker influences meeting structure through topic change, so the optimal recording strategy is to separate each speaker's voice through the recording devices. Throat microphones are superior to traditional microphones for this task because they capture the sound wave more directly from the vocal chords thus reducing outside noise interference. Audio signals from one throat microphone can be recorded in a single channel, on which the target speaker's voice is prominent over peripheral speakers and noise. In post-capture processing, the audio files can be filtered through speaker diarisation so as to generate files containing the voice of a single speaker. As we note below however, despite the fact that throat microphones would have been an optimal choice from an audio processing perspective, they proved unacceptable for this particular data gathering project.

In the current MDTM setup (see Figure 1), two cardioid condenser boundary microphones are mounted on the front wall and side wall of the meeting room. The distance between microphone and the main speakers is about 3 meters. Cardioid microphones pick up sounds from all directions, so that voices from all speakers are recorded in a single channel. In comparison with the throat microphones, cardioid microphones record lower quality audio, but they do not interfere with the meeting participants. In order to locate vocalization boundaries of each speaker from the cardioid microphones recordings, speaker diarisation and segmentation algorithms can also be executed, though the results will be far less satisfactory. These proceedures are required if one aims to perform high-level segmentation, categorisation, and other forms of indexing on the meeting data (Bouamrane and Luz, 2007). Chen (Chen and Gopalakrishnan, 1998) suggests Bayesian Information Criterion (BIC) (Schwarz, 1978) as a standard to evaluate the coherence of continuous speech, for speaker segmentation. Speaker identification techniques can be used to label all vocalizations from the same speaker. Gaussian Mixture Models (GMM) can be used for this task (Reynolds and Rose, 1995). The use of Gaussian Mixture Models for modeling speaker identity is motivated by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes. In other words, these are acoustic classes, which are useful for modeling speaker identity.

3.2. Data gathering in practice

Due to the fact that we aimed to record real MDTMs, subjected to the stringent constraints of a medical setting, the actual practice of data gathering deviated significantly from the requirements scenario outlined above. The multimodal corpus was gathered from two media sources. The first was an S-VHS recording facility in the Telesynergy®system (Martino et al., 2003) of the audio in videoconference together with the screen display being broadcast to the meeting. A proportion of the meetings were held in videoconference, and for those meetings the recording captured the incoming video stream together with the audio discussion in videoconference. Outgoing video data were captured through the picture-in-a-picture view that was displayed on a TV monitor during the conferences. Given that the research involved a second institution, approval was also

Figure 1: Multidisciplinary medical team meeting

sought from the staff at the second hospital site.

The second recording source was a video camera placed at the back of the room, in the same location from which the still image in Figure 1 was taken. This recording captured gestures and movements among the participants as they pointed, turned to gaze in particular directions or engaged in personal note-taking. These observations were valuable in helping identify information gathering needs of individual roles within the group.

Participants at the meeting did not wear any microphone devices, because of the principle agreed at the outset not to interfere in the work of the staff in any way (Section 3.3.). This resulted in less than perfect recordings, because of background noise from people handling papers, coughing, sneezing, moving in seats, etc. Furthermore, while the staff were agreeable to the recording of the meeting in order to gather annotations and verify observations, there were some reservations expressed concerning potential breach in confidentiality. Once assurances were given and trust was established, then cooperation was achieved.

The wall-mounted cardioid condenser boundary microphones were used to capture the speech from the participants in discussion. As mentioned above, we would have preferred if the participants had worn throat microphones, to improve the signal to noise ratio of the recorded audio. However because throat microphones exert pressure on the throat and may distract clinical specialists, and the danger of this having a detrimental indirect effect on patient care, it was decided to use the wall mounted cardioid condenser boundary microphones only. While this method proved satisfactory for our purposes, we noted that the audio recordings from videoconference was superior to that of co-located discussion. This may be due to the way the Telesynergy system (McAleer et al., 2001) used in the recordings was configured.

The S-VHS recording was transferred to digital tape and both recordings were converted to MPEG-4 format. These media files were imported into the Elan annotation tool (MPI, 2005), synchronised and annotated. Annotations were prepared manually in the first instance at the level of individual patient discussion. Following the full meeting being segmented into individual patient discussions, each discussion was then sub-divided into its natural sections, defined as D-Stages, and fully described in (Kane and Luz, 2009). The identifiable discussion tasks were further subdivided into four sub-sections. At a deeper level of detail, vocalisation events by participants were annotated and labelled with the individual's identifier and professional role. Analysis was subsequently conducted at individual and specialist role levels (reported elsewhere), since there was more than one individual for any particular role. For example there were two consultant radiologists in attendance and three respiratory physicians at most meetings.

Automatic annotation has also been performed at different

levels based on the recorded speech. These included speech segmentation and speaker diarisation (Su et al., 2008), segmentation of meetings into patient case discussions (Luz, 2009), categorisation of such discussions (Luz and Kane, 2009), and speaker role identification (Su et al., 2010). Although speaker diarisation in noise settings remains a difficult problem, higher level segmentation tasks can be performed accurately enough to facilitate the process of manual annotation by researchers (e.g. as an add-on to tools such as ELAN) or to support browsing of meeting records by users.

Following collection of the detailed annotations the original recordings containing confidential patient data and actual medical discussions were destroyed, as agreed at the outset.

3.3. Privacy and Ethics

In the first instance ethical approval was required by the hospital Board before commencing any research with the multidisciplinary team. Two areas of ethical concerns were required to be addressed: i) The protection of the business interests of the hospital as would be required in any company, and ii) concerns for patient confidentiality.

In gaining confidence of the hospital staff, the ethical committee required that one of the senior medical consultants (staff member) vouched for and supported the research proposal. This individual agreed to mentor the study and also accepted responsibility for maintaining the highest ethical standards on behalf of the hospital. All staff were informed of the nature of the study by the lead researcher at the outset. As part of maintaining on-going cooperation and trust, regular progress reports and result data are provided to the staff.

Because our primary interest was in the hospital work systems and the role and effectiveness of the multidisciplinary team meeting in those processes, our research was given approval. Had this project focussed on any individual patient data, further ethical process would have been required as part of the research approval procedures of the hospital.

The researchers undertook not to interfere in any way in patient management, as well as giving undertakings that any patient information that was incidentally learned in the course of the research would be respected and maintained in the strictest confidence. The issue of how to preserve anonymity and the sensitive content in recording medical and other types of meetings while maintaining enough of the original data (speech and video signals) to allow researchers to investigate automatic meeting indexing methods has received attention from the research community in recent years. Promising directions include the gathering of "sociometric" signals though unobtrusive wearable devices (Olguin et al., 2009), and the use of digtal signal processing techniques for anonymising speech data (Parthasarathi et al., 2009). Achieving an acceptable way of recording and storing multimodal data is crucial to corpus gathering and data indexing research in medical settings.

4. Structure of the meeting

In an analysis of the collaboration and interaction exhibited at MDTMs, an apparent structure was identified, described in (Kane and Luz, 2009). MDTM are composed of several patient case discussions (PCDs) and internal structures have been identified. These structures reflect the highly structured tasks undertaken in the discussion and are a testament to the medical tradition of conducting a patient assessment in a predictable way, i.e. information is methodologically reviewed and the underlying cause of the presenting problem is assessed in the first instance, before treatment is prescribed. During PCDs the narrative follows with tradition in the conduct of the two main tasks, namely, the patient diagnosis and the next step in the patient's management. For each of these tasks specialities interact, collaborate, exchange and share information. Images are used by some of these specialities, for example radiologists use patient radiological imaging, pathologists show microscopic images and surgeons may use video to demonstrate their findings or procedures. Participants have been observed to point at images, and use their hands to describe the complexities of size and shape of tumours. Drawings have also been used by participants to explain the orientation of a tumour, or finding, at MDTMs. Representational gestures have been found to play an important role in medical meetings (Becvar et al., 2008). Therefore, multimodal meeting corpora from MDTMs would include the artefacts used, gestures and annotations, but in the corpora described here, we confine ourselves to audio recordings of the speech interactions.

As well as identifying internal structures of MDTMs and PCDs through our ethnographic observations, we are investigating automatic methods, as outlined in Section 3.. PCDs segments in MDTMs are analogous to *topics segments* in more general meeting corpora such as AMI (Carletta, 2007). We found that features of the vocal interactions such as the speaker ID, role, length of vocalisation, pauses and overlaps are useful in helping segment the meeting data into individual topics, or individual patient case discussions (PCDs). These features may also be useful in segmentation of the internal PCD sub-section boundaries which we define as D-Stages.

5. Utility of MDTM Records

The fact that an internal structure is identifiable through automatic means suggests to us that these techniques could be applied for the retrieval of individual discussions that match particular criteria. Such a development would potentially facilitate a number of important hospital functions listed in Table 1, particularly review of meeting proceedings without necessitating a full meeting review and the development of a corpus of patient cases that would inform future decisions, including the development of clinical practice guidelines. These potential uses are discussed below.

Individual Contributors would be able to review or check their input to a discussion. Sometimes a patient's results might be given an emphasis in discussion that is not reflected in the formal written report in the patient's file and this can lead to later confusion. A PCD record would allow for the contributor to check that their contribution was not misleading, or in contradiction to any formal written reports.

Perspective	Utility
	Current
Individual contributor to PCD	Record of contribution made and context of any comments Evidence of image data provided that informed discussion
Individual listener	Ability to expert advice given in PCD Facility to review any task assigned
Specialist in training	Bank of cases for educational purposes
Hospital	Record for individual patient's record Audit Development of improved practice guidelines
	Potential
Hospital	Automatic data collection for National Statistics Health Insurers Department of Health
MDTM	Facilitate real time review of similar case to current discussion

Table 1: Potential Utility for MDT meeting records

Individual Listeners could review a meeting record and the need for individual note taking, which might be a distraction at a meeting, would be obviated.

Specialists in training would be able to review a corpus of cases of a particular type in order to educated themselves in a particular type of problem.

Hospitals would reduce the risk of errors by being assured that decisions taken at MDTM were fully documented. Furthermore, the availability of MDTM records would improve current audit practices and provide useful data for the development of clinical practice guidelines. The MDTM is also a potentially a very useful forum for data gathering for National Registries and required by the Department of Health and other agencies.

MDTMs could potentially access prior similar cases to the case under discussion which would help in making the decision about the current case. It may be, for example, that a similar case had an unexpected outcome during treatment, which may moderate the treatment decision on the current patient. Having a corpus of PCDs together with follow-up data on the outcome of their treatment undertaken would provide evidence (data) for the development of clinical practice guidelines that would influence future decisions.

6. Discussion and Conclusion

The opportunities that a multimodal meeting record would provide to a multidisciplinary medical team are well recognised. However, difficulties are experienced in capturing such a record from technical and behavioural perspectives. The technical difficulties in making the recordings could be overcome through the use of a dedicated meeting room with suitable microphones and recording devices to capture the images used to inform the discussion. Difficulties however in maintaining data security, including respecting patient privacy and confidentiality, while making electronic records available in a hospital network poses a great challenge.

The skepticism demonstrated by medical specialists in the adoption of technology into the healthcare workflow has been documented. Staff experience of failed IT projects and lack of knowledge of the potential contribution that IT might bring to MDTMs are both significant and are not to be underestimated (Heeks, 2006; Southon et al., 1999). Establishing and maintaining trust between the researchers, developers and hospital staff is a key factor in the success of any study. Involving individuals from the group, inviting research mentors and providing frequent progress reports to the group was critical to undertaking this study. We believe that if the technical issues can be satisfactorily addressed and potential benefits demonstrated, then the development of multimodal meeting records will be found to directly improve patient care and make the health services more efficient in the long term.

Acknowledgements

We wish to thank the multidisciplinary teams at St. James's hospital Dublin for their co-operation in this on-going study. We especially thank the members of the lung MDT for facilitating this research, and our mentors Dr. F. O'Connell, Prof. K. O'Byrne, Prof. D. Hollywood and Mr.

M. Buckley. This work is funded under the IRCSET Enterprise Partnership scheme with St. James's hospital.

7. References

Amaya Becvar, James Hollan, and Edwin Hutchins. 2008. Representational gestures as cognitive artifacts for developing theories in a scientific laboratory. In *Resources, Co-Evolution and Artifacts: Theory in CSCW*, pages 117–143. Springer-Verlag, London.

Matt-M. Bouamrane and Saturnino Luz. 2007. Meeting browsing. *Multimedia Systems*, 12(4–5):439–457.

Kenneth Calman and Deirdre Hine. 1995. *A Policy Framework for Commissioning Cancer Services*. Department of Health, Welsh Office.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

S. Chen and P. Gopalakrishnan. 1998. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. of DARPA Broadcast News Transcription and Understanding Workshop*.

Richard Heeks. 2006. Health information systems: Failure, success and improvisation. *International Journal of Medical Informatics*, 75(2):125 – 137.

Bridget Kane and Saturnino Luz. 2006a. Multidisciplinary medical team meetings: An analysis of collaborative working with special attention to timing and teleconferencing. *Computer Supported Co-operative Work (CSCW)*, 15(5-6):501 – 535, December.

Bridget Kane and Saturnino Luz. 2006b. Probing the use and value of video for multi-disciplinary medical teams in teleconference. In *Proceedings of the 19th IEEE International Symposium on Computer-Based Medical Systems*, pages 518–523. IEEE Computer Society, July.

Bridget Kane and Saturnino Luz. 2009. Achieving diagnosis by consensus. *Computer Supported Cooperative Work (CSCW)*, 18(4):357 – 392, April. DOI:10.1007/s10606-009-9094-y.

Kai-Fu Lee, Hsiao-Wuen Hon, and Mei-Yuh Hwang. 1989. Recent progress in the sphinx speech recognition system. In *HLT '89: Proceedings of the workshop on Speech and Natural Language*, pages 125–130, Morristown, NJ, USA. Association for Computational Linguistics.

Saturnino Luz and Bridget Kane. 2009. Classification of patient case discussions through analysis of vocalisation graphs. In *Proceedings of the 11th International Conference on Multimodal Interfaces and Machine Learning for Multimodal Interaction (ICMI-MLMI'09)*, pages 107–114, New York, NY, USA. Association for Computing Machinery, ACM.

Saturnino Luz. 2009. Locating case discussion segments in recorded medical team meetings. In *SSCS '09: Proceedings of the ACM Multimedia Workshop on Searching Spontaneous Conversational Speech*, pages 21–30, Beijing, China, October. ACM Press.

R L Martino, K M Kempner, F S McGovern, D Chow, M E Steele, J E Elson, and C N Coleman. 2003. A collaborative telemedicine environment for the ireland - northern ireland - national cancer institute international partnership in cancer care. In *25th Annual International Conference of the IEEE EMBS*. IEEE Computer Society, Sept 17-21.

J McAleer, D O'Loan, and D Hollywood. 2001. Broadcast quality teleconferencing for oncology. *Oncologist*, 6(5):459–462.

MPI. 2005. ELAN: Eucido Linguistic Annotator. Max Planck Institute for Psycholinguistics, March. http://www.lat-mpi.eu/tools/elan/.

D.O. Olguin, B.N. Waber, Taemie Kim, A. Mohan, K. Ara, and A. Pentland. 2009. Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *IEEE Transactions on Systems, Man, and Cybernetics*, 39(1):43–55, February.

Sree Hari Krishnan Parthasarathi, Mathew Magimai.-Doss, Daniel Gatica-Perez, and Hervé Bourlard. 2009. Speaker change detection with privacy-preserving audio cues. In *ICMI-MLMI '09: Proceedings of the 2009 international conference on Multimodal interfaces*, pages 343–346, New York, NY, USA. ACM.

D.A. Reynolds and R.C. Rose. 1995. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83.

Gideon Schwarz. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Gray Southon, Chris Sauer, and Kit Dampney. 1999. Lessons from a failed information systems initiative: issues for complex organisations. *International Journal of Medical Informatics*, 55(1):33 – 46.

Jing Su, Bridget Kane, and Saturnino Luz. 2008. Automatic content segmentation of audio recordings at multidisciplinary medical team meetings. In *International Conference on Information Technology*, Gdansk, Poland.

J. Su, B. Kane, and S. Luz. 2010. Automatic meeting participant role detection by dialogue patterns. In *Proceedings of COST 2102 Int. School - Development of Multimodal Interfaces: Active Listening and Synchrony*, volume 5967.

The role of redundancy in the emergence of sign language prosody

Svetlana Dachkovsky
Sign Language Research Laboratory
University of Haifa, Israel
E-mail: dachkov@yahoo.com

Abstract

This paper will focus on the role of redundancy for the development and self-organization of a new language. The study underscores the importance of interchannel redundancy for the detection of prosodic constituents. Alignment of multichannel cues – facial intonation, manual rhythm, and head and body positions -- within prosodic constituents not only lends salience through redundancy to the signal, it also "creates" a system of discrete linguistic constituents. In spoken languages and in established sign languages, such as Israeli Sign Language, prosodic cues align with the text (i.e., with manual rhythmic cues), while affective cues do not (Dachkovsky 2005). Yet, in a new sign language formed in an insulated community, signal redundancy in the form of coordination and alignment of multiple prosodic features has not yet self-organized. These findings from joint work (Sandler et al, to appear) are analyzed here in light of the concept of redundancy, both in the phonetic signal and in higher linguistic structure. The findings run contrary to predictions of a theory such as the Smooth Signal Redundancy Hypothesis (Aylett and Turk 2004, 2006), which posits an inverse relation between redundancy in the linguistic structure and redundancy in the phonetic signal. Implementing a Redundancy Index developed for this study, I show that in the new sign language that is the object of this analysis, phonetic redundancy (at the level of sentence prosody) and language redundancy increase together across generations, and do not compensate for one another. The lack of balance between the two types of redundancy at the early stages of language development is explained by high levels of extra-linguistic redundancy initially, whose role in communication diminishes as the language matures.

1. Introduction

Although redundancy is agreed to be an essential feature of all biological systems, the role of redundancy for language structure and complexity is not self-evident. For example, Chomsky (1991) asserts that a lack of redundancy characterizes language and contributes to its elegance as a system. Yet he cannot point to any logical explanation for why language is so different from other biological systems on this view. At the same time other scholars emphasize the importance of redundancy as a mechanism that protects communication from noise and disruption.(Jackendoff and Pinker, 2005). The present paper will address this problem by examining the relevance of redundancy for language organization, in particular, for the development of a prosodic system. This will be done through the perspective of the Smooth Signal Redundancy hypothesis which predicts the inverse relationship between acoustic and language redundancy. This will be done by analyzing the emergence of prosodic redundancy in a nascent sign language.

Section 2 of this paper discusses the notion of redundancy as a feature characterizing cognitive systems in general and language in particular. The realization of redundancy in prosodic system of spoken languages is considered in Section 3. Section 4 reviews the system of prosodic cues in established sign languages. Section 5 summarizes results from a joint study on a new sign language, Al-Sayyid Bedouin Sign Language (Sandler et al, to appear.) That study demonstrated that this new language, which arose recently in an insular Bedouin village with little or no outside influence in the early stages (Sandler et al., 2005), shows evidence for gradual development of complexity in both prosody and syntax. Having described those results, we move on to the present analysis. Section 6 considers the attempt of the Smooth Signal Redundancy hypothesis to account for the complex interaction between acoustic (phonetic/ prosodic) and language (semantic, syntactic and pragmatic) redundancy in language, which can be explained by the drive for speakers to achieve robust information transfer in a potentially noisy environment while conserving effort. Section 7, the body of the present study, addresses the problem of the balance between the two types of redundancy in the two age groups of ABSL signers by introducing an index of prosodic redundancy and evaluating it against some aspects of language redundancy. The paper concludes that redundancy is crucial for the organization of a prosodic system. The numerous signals produced by various multi-channel articulators -- hands, facial expressions, and body movements -- contribute to the crystallization of prosodic constituents as discrete linguistic units. Yet, redundancy takes time to develop and is not present in a language at the outset.

2. Redundancy in language and beyond

Before we discuss the notion of redundancy in language, and more specifically, in prosody, a more general definition of redundancy will be briefly considered below.

2.1 The notion of redundancy

The concept of redundancy is often associated with the idea of superfluity, overabundance and surplus, being thus perceived negatively in its general meaning. Yet, redundancy is absolutely essential for all cognitive

systems, for example, in the visual perception of depth (Jackendoff and Pinker, 2005), where multiple mechanisms compute the same output – the relative distance of objects in the visual field.

2.2 Redundancy in language systems

In this sense, language follows the same organizational principle when it employs several devices to fulfill a particular linguistic function. Thus, we observe redundancy where more than one linguistic element plays the same role.

Redundancy in language can play different roles, all tending towards a better compromise between the speaker's and the listener's need for effective communication and intelligibility. It has been shown to be crucial for the facilitation of cognitive development and language acquisition. Bahrick (1992, 1994) proposes that redundancy across the senses captures infants' attention, and directs and constrains intermodal learning in the first months of life. For example, several studies have shown that infants detect the arbitrary relation between the visible object and the type of sound it produces when amodal information such as temporal synchrony between the visual and auditory signal, as well as rhythm and tempo are present (e.g., Bahrick, 1994). As far as the foundations of early lexical acquisition are concerned, 7-month-olds detect the arbitrary relations between simple speech sounds and objects when the visual and acoustic information is dynamic and temporally coordinated.

Redundancy works at all levels of language, from phonology to morphology and syntax. As an example of redundancy in morphology and syntax we can observe agreement which operates by copying grammatical features of a nominal argument, for example, on a verb.. Romance languages, for example, exhibit gender and number agreement for nouns, while Russian shows gender, number and case marking

At the level of phonetics, the redundancy of the acoustic signal makes speech more robust in a noisy environment. This means that speakers tend, for example, to hyperarticulate in an environment with a lot of auditory interference. All natural languages are subject to noise factors that are intrinsic to communication activities, and thus the redundancy of the whole communication cannot be removed without some loss in the outcome of the performance. The notion of redundancy as a language security mechanism becomes even more relevant when we think about prosody -- the language component which contributes to the parsing and chunking of speech into intelligible units.

3. Prosody and redundancy

By imposing rhythmic structure on the language stream, prosody signals the division of our utterances into interpretable pieces, or constituents. Intonation, or tune, is superimposed on these rhythmic constituents, in order to convey semantic, syntactic and pragmatic information, such as whether we are asserting or questioning.

Together, rhythmic and intonational structure also signals relations between constituents, as in the two clauses of the complex conditional sentence, *If it rains, the fireworks are off.* In addition, as demonstrated in Swerts and Krahmer's studies, seemingly extra-linguistic aspects, such as head movements and facial expressions that accompany speech, aid the comprehension of linguistic content and facilitate speech stream parsing and emphasis detection (e.g., Krahmer and Swerts, 2007). In sum, since the clear parsing of the speech stream is essential to efficient communication, multiple features, such as changes in duration, f0, amplitude and voice quality, as well as co-speech visual signals, cue the boundaries of prosodic chunks. The investigation of prosody in a physical modality other than auditory sheds light on the universal and essential properties of prosody. The next section will overview prosody research in sign languages, that transmit their signal in the visual modality.

4. Sign language prosody

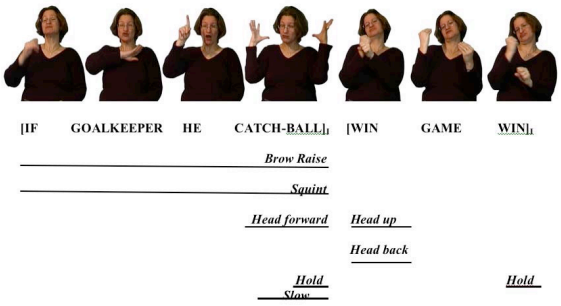
Sign languages, the natural languages that develop spontaneously in deaf communities, have grammatical organization, and many of their structural properties overlap with those of spoken languages (Sandler and Lillo-Martin, 2006). Once that has been established, it should not be surprising that the utterances of sign language have prosodic organization, devices for marking rhythm, stress, and the visual equivalent of intonation.

Sign languages are invaluable objects of linguistic investigation since they are the only languages that allow us to observe the way such a language system and its properties emerge. Not only are all known sign languages relatively young (under 300 years old), but some have arisen quite recently, and their development can be observed in real time**.** We start the discussion of the development of prosodic redundancy in more established sign languages by focusing particularly on Israeli Sign Language (ISL), which is the sign language used by most deaf people in Israel.

By studying the distribution of nonmanual markers together with that of rhythmic patterning in ISL, Nespor and Sandler (1999) developed a theory of sign language prosody according to which rhythmic constituency is demarcated by the hands, while the functions of intonation are manifested on the face. Body movements are also coordinated with prosodic constituents. For example, the boundaries of intonational phrases are marked by rhythmic manual cues as well two additional salient cues: change of head or body position, and across-the-board change in facial expression. The juxtapositon of manual rhythmic cues with face and body articulations and changes is what makes this boundary so salient.

Figure (1) below shows the juncture of two IPs in an ISL counterfactual conditional sentence meaning, 'If the goalie had caught the ball, they would have won the game' (Dachkovsky 2005; Dachkovsky and Sandler 2009). The contrastive change in head postures and facial expressions between the last sign of the first clause and the first sign of the second clause, as well as rhythmic

cues at the end of each intonational phrase make the



prosodic boundary very prominent.

Figure 1. The alignment of multiple prosodic cues in an ISL counterfactual conditional

Alignment of multichannel cues – facial intonation, manual rhythm, and head and body position -- within prosodic constituents not only lends salience through redundancy to the signal, it also "creates" a system of discrete linguistic constituents. In established sign languages, like ISL, linguistic facial expressions align with the text (i.e., with manual rhythmic cues), while affective or paralinguistic facial expressions do not (Dachkovsky 2005). Yet, as explained in Section 5, Sandler et al (to appear) found that coordination and alignment of multiple features is one of the properties of prosody that has not yet self-organized in the early stages of a new sign language arising in an insulated community. Investigating such languages allows researchers to ask the questions about the nature of the earliest kinds of structuring to arise in a human language. Here I address the following question in the context of a newly emerging sign language: Is redundancy present in the prosody of a human language from the very beginning? The discussion of these issues will be placed in the broader context of the Smooth Signal Redundancy Hypothesis in Section 7.

5. Prosody and Syntax in Al-Sayyid Bedouin Sign Language

Having demonstrated that sign languages have such properties as prosodic constituency signaled by multiple, redundant rhythmic and intonational cues, I move on to describe the new sign language, Al-Sayyid Bedouin Sign Language (ABSL), and to summarize our study of syntax and prosody in that language (Sandler et al, to appear), as context for the present analysis.

5.1 A new sign language and its community

The language is found in a Bedouin village in the south of Israel, where the presence of a gene for deafness and marriage patterns within the community have resulted in the birth of a proportionately large population of deaf people over the past 75 years -- ~ 150 out of ~ 4,000, almost fifty times the proportion of deaf people in the United States, for example. This sign language developed in relative isolation in this village, and today all deaf people and a large number of hearing people use it. The first generation of deaf people in the village was made up of four children born into a single family who used some form of home sign. The language of the second generation of deaf people (now from early 30s to over 50 in age) has robust word order patterns: SOV and Noun-Modifier (Sandler et al., 2005). Hardly any of the kinds of morphology found commonly in sign languages, such as verb agreement, rich aspectual morphology, and complex classifier predicates, have been found in the language (Aronoff et al. 2004). A new look at the development of prosody and syntax in this language (Sandler et al, to appear) is summarized here as the basis and background for the present redundancy analysis.

5.2 Gradual development of prosody and syntax in ABSL

This section will summarize the prosodic and syntactic differences found between older and the younger ABSL signers that are reported and analyzed in detail in Sandler et al. (to appear). The methodology of the study, its findings and their implications are briefly summarized here.

One minute of narrative was analyzed for each of four signers, two in their forties, and two 12-17 years younger. The two older subjects are O_s and O_T , and the younger subjects are Y_N and Y_A. O_S is a man about 40 years old at the time of recording, and O_T is a woman then aged 42. Y_N and Y_A are both women, aged about 28 and 25, respectively, when videotaped.

In attributing the linguistic differences to the differences in the development of the language, we adhere to Labov's Apparent Time approach (Labov, 1963, 1966), supported by his study of New York English. Under this approach, the regular increase in the use of a particular language feature across the age cohorts would represent a generational change in process. This approach reasonably assumes that each generation of language users reflects the state of the language when they acquired it as children, relying on the assumption that people do not significantly alter their language structure over their adult lifetimes.

The transcription and coding of the data was performed with the ELAN program. The data was coded and organized in the following categories of ELAN tiers: 1) glosses; 2) manual rhythmical features.; 3) facial actions.; 4) head positions; 5) torso positions and movements, and 6) blinks.The boundaries of intonational phrases in the coding were established in accordance with the presence of the manual rhythmic cues, such as holds, pauses, reduplications of sign movements, or sign movements prolonged through slower signing and exaggerated sign size. The manual signals were considered decisive in the prosodic parsing, since they constitute the main channel of sign language word production. On the ELAN annotation tiers, the onset and offset of all manual and nonmanual articulations are marked in relation to the time code of the recorded media, as shown in Appendix 1 for the younger signer O_S and Appendix 2 for the older signer Y_A. In a similar vein, in order to examine the development of syntactic marking in ABSL, several syntactic parameters -- the number and content of noun phrases, their ratio to the predicates, the number of pronouns – were coded and

analyzed. The findings from Sandler et al.'s study relevant for the discussion of redundancy will be summarized below.

The analysis of the development of ABSL prosody and syntax revealed that while the older signers do separate prosodic constituents rhythmically, they tend to use few boundary-aligned prosodic cues – the average of 2.3 and 2.7 cues per intonational phrase, whereas for the younger ones the number is much higher – 3.7 and 4. The disparity between the older and younger signers is especially vivid at the level of linguistic intonation (facial expression): while the older ABSL users practically do not use any linguistic intonation (facial expression), the younger signers use plentiful linguistic facial expressions – 30 and 18 percent of their intonational phrases are marked with linguistic intonation.

As far as the syntactic characteristics of the narratives are concerned, third person pronouns do not appear in the older signers' narratives at all. Along with the absence of third person pronouns, Y_S and Y_T's signing is characterized by a low number of noun phrases in relation to the number of predicates: on average, one NP is associated with 2.25 or 3 predicates. This can imply that in the narratives of the older signers, there are many predicates that are not overtly associated with any arguments. In contrast with the older group, ABSL signers only 15-20 years their junior use the same number of noun phrases as the number of predicates. In other words, unlike the older signers, their sentences tend to have overt subjects, often in the form of pronouns, including third person pronouns, the latter completely absent from the older signers' narrative in our corpus. In sum, this study convincingly demonstrated that both syntactic and prosodic complexity in a language develop gradually and in tandem.

These two stages of the language development were exemplified with two short segments, one produced by older second-generation signer O_S (Figure 2), and the other by younger signer Y_A (Figure 3).

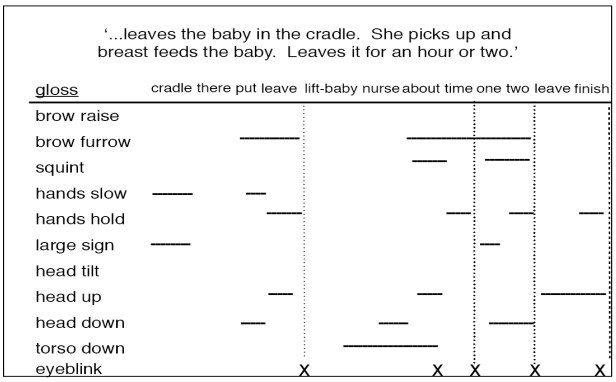


Figure 2. A stretch of signing produced by an older signer.

This segment of O_S's narrative consists of four IPs, marked by holds and pauses of the hands. All the final prosodic boundaries are marked by blinks. The final boundary of the first intonational phrase is also associated with a downward head movement. There are no facial expressions or additional head movements that align with the rhythmic boundaries of the prosodic units. The distribution of number of prosodic cues characterizing each constituent in this stretch is 3-2-2-2, with a mean of 2.2 boundary-aligned cues per intonational phrase in this segment. At the same time, a significant number of facial and head cues in this stretch are not aligned with the manually determined rhythmic boundaries. Their appearance is governed by extralinguistic rather than by linguistic factors. For example, sign PUT (baby) has a slower tempo, which mimics the motherly manner of placing a baby in a crib. The mimetic rhythm modifications disrupt the general rhythm of the intonational phrases. In addition to creating rhythmic "noise", mimetic signs like PUT create intonational "noise" as well, since the accompanying face and body movements are governed by mimetic or emotional rather than linguistic factors. In these ways, pantomimic signs interrupt the general prosodic pattern of the utterance.

The example in Figure (3) presents a short stretch of discourse signed by a younger person, Y_A. The rhythmic markings and the alignment of prosodic cues look very different from that in the previous example.

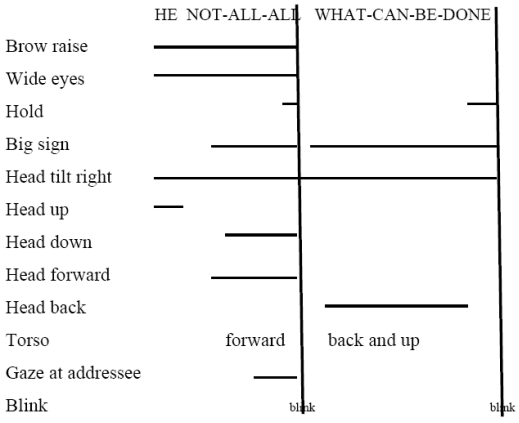


Figure 3. A coded example of a conditional produced by a younger signer.

In contrast with the previous example, in Figure (3) the boundary between intonational pharses [HE NOT-AT-ALL] and [NOTHING-CAN-BE-DONE] can be easily detected. The right edge rhythmic boundaries of each intonational phrase are signaled by several perceptually clear manual cues. The final sign in each prosodic unit is lengthened by holds, and, in addition, is judged to be larger than a citation form of the sign. The number of non-manual prosodic cues and their alignment in this utterance are similarly indicative of the signing of the younger pair, in contrast with that of the older signers. In sum, the final boundary of the first intonational phrase is associated with eight prosodic cues – raised brows, widely open eyes, large sign with hold, head down, head and torso forward, prolonged gaze at the addressee, and blink. The final boundary of the second phrase is marked by five prosodic signals – hold and large size on the last sign, head and torso tilted back, and blink, yielding a mean number of about 6 cues per intonational phrase in this utterance.

The findings of Sandler et al.'s study, while shedding light on the principal problems of language emergence, raise other intriguing questions, which are no less important than the answers. For example: What is the nature of this "partnership" between prosody and syntax in the development of a new language? What principle(s) lie(s) at the basis of their interaction? The present study will extend the earlier study by trying to address these issues through the notion of redundancy in language. The perspective taken by The Smooth Signal Hypothesis on redundancy will help us to interpret prosodic aspects of acoustic redundancy and their interaction with language redundancy as it is realized in syntax.

6 The Smooth Signal Redundancy Hypothesis

An attempt to show the interaction between acoustic redundancy and language redundancy is made by Aylett and Turk in their Smooth Signal Redundancy Hypothesis (2004, 2006). Their work demonstrates that the language redundancy of a phoneme, a syllable or a word has a strong inverse relationship with their acoustic redundancy. Language redundancy is measured by the predictability of a linguistic unit given its context and inherent frequency (i.e., highly predictable from lexical, syntactic, semantic, and pragmatic factors, Jurafsky et al., 2001). Acoustic redundancy is understood as the predictability of acoustic signals, such as duration or spectral effects of a vowel. For example, a longer syllable is more acoustically redundant than a shorter syllable, or the more distinctive a vowel, the more acoustic information can be said to be present. For example, taking the utterance “I’m going to the beach”, the word 'to' is more likely to occur than the word 'beach', therefore, the language redundancy of the former is higher than the language of the latter. In addition, the amount of acoustic information both in terms of duration and in terms of spectral distinctiveness is lower in 'to' than in 'beach'. This means that, vice versa, the acoustic redundancy of 'to' is higher than that of 'beach'. Aylett and Turk claim that the overall signal probability is smoothed by this inverse relationship. This smoothness adds robustness because the information content is spread more evenly across the signal. This inverse correlation is hypothesized to make speech more robust in a noisy environment.

This proposal can be explained functionally within an information theoretical framework, by the drive for speakers to achieve robust information transfer in a potentially noisy environment while conserving effort (Lindblom, 1990). These pressures encourage speakers to produce utterances whose elements have similar probabilities of recognition, that is, utterances with what the authors call a smooth signal redundancy profile. Aylett and Turk’s study shows that phrase-medial syllables with high language redundancy were shorter than less predictable elements. In addition, the Smooth Signal Redundancy Hypothesis (SSRH) makes a claim that redundancy is implemented through prosodic structure, since variability in length and spectral characteristics was strongly related to variability in prosodic prominence. Their 2004 and 2006 studies presents data based on half million syllables of English Rhetorical Corpus, and provides a convincing argument for the relation between redundancy in language structure and redundancy in the acoustic signal. For this reason, it offers a good starting point for investigating an emergence in a new language. Since prosody interacts with syntactic, semantic and pragmatic levels of language and is at the same time realized at the acoustic level, it can be thought of as an interface between language redundancy and acoustic redundancy.

As stated above, the present study investigates the interaction between redundancy and prosody in a new sign language. New sign languages that exist in communities of deaf people give us an opportunity to investigate the interaction of language and signal redundancy in the earliest stages of language emergence. An investigation of this kind on a spoken language is not possible, as all spoken languages are many thousands of years old or descended from old languages. Therefore, sign languages offer the only opportunity to watch language develop from the beginning. (Senghas & Coppola, 2001; Sandler et al., 2005 (PNAS); Aronoff et al., 2008). In the continuation of the paper, the results from the previous section are analyzed in the context of redundancy introduced above. The analysis offers an algorithm or Redundancy Index for determining the amount of redundancy in prosodic signals, and reveals the way in which signal redundancy emerges over time. It is suggested that at the initial stages of the language development lower levels of linguistic redundancy are compensated for by higher levels of extra-linguistic redundancy. Yet, as the evidence shows, in the progress of the language development the linguistic aspects of redundancy take over extra-linguistic ones, and initiate their own "game".

7. Redundancy in a new language

As explained in the previous section, the SSRH proposes that the language redundancy of a linguistic unit has a strong inverse relationship with its acoustic redundancy in an utterance or intonational phrase. This relation can be represented as the balance between two parts of the scales: the higher the acoustic signal is on the one scale, the lower its predictability from semantic, pragmatic and syntactic contexts. The authors demonstrate that this is a normal and optimal state of affairs in an established language, where the speaker strives to conserve his/ her effort while at the same time not compromising the efficiency and robustness of communication.

How will this relationship look in a language at its inception? One possibility is that the relation is the same as for older languages. Yet a second hypothesis presents itself as well. It may be that while a language is fragile and variable at the very start, its speakers will make an effort to keep both sides of the redundancy scales high in order to ensure communication with minimal disruption. The latter hypothesis gains extra credibility in the case of a sign language. In contrast with spoken languages, which can augment auditory signals with visual ones, sign languages have only visual signals at their disposal. This might mean that users of newly developing sign languages from the very start would enhance both "acoustic" – in

this case, visual -- and language redundancy to protect their signing from disfluencies and visual noise. As we will see, applying the Redundancy Index to our data yields different results depending on the stages of the language development.

7.1 Calculating the index of Prosodic Redundancy

In order to evaluate these hypotheses, the present study interprets the findings regarding prosody and syntax in the new sign language in the context of redundancy. First of all, an attempt will be made to quantify prosodic redundancy as a manifestation of aspects of visual/ phonetic redundancy. Prosodic redundancy will be calculated as the ratio between prosodic cues that align with manual rhythmic cues and those that do not align, to provide a Redundancy Index (RI). For example, if there are four prosodic cues that align with prosodic boundaries as determined by the manual behavior, whereas there is only one cue that didn't align with them, the RI of prosodic redundancy will be 3. This index shows to what extent the cues grouped to signal the boundaries of a linguistic constituent are stronger that those disfluent cues that disrupt its unity and dissolve its borders. The assumption is that the greater the number of prosodic cues aligned with a rhythmic boundary, the more salient the boundary. If these aligned prosodic cues are produced by different articulators – face, head and torso, the effect is even stronger. In addition, the number of mimetic signs produced by the two age groups of signers will be calculated, since mimetic signs were found to disrupt the general rhythm of the utterances. Therefore, they are one of the factors influencing the proportion of non-aligned cues in the signing samples. By comparing the average number of aligned cues from Sandler et al.'s study (to appear) with the average number of nonaligned cues, Table 2 provides an index of prosodic redundancy. In Section 7.2., we will compare this index with the amount of syntactic redundancy manifested in the findings on the number and content of noun phrases in Section 5 above. This way signal redundancy will be compared to language redundancy.

Signer & age	Avg. # aligned cues	Avg. # non-aligned prosodic cues	Index of prosodic redundancy
Os (40)	2.3	3.7	-1.4
OT (42)	2.7	0.7	2.0
YN (28)	3.7	0.7	3
YA (25)	4	1.1	2.9

Table 1. Prosodic redundancy as ratio between aligned and non-aligned cues

As explained above, the index of prosodic redundancy will show us to what extent a prosodic constituent as a discrete unit with clear boundaries is protected from the disruption by visual disfluencies not related to its inherent structure. On average, the relation between the older and younger signers for the non-aligned cues is inversed. So, the older signers' narratives display 3.7 and 0.7 non-aligned prosodic cues per intonational phrase, whereas the younger signers' numbers are lower on average – 0.7 and 1.1. As far as the index of prosodic redundancy is concerned, for older signer O_T it is computed as 2.7 (aligned cues) - 0.7 (non-aligned cues) = 2, whereas the same index for younger singer O_N will be 3.7 – 0.7 = 3. This means that the relative strength and prominence of the prosodic boundaries for younger signer O_N are higher. It is worth emphasizing that for older signer O_S this ratio between the aligned and non-aligned prosodic cues is even negative -- -1.4. This can mean that the prosodic boundaries are almost not distinguishable and non-discernable due to the high number of non-aligned cues. Partly, the disruption of prosodic organization is the result of mimesis. As explained in Section 5, the signing of the two older signers includes quite a number of signs with pantomimic elements, while these are completely absent from the narratives of the two younger signers. In sum, the number of the non-aligned cues in the older signers' stretch of narrative outweighs the number of the aligned cues, thereby making the prosodic boundaries barely discernible, and sometimes indiscernible. In addition, in many cases the low degree of prosodic redundancy due the high number of non-aligned cues is undermined even more by unclear and gradient character of the manual cues that are supposed to be the principal markers of prosodic constituents. For example, in Figure (2) above the hold after LEAVE, which, together with the meaning, groups it with PUT and separates it from PICK-UP-BABY is very short and unclear. All these make the demarcation line between the prosodic constituents very unclear as well, both because the signals are not salient and because manual rhythm and non-manual articualtions are not coordinated. This is comparable to an orchestra where each instrument, though playing the same melody, starts and proceeds at its own tempo, not constrained by the conductor's directions. As a result, the overall melody cannot be picked out from the cacophony of sounds.

7.2 Indications of ABSL language redundancy

Overt syntactic markers such as pronouns, as well as the ratio between noun phrases and predicates, can signal language redundancy in the sense of Aylett and Turk, and will be considered in this analysis as crucial aspects of language redundancy. Third person pronouns are redundant language elements by definition since they do not have independent lexical content but function to establish co-reference with referents mentioned previously in the discourse. Therefore, their absence in the older signers' discourse underscores the overall lack of redundancy in their language. On the other hand, third person pronouns are abundant in the younger signers' narratives -- 11 and 18 tokens were found in the data, as reported above.

The low number of noun phrases in comparison with the number of predicates in older signers' data implies that in there are many predicates that are not overtly associated with any arguments. The deficit of noun phrases hampers efficient determination of the discourse participants. This

indeterminacy of argument roles decreases the language redundancy of the whole message. In contrast, in the two younger signers, one noun phrase corresponds to approximately one predicate, which means that thematic roles of the propositions are explicitly marked, and do not need to be inferred by interlocutors from the context**.** For example, in Figure 3 above, the first intonational phrase contains third person pronoun HE which is co-referential with the argument in the preceding discourse, and clearly indicates who is the subject of the negative speech act. The predicate of the other intonational phrase WHAT-CAN-BE-DONE is not associated with any noun phrase, but its denotation is rather impersonal.

7.3 General discussion of ABSL signal redundancy

Through careful analysis of prosodic cues and their organization, we are able to see clear differences in the prosodic redundancy of the older and younger pairs of signers. There seem to be several reasons for these differences. One is the fact that younger signers are more likely to use linguistic facial expressions, thereby increasing the number of prosodic cues overall. The other is a general disruption of prosodic organization in the older signers, as clear from the numbers of the non-aligned cues. Partly it is accounted by a high number of mimetic signs. Another reason for a high number of alignment mismatches in the older group is the type and distribution of rhythmic cues marking prosodic boundaries. Some cues are clearer than others, and the younger signers favor the clearer cues in combination with those that are more gradient. For example, holds and pauses are clear signals of prosodic phrasing, whereas slowing down of the signing tempo or increasing the size of a sign is more gradient and harder to perceive. As we have shown in Figure 2, the latter two cues are also used to mimic the slow speed of an action or its extended size in reality. Along with the scarcity of aligned manual and nonmanual cues, such size and rhythm indeterminacies create ambiguity and opacity in the signal, contributing to a general lack of consistent and easily perceptible "anchors" for boundary detection. This all leads to a lower index of prosodic redundancy in the older group of signers. Contrary to the Smooth Signal Redundancy Hypothesis, which predicts that in fully developed languages lower degrees of acoustic (prosodic) redundancy should be compensated for by higher degrees of language redundancy, the results obtained in this study show that in a nascent language both types of signal redundancy are low. For instance, in Figure 2 the low prosodic redundancy is not compensated for by language redundancy, since many predicates, like LEAVE, NURSE, LEAVE, FINISH, are not associated with any arguments. Actually, none of the predicates produced in this passage is overtly associated with its doer. As a result, the interlocutor has to infer who did what to whom from contextual hints rather than from explicit language cues. Therefore, both the content and the boundaries of proposition are left vague.

It takes at least a generation to develop a redundant and "disruption-proof" grammar which is able to neutralize natural signal disfluencies and inaccuracies by redundant prosodic signals clustering together to mark prosodic constituents, as well as clear semantic, syntactic and pragmatic cues that would assign explicit semantic roles to the referents in propositions, and signal grammatical relations between the propositions.

Nevertheless, the older signers are able to understand each other and to get their message across among their peers. No effective communication can do without redundancy, and a closer look reveals some particular types of redundancy in the narratives of the older generation of signers. Signs that involve mimetic elements are redundant in a sense that they usually involve big and slow or reduplicated movements. Along with the enhanced manual aspects, the sign production engages the whole body and head postures, as well as numerous facial expressions. In addition to the redundancy of mimetic signs, the older signers' communication is accompanied by what can be called the redundancy of a self-conatined and isolated speech community, where the number of people living in the same communicative context is very small, so that everybody shares a great deal of information across the community. It is important to note that these two types of redundancy are extra-linguistic in nature. Therefore, they do not contribute to the robustness and richness of the linguistic system, but they facilitate efficient communication as an interaction of purely linguistic and extra-linguistic factors, and explain the relative ease with which the members of the older generation appear to understand each other.

8. Conclusion

By closely observing a new language as it emerges, we are able to see the stages of a linguistic system as it evolves. Analyzing the stages of this process in detail, as we have done here, reveals properties of our language capacity that are usually inaccessible, for example, the emergence of redundancy in a language system. A fundamental result of this study is the finding that redundancy is crucial for the organization and development of language in general and prosody in particular, since the latter breaks up thought units rhythmically before there is any other structure to speak of. Yet, linguistic differences between the age groups convincingly demonstrate that even such a seemingly universal and automatic matter as the accumulation of multiple rhythmic and intonational cues at prosodically meaningful points in the language stream is not present from the very start of a language, but rather takes time to be developed and elaborated. For the older second generation signers of this study, we see an attempt to mark thought units by manual rhythmic markers and the sporadic addition of other cues, such as occasional facial expressions and body movements. For those signers, these scarce, minimal elements are not yet coordinated in such a way as to clearly cue prosodic constituents. The unclear, barely distinguishable prosodic structure is constantly disrupted by mimetic and extra-linguistic visual "noise". The lack of prosodic redundancy is not compensated for by higher degrees of language redundancy, as would be expected in established languages according to the Smooth Signal Redundancy Hypothesis. Apparently,

extra-linguistic cues are relied upon in order to compensate for the low levels of linguistic redundancy.

Analyzing ABSL in signers twenty years younger, we can see important differences. The various and numerous signals produced by different articulators -- hands, face, and body, are all coordinated and temporally aligned at the prosodic boundaries, thus so that a prosodic constituent crystallizes as a discrete linguistic unit. It seems that, as noted Bahrick (1992, 1994) for language acquisition, dynamic redundant multi-channel cues that are temporally coordinated captures attention more efficiently, directs and constrains the detection of discrete linguistic structure. The application of ELAN as a coding and analyzing device facilitated the capturing of both qualitative and quantitative features of various visual prosodic cues produced by numerous articulators. Higher levels of prosodic redundancy are supported by gradually increasing levels of language redundancy.

In sum, contrary to our initial hypotheses, at the very early stages of ABSL development, the interplay between different types of signal redundancy is between "acoustic" redundancy and language redundancy, but rather between linguistic redundancy and extra-linguistic redundancy. The higher levels of extra-linguistic redundancy compensate for lower levels of linguistic redundancy. With the language maturation the role of extra-linguistic factors seems to drop, and then acoustic and language redundancies start striving to reach their equilibrium. If the SSRH is correct, the two types of redundancy are expected to assume a more complementary and optimal relationship as the language matures.

There are two main messages to be learned from the present study. First, we understand that even such a seemingly low-level language property as signal redundancy does not appear at the moment of the inception of language, although the language functions effectively within a community. The second message is that both language and phonetic redundancy are crucial for the development of language complexity and maturation.

References

Aronoff, M., Meir, I., Padden, C. and Sandler, W. 2004. Morphological universals and the sign language type. Yearbook of Morphology 2004, 19-39.

Aronoff, M., Meir, I., Padden, C. and Sandler, W. 2008. The roots of linguistic organization in a new language. Interaction Studies 9, 133-153.

Aylett, M. and Turk, A. 2004. The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. Language and Speech 41, 31-57.

Aylett, M. and Turk, A. 2006. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. Journal of Acoustic Society of America 199, 3048-3058.

Bahrick, L. E. 1992. Infants' perceptual differentiation of amodal and modality-specific audiovisual relations. *Journal of Experimental Child Psychology 53*, 180-199.

Senhgas, A. and Coppola, M. 2001. Children Creating Language: How Nicaraguan Sign Language Acquired a Spatial Grammar

Bahrick, L. E. 1994. The development of infants' sensitivity to arbitrary intennodal relations. *Ecological Psychology 6(2),* 111-123.

Chiari, I. 2007. Redundancy elimination: The case of artificial languages. *Journal of Universal Language 8*, 7-38.

Dachkovsky, S. 2005. Facial Expression as Intonation in Israeli Sign Language: The Case of Conditionals. MA Thesis, University of Haifa.

Dachkovsky, S., Sandler, W. 2009. Visual intonation in the prosody of a sign language. Language and Speech 52, 287-314.

Hauser, M., Chomsky, N., and Fitch, T. 2002. The faculty of language: What is it, who has it, and how did it evolve? Science vol. 298, pp. 1569-1578.

Jackendoff, R. and Pinker, 2005. The nature of the language faculty and its implications for evolution of language (Reply to Fitch, Hauser, and Chomsky). Cognition 97, 211-225.

Jusczyk, P. W. et al. 1992. Perception of acoustic correlates of major phrasal units by young infants. Cognitive Psychology 24, 252-293.

Krahmer, E. and Swerts. M. 2007. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. Journal of Memory and Language 57, 396-414.

Labov, W. 1963. The social motivation of a sound change. Word 19, 273-307.

Labov, W. 1966. The Social Stratification of English in New York City. Center for Applied Linguistics, Washington.

Meir, I., Sandler, W. 2008. A Language in Space: The Story of Israeli Sign Language. Lawrence Erlbaum Associates, New York.

Nespor, M., Sandler, W. 1999. Prosody in Israeli Sign Language. Language and Speech 42, 143- 176.

Padden, C. A., Meir, I., Sandler, W. and Aronoff, M. 2005. Against all expectations: Encoding subjects and objects in a new language. In D. Gerdts, J. Moore and M. Polinsky (Ed.), *Hypothesis A/Hypothesis B*. Cambridge, MA: MIT Press.

Padden, C. Meir, I., Aronoff, M., Sandler, W. The grammar of space in two sign languages. 2006. In D. Brentari (Ed.), *Sign languages: A Cambridge language survey*. Cambridge, UK: Cambridge University Press.

Sandler, W. 1999a. Prosody in two natural language modalities. Language and Speech 42, 127-142.

Sandler, W. 2009. Symbiotic symbolization by hand and mouth in sign language. Semiotica 174, 241-275.

Sandler, W. and Lillo-Martin, D. 2006. Sign Languages and Linguistic Universals. Cambridge: Cambridge University Press.

Sandler, W., Meir, I., Padden, C. and Aronoff, M. 2005. The emergence of grammar: Systematic structure in a new language. Proceedings of the National Academy of Sciences 102, 2661-2665.

Sandler, W., Dachkovsky, S., Meir, I., Aronoff, M. and Padden, C. (To appear). The emergence of complexity in prosody and syntax. *Lingua.*

Automatic face tracking in Anvil

Bart Jongejan
University of Copenhagen, Center for Language Technology (CST)
Njalsgade 140, 2300 København S, Denmark
E-mail: bartj@hum.ku.dk

Abstract

We describe the current state of our work on integration of the OpenCV face tracking functionality into Anvil as a Java plug-in. Our goal is to save time for the annotator by automatically annotating salient head movements such as nodding and shaking. Problems due to limitations to the availability of the OpenCV functionality from Java are discussed.

1. Introduction

In the construction of a multimodal corpus the manual annotation of head movements is a very time consuming task. Moreover, the onset and end of head movements are not as sharply defined as e.g. speech utterances, and therefore manual head movement annotations may suffer from personal preferences and shifting states of concentration of the annotator. For these reasons, automatic detection of head movements has the potential of making the annotation process swifter and less prone to personal choices.

2. Augmenting Anvil with OpenCV

Anvil (Kipp 2008) is a generic annotation tool for multimodal dialogue. The user interface consists of several windows, the most central of which are the video window and the annotation board. The latter consists of several tracks that the user can specify. The video window has an overlay component that can be used to visually mark anything of interest. Anvil has no built-in means of automated annotation, but the software's functionality can be enhanced by means of plug-ins. The Anvil software is written in Java, and so are any plug-ins the user wants to employ. The author of Anvil gave us the opportunity to use the coming version of the software, 5.0.

OpenCV (Open Source Computer Vision) (Bradski & Kaehler 2008) is a library of programming functions for real time computer vision, and includes ready-to-use methods for face tracking using Haar wavelets. OpenCV is coded in C++.

Bringing the power of OpenCV into Anvil is not straightforward, because Java and C++ do not easily integrate. There has been an attempt in Michael Kipp's group to integrate the OpenCV software into Anvil (Kipp, personal communication).

Nowadays, the Java world of Anvil and the C++ world of OpenCV can be combined with an additional piece of software, the OpenCV Processing and Java Library (OPJL from hereon). The OPJL is a project of the Atelier Hypermédia at the École Supérieure d'Art in Aix-en-Provence, France. The OPJL only supports part of the full OpenCV functionality. Most importantly, this library enables the Java programmer to use OpenCV's face tracking abilities. However, algorithms for detecting changes in face orientation are not accessible using this library. For an example of using OpenCV's Lucas-Kanade algorithm for tracking this kind of face movement, see Moubayed et al (2009).

3. Method

We have written an Anvil plug-in in Java that uses the OPJL, which itself is written in Java and therefore is easily incorporated into the plug-in. OpenCV can analyse a video stream from a video camera or from a video file, but it can also analyse single video frames. Analysing video frames with a face tracking algorithm is very CPU intensive, but luckily OpenCV's face tracking software does not require the highest possible pixel resolution – the output from a simple webcam is sufficient. On the downside, Anvil does not give direct access to the current video frame, leaving us no other choice than to introduce an intermediate step in which the contents of Anvil's video window are captured using a standard Java method and then stored in memory. In a second intermediate process each captured image is copied to an OpenCV buffer. Because the input to OpenCV originates from Anvil's video window, the resolution and colour depth of the image depends on the size of this video window. If the user resizes the video window, the face tracking analysis can be considerably speeded up or slowed down.

We use the OpenCV Haar-based routine for frontal face detection. The routine can detect many faces in the same image. Each detected face is expressed by a square that tightly fits the area covering eyes, nose and mouth. To save processing resources we try to analyse only one face at a time. The software follows a single face by zooming in on a region of the visual field where a face was detected in an earlier frame in the video stream. If, contrary to expectation, the face tracking algorithm does not detect a face in the restricted area, the software zooms out before analysing the next frame. The failure to detect a face can have several reasons, of which the absence of a face in the analysed area is an obvious one. Other reasons are that the face is turned away from the camera, that part of the face is covered by a hand or by long hair, or that the face is rotated away from a horizontal posture.

It is a problem that the OPJL does not include routines for the detection of changes of face orientation. We try to

compensate for this defect by analysing the change in time of the head positions. The hypothesis is that e.g. a head nod not only changes the orientation of a face, but also its lateral and vertical position.

The most straightforward way to "annotate" head movements is to just record the horizontal and vertical coordinates of the detected face in an annotation element. This results in a non-distinct sequence of annotation elements. Slightly more informative is to also record whether the face is moving and if so, in which direction and with which velocity. In this way, the user would be able to quickly find all the episodes where a face is moving, say, leftwards at a moderate speed. Track elements annotating face movements may stretch over more frames than elements annotating face positions, because whereas face positions change during a head movement, the velocity and direction of the face movement may remain stable during many frames. Still, a face movement does not tell very much. A face can be moving left because the person nods or because the person moves the whole body. By taking the second time derivative of the face position - the acceleration and deceleration of the face movement - we get somewhat closer to a meaningful automatic categorisation of face movement. See Erdem & Sclaroff (2002) and Chippendale (2006). A face that is quickly gaining or losing velocity is probably indicative of a head movement caused by nodding or head turning, while we expect low acceleration and deceleration rates to be possibly connected to bodily movements. One may continue using third and higher time derivatives of the position. A high third derivative, for example, is indicative of two consecutive movements in opposite directions – a head nod or a head shake, for example. As we analyse higher derivatives, more and more data points are needed, stretching the annotation-worthy phenomena over longer spans of time. This technique is not unlike a Fourier analysis or a Haar wavelet based approach to face movement in the temporal domain. See Moubayed et al (2009) for an application of Haar wavelets to face movement detection.

We have succeeded in extracting an acceleration vector that seems to be quite well correlated with head nods and shakes. Each vector is based on a sequence of face positions of three or more frames – the optimum number is still an object of investigation. The mean acceleration of the head movement during the time lapse covering the analysed frames is computed from the change in head velocity during this time lapse. The velocities, in turn, are measured in terms of differences in head positions between two consecutive frames.

The Anvil tool allows the user to play the video at slower frame rates than normal. Slowing down the frame rate causes the face tracking to take place in frames that are closer together in time. At normal frame rate, many frames are skipped between face tracking analyses, because the hardware cannot catch up. However, the user gets some feeling for the amount of skipped frames, because the face tracking processes causes the video to somewhat stagger. Slowing down the frame rate will allow the software to detect finer details of the head movements, but may on the other hand cause the software to miss some gross movements that take place over a longer stretch of time and more slowly.

According to Matsusaka et al (2009) we need to annotate every frame in a video sequence. We think the optimal number of annotations per second depends on one's purpose. Coarse graining the observation of bodily movements may bring phenomena to light that are not seen with fine-grained observations.

4. Results

This work is still very much in progress. There seems to be useful information in the detection of head acceleration, both its magnitude and its direction. Regarding the latter, improvements could be gained from using more than one camera directed at the same person, so that a truly 3-D acceleration vector could be established.

The hardware we used is a Dell Precision PW 380 with an Intel Pentium 4 CPU running at 3.79 GHz. The machine has 2 GB of RAM and runs Windows XP SP3.

5. Conclusion

With the restricted functionality of the Java port of OpenCV it is possible to automatically detect gross head movements. It is hoped that peaks in head acceleration data are indicative of head nods, shakes and other movements that leave the lower body at rest. We will test the new Anvil plug-in during the annotation of the corpus that is being built in the NOMCO project.

6. Acknowledgments

This research has been supported by the Danish Council for Independent Research in the Humanities and by the NOMCO project (http://sskkii.gu.se/nomco/), a collaborative Nordic project with participating research groups at the universities of Gothenburg, Copenhagen and Helsinki. The project is funded by the NOS-HS NORDCORP programme under the Danish Agency for Science, Technology and Innovation.

7. References

Bradski, G.; Kaehler, A. (2008). *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly

Chippendale, P. (2006). Towards Automatic Body Language Annotation. In *Proceedings of the 7th Int. Conference on Automatic Face and Gesture Recognition*, Southampton, UK: IEEE, pp. 487–492 (2006).

Erdem, U. M.; Sclaroff, S (2002). Automatic detection of relevant head gestures in American sign language communication. In *Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02)*, Quebec City, QC, Canada, vol. 1, pp. 10460

Kipp, M. (2008). Spatiotemporal Coding in ANVIL. In *Proceedings of the 6th international conference on Language Resources and Evaluation (LREC-08)*

Matsusaka, Y; Katagiri, Y; Ishizaki, M, Enomoto, M (2009). Unsupervised Clustering in Multiparty Meeting Analysis. In M. Kipp, J.-C., Martin, P. Paggio & D. Heylen (Eds.), *Multimodal Corpora. From Models of Natural Intercation to Systems and Applications*, Berlin Heidelberg: Springer Verlag, pp. 93-108

Moubayed, S. A.; Baklouti, M.; Chetouani, M.; Dutoit, T.; Mahdhaoui, A.; Martin, J.-C.; Ondás, S.; Pelachaud, C.; Urbain, J. & Yilmaz, M. (2009). Generating Robot/Agent backchannels during a storytelling experiment. In *Proceedings of the 2009 IEEE international conference on Robotics and Automation*, Kobe, Japan: IEEE Press, NJ, USA, pp. 2477-2482

InSight Interaction

A multimodal and multifocal dialogue corpus

Geert Brône[1], Bert Oben[2], Kurt Feyaerts[2]
[1]Department of Applied Linguistics, Lessius University College, Antwerp (Belgium)
[2]Department of Linguistics, University of Leuven (Belgium)
E-mail: geert.brone@arts.kuleuven.be, bert.oben@arts.kuleuven.be, kurt.feyaerts@arts.kuleuven.be

Abstract

Research on the multimodal aspects of interactional language use requires high-quality multimodal resources. In contrast to the vast amount of available written language corpora and collections of transcribed spoken language, truly multimodal corpora including visual as well as auditory data are notoriously scarce. In the present project, we propose two significant extensions to the few notable exceptions that do provide high-quality and multiple-angle video recordings of face-to-face conversations. First, the recording set-up was designed in such a way as to have a full view of the dialogue partners' gestural behaviour, including facial expressions, hand gestures and body posture. Second, by recording the interlocutors' internal perspective and behaviour during conversation, using head-mounted scene cameras and eye-trackers, we obtained a 3D landscape of the conversation, with detailed production (scene camera and sound) and processing (eye movement for gaze analysis) information for both participants. In its current form, the resulting *Insight Interaction Corpus* consists of 15 recorded face-to-face interactions of 30 minutes each, of which 5 have been transcribed and annotated for a range of linguistic and gestural features, using the ELAN multimodal annotation tool.

1. Introduction

Research on the various aspects of human communicative interaction increasingly draws on the visual as well as the auditory input of the speech event in order to get a full view of the trade-off between different modalities. Researchers focusing on multimodal aspects of interactional language use thus require high-quality multimodal resources, which however are notoriously hard to come by. Compiling, annotating and processing a corpus of audio-video recordings of conversational speech is a time-consuming and complex undertaking even for a relatively restricted set of conversational data.

In order to meet the growing need for a multimodal dialogue corpus, a number of recent projects have started to collect and annotate video recordings of conversational speech, with the aim to provide a fine-grained data collection for theoretical, descriptive and experimental work. For the purpose of the present paper, we refer to two such projects that bear a number of similarities. The first, the HeadTalk project (Knight et al., 2008, 2009) focused on the verbal and gestural signalling of active listenership in conversation. In order to get a high-quality view on the various semiotic channels that language users employ, natural face-to-face conversations were recorded from two perspectives (rather than an external bird's eye view), with two cameras directly facing the participants. The resulting recordings were used as input for computer vision techniques (including a 3D head tracking model), which allow for the (semi-)automatic recognition of specific gesture types (and more specifically head nods). The second project, the IFADV corpus (*IFA Dialogue Video Corpus*, Van Son et al., 2008), consists of 5 hours of annotated video recordings of face-to-face conversations in Dutch. Comparable to the HeadTalk corpus, the interactions were recorded using two cameras positioned next to the speakers and facing the other. The recorded data were annotated automatically for the parameters included in the Dutch CGN spoken language corpus (2006) (e.g. POS tagging and word alignment) and manually for the pragmatic function of utterances (e.g. reactions, grounding acts, etc.) and gaze direction.

Despite the rich source of information they provide, the scope of both the HeadTalk and the IFADV project is limited in a number of ways. First, both corpora were designed primarily to gain a high-resolution picture of facial expressions (including gaze and head nods), as one significant dimension of gestural behaviour. As a consequence, however, the recordings do not always provide visual access to other gestural behaviour types, including hand movement and body posture. In order to maximally disclose the various modalities involved in dialogue, we need a fully-fledged multimodal interaction corpus that incorporates all dimensions of bodily semiotics. Second, studying gaze behaviour on the basis of 'external' video data of a participant's face is problematic as it is impossible to pinpoint the focus of attention (Kendon, 2004; Streeck, 2009). One way to overcome this problem is to track participants' gaze behaviour during face-to-face conversation using head-mounted eye-tracking equipment.

In our project, we aim at developing a dialogue corpus that is both multimodal and multifocal in nature. Using recently developed eye-tracking methods from discourse psychology (Pickering & Garrod, 2004, 2006; Tanenhaus & Brown-Schmidt, 2008), we recorded interlocutors' perspective and behaviour during conversation, using head-mounted scene cameras and eye-trackers. In doing so, we obtained a 3D landscape of the conversation, including production (scene camera, sound) and processing (eye movements for gaze analysis) information for both participants.

2. Recording design and set-up

2.1 Physical set-up: multi-angle recording

In order to maximally capture the various modalities involved in interaction, we used a multi-angle recording technique. The different recording angles comprise one static viewpoint, capturing the participants in profile with a fixed camera for each of the two participants, and two dynamic internalized viewpoints, using head mounted eye-tracking devices (figure 1). For each of the recordings, two speakers are seated face-to-face. To maximize the freedom of hand gesture use, no objects are placed within arm's reach of the participants. Because the eye-tracking devices are head-mounted, the speakers can move freely and don't need to restrict themselves to a certain position or virtual frame.

Figure 1: Multi-angle recording set-up

2.2 Recordings: continuum of interaction types

Because multimodal techniques such as our multi-angle recordings yield such a vastness of data for which no integrated annotation schemes exist, we have chosen an incremental build-up in the type and complexity of the recorded interactions. This contributes to a more controlled coding cycle. On the one hand, we use targeted collaborative tasks because they allow for appropriate trials and baseline conditions to emerge from the interactive dialogue, i.e. they elicit specific types of utterances that can be compared across subjects (Tanenhaus & Brown-Smith, 2008). On the other hand, we let the participants have free-range conversations because these yield the most natural picture of multimodal interaction. The former step of controlled circumstances is useful and even necessary for the latter step of coding truly spontaneous conversation.

The incremental build-up of the interaction types was implemented in our recording sessions as follows:

(i) Participants were asked to describe and discuss *simple and schematic* spatial scenes. We have chosen this type of collaborative task because studies have shown that describing spatial scenes or relations strongly elicits gestural next to verbal communication (Cienki, 2005; Mittelberg, 2007; Sweetser, 2007).

(ii) Participants were asked to describe and discuss *more complex and real life* spatial scenes.

(iii) Participants were asked to have a *free conversation* on a given topic.

The combination of the three interaction types, with an increasing degree of interactional freedom, provides sufficient leverage to elicit input from and interaction between multiple channels involved in dialogue.

3. Materials

For each of the recorded sessions we have joined and synchronised the signals of five different input devices:

(i) 1 fixed colour video camera: Sony HDR-FX1000E

(ii) 2 head-mounted eye-trackers with scene camera: Arrington Gig-E60 Eye Frame Scenecamera System

(iii) 2 directed microphones: Zoom H2

4. Participants

At this stage, the corpus consists of 15 recorded dialogues of about 30 minutes each, of which 5 sessions are fully transcribed and annotated. The participants are undergraduate students from the University of Leuven and Lessius University College, all native Dutch speakers. All participants received a financial compensation for their effort. All of the dialogue pairs knew each other before entering the recording session.

5. Annotations

Part of the processing of our data is based on our recent work on the CORINTH-corpus (CORINTH, 2010). Although the conversations in our video data are partly scripted and directed, they still can be categorized as "spontaneous speech". Since the GAT transcription norm (Selting et al. 1998) has proven to be very well suited for spontaneous conversations and the Tadpole POS-tagger (Van den Bosch et al. 2007) is well trained on conversations in Dutch, we have used both to obtain an alpha version of our corpus. The overview in figure 2 shows how we added multimodal annotation layers to the basic alpha version of the corpus.

Gesture	(McNeill, 1992)
Typology of hand gestures based on physical form and contextual function of the hand movements	
Intonation	(Boersma & Weeninck, 2009)
Frequency/amplitude patterns based on PRAAT	
Gaze	(a.o. Bavelas et al. 2002)
Determination of regions of interest and calculation of relevant overlap (mutual gaze, mutual object of focalization)	
Head nods	(Carter et al., 2006)
Functional typology of head nods	

Figure 2: Multimodal annotation scheme

All of the transcriptions and annotations were done with the ELAN multimodal annotation tool (ELAN, 2002-2009). ELAN is an XML-based tool that allows for a series of search strategies both within single files and over the complete corpus. For further statistical processing of the data, ELAN allows for exports to numerous types of databases, the tab-delimited format being the most useful in that respect.

6. Acknowledgements

The Insight Interaction corpus is supported by grant number 3H090339 STIM/09/03 of the University of Leuven.

7. References

Boersma, P., Weenink, D. (2009). PRAAT: doing phonetics by computer (Version 5.1.02) [Computer program]. http://www.praat.org/.

Carter, R., Knight, D., Adolphs S. (2006). HeadTalk: Towards a multimodal corpus. Presentation at *BAAL: From Applied Linguistics to Linguistics Applied*. Cork: 7-9 September 2006. http://www.ncess.ac.uk/research/sgp/headtalk/20060907_carter_multi_modal_corpus.pdf

CGN. (2006). The Spoken Dutch Corpus project. http://www.tst.inl.nl/cgndocs/doc_English/topics/index.htm.

Cienki, A. (2005). Image schemas and gesture. In B. Hampe (Ed.). *From Perception to Meaning: Image Schemas in Cognitive Linguistics*. Berlin/New York: Mouton de Gruyter, 2005, pp. 421--441.

CORINTH. (2010). Corpus Interactional Humour. Department of Linguistics. University of Leuven.

ELAN. (2002-2009). ELAN. A free and open source multimedia annotation tool. http://www.lat-mpi.eu/tools/elan

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Knight, D., Evans, D., Carter, R., Adolphs, S. (2009). HeadTalk, HandTalk and the corpus: towards a framework for multi-modal, multi-media corpus development. *Corpora*, 4(1), pp. 1--32.

Knight, D., Adolphs, S., Tennent, P., Carter, R. (2008). The Nottingham multi-modal corpus: a demonstration. In *Proceedings of LREC2008* (Workshop on Multi-modal Corpora). Palais des Congrés, Mansour Eddahbi, May 28-30, Marrakech: ELRA.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. Chicago: University of Chicago Press.

Mittelberg, I. (2007). Methodology for multimodality. One way of working with speech and gesture data. In: M. Gonzalez-Marquez, I. Mittelberg, M. Spivey, S. Coulson (Eds.), *Empirical Methods in Cognitive Linguistics*. Amsterdam: John Benjamins, pp.225--248.

Pickering, M., Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27, pp. 169--226.

Pickering, M., Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation,* 4, pp. 203--228.

Selting, M. et al. (1998). Gesprachsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte*, 173, pp. 91--122.

Streeck, J. (2009). *Gesturecraft. The Manu-Facture of Meaning*. Amsterdam: John Benjamins.

Sweetser, E. (2007). Looking at space to study mental spaces. Co-speech gestures as a crucial data source in cognitive linguistics. In: M. Gonzalez-Marquez, I. Mittelberg, M. Spivey, S. Coulson (Eds.), *Empirical Methods in Cognitive Linguistics*. Amsterdam: John Benjamins, pp.203--226.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., Sedivy, J. E. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science,* 268, pp. 632--634.

Van den Bosch et al. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*. University of Leuven, pp. 99--114.

Van Son, R.J.J.H, Wesseling, W, Sanders, E., van den Heuvel, H. (2008). The IFADV corpus: A free dialog video corpus. In *Proceedings of LREC2008*. Palais des Congrés, Mansour Eddahbi, May 28-30, Marrakech: ELRA.

Capturing massively multimodal dialogues: affordable synchronization and visualization

Jens Edlund and Jonas Beskow
KTH Speech Music and Hearing
Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden
E-mail: edlund@speech.kth.se, beskow@speech.kth.se

Abstract

In this demo, we show (a) affordable and relatively easy-to-implement means to facilitate synchronization of audio, video and motion capture data in post processing, and (b) a flexible tool for 3D visualization of recorded motion capture data aligned with audio and video sequences. The synchronisation is made possible by the use of two simple and analogues devices: a turntable and an easy to build electronic clapper board. The demo shows examples of how the signals from the turntable and the clapper board are traced over the three modalities, using the 3D visualisation tool. We also demonstrate how the visualisation tool shows head and torso movements captured by the motion capture system.

1. Introduction

In this demo, we show (a) affordable and relatively easy-to-implement means to facilitate synchronization of audio, video and motion capture data in post processing, and (b) a flexible tool for 3D visualization of recorded motion capture data aligned with audio and video sequences. The tools were developed during the collection of a large massively multimodal corpus of Swedish dialogue. This short paper describes, briefly, the background and motivation for developing these tools, and gives a short presentation of how they work.

2. Background

Recordings of humans talking to each other make up the foundation for much fruitful research into human communicative and social behaviour. As capturing technology develops and our insights further, recordings of increasing complexity are required on many levels: modalities move from text or audio only through audio-visual to what we might call massively multimodal recordings capturing audio, video, 3D motion as well as other modalities such as gaze tracking, EEG, and breathing; what is captured move from single speaker read speech through controlled or task oriented dialogues to spontaneous social conversation; the environment move from anechoic chambers through controlled environments with immobile participants to environments where the participants can move around freely; and the participants move from a fixed number of preselected speakers to a variable number of unknown speakers. This complexity is daunting to handle, and the technology required can be very expensive. At the same time, the richness of the data makes it difficult and time-consuming to overview. In this demonstration, we will show how affordable and robust synchronization is facilitated within the Spontal project (Edlund et al., 2010).

Synchronization between audio and video is commonly achieved using mature technologies. Nevertheless, it is fairly expensive to set up a system that synchronizes one video feed with a separate audio feed, and more so as additional video feeds are added. When we move to massively multimodal capture by adding less conventional modalities, the cost is severely increased, and it becomes harder to find mature technologies for the task. Some modalities, such as measures of breathing using piezoelectric film-elastic belts or of fundamental frequency using laryngographs can be fairly easily synchronized by capturing them on a separate channel on a frame synchronized audio board together with the audio. Others, such as the output from the evermore accessible and affordable 3D motion capture systems, may be considerably harder to integrate in an existing system for synchronization given a standard research budget. Naturally, this can be expected to improve over time. Whilst we wait for improvement, the present paper describes the two analogous devices we designed when recording the Spontal database. The Spontal project has seen a succession of recording series, and the systems described here evolved with each of these. What is described here is the system as used in the final batch of recording sessions in the project.

3. Synchronization

Two devices were developed to facilitate post-recording validation and resynchronization as well as data navigation. Both devices produce a signal that can be captured simultaneously on audio track, video and in motion capture, but that is produced robustly independently of these capturing systems. This independency makes the entire recording configuration robust against failure of the individual systems it is made up of: if the motion capture fails, it can simply be restarted while audio and video capture continues.

2.1 The clapper board

The first device is a replacement for the clapper board or sync slate used on film sets. It consists of a simple switch that simultaneously controls two green diodes, an IR diode and one channel of audio throughput (see Figure 1). The diodes are mounted in separate devices, and the audio throughput takes a sine signal as its input. The switch in itself is placed in the control room, whereas the diodes are placed in the recording studio, one green diode in the line of sight of each video camera (SPontal uses two HD video camera; more diodes could be easily added in a different

setup), and the infrared diode clearly visible to the infrared 3D motion capture cameras. When the switch is flipped, the infrared diode is captured by the motion capture system (which perceives it as one of the reflecting markers it is designed to capture), the green diodes are captured by the video cameras, and the sine tone is captured on a separate frame synchronized track in the audio recordings. Each of these signals can be detected in the resulting data using automatic methods. In the Spontal recordings, the switch is used at the beginning and end of each session (3 blinks), at each ten minute interval of the recording (2 blinks), and anytime something noteworthy happens (1 blink, followed by a spoken explanation by the person managing the recording).

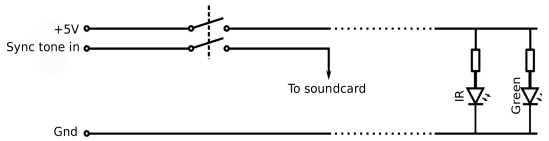


Figure 1: Schematic of the clapper board: the left side depicts the switch (positioned in the control room) and the right part the diodes (places in the recording studio).

2.2 The sync signal

The second device makes it possible to resynchronize the data in post processing by stretching and compressing the data, should this prove necessary. The device is basically a simple turn-table placed in plain view of both video and 3D motion capture cameras. The turn-table's audio output is recorded on an audio cannel that is frame synchronized with the audio recordings. On the turn-table, which rotates at a constant speed of 33 rpm, a record is placed. The record has been deliberately scratched deeply at one place, causing it to produce a sharp noise on each rotation. On the record, a reflecting marker that is readily captured by both 3D motion capture and video is placed. The result is that the position of the marker in each audio, video and motion capture frame can be compared to its expected position, and the data can be temporally stretched or compressed if any discrepancies are detected.

Figure 2: The turntable in action. In addition to the reflecting marker mounted on the record, two markers fixed at a set distance on the side of the turntable serve as a distance reference.

4. Visualization

When recording and processing motion capture data, visualization of the data and of the effects of data manipulation and post-synchronisation is invaluable. In addition to demonstrations of the synchronization devices, the demo includes a tool called *DotPlot*, which allows visualization recorded of motion capture data in synchronization with time audio and video sequences. The tool is implemented as a plugin to the WaveSurfer open source tool (Sjölander & Beskow, 2000) for visualization and manipulation of sound and video. The DotPlot plugin can be configured to automatically detect motion capture files in a given directory, and opens them automatically when the corresponding audio file is opened in WaveSurfer. DotPlot plots each marker as a coloured sphere in 3D, optionally connecting the spheres with lines to produce stick figures, given that a skeleton has been defined by the user. The 3D plot can be freely rotated, panned and zoomed. When the cursor moves back and forth in the WaveSurfer time view, or when the sound is played, the plot is updated accordingly. DotPlot can also co-exist with WaveSurfer's video plugin, thus allowing a simultaneous view of audio, video and motion capture.

5. Concluding remarks

The researchers working on the recordings of the Spontal database, including the present authors, testify that using the clapper board and turntable presented here made the recording task considerably less stressful and much less susceptible to technical errors. Currently, the demonstration is as far as we get in terms of validation, however, as the Spontal project has just recently finished recording, and the end results of the synchronization are not yet in.

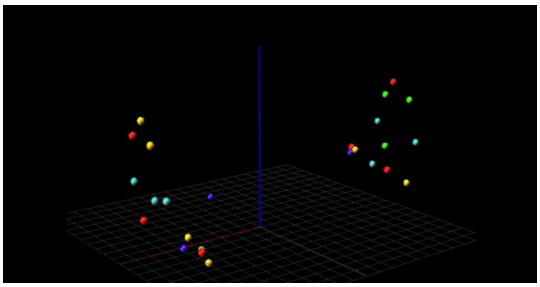


Figure 3: View from DotPlot: still of two Spontal participants chatting.

6. References

Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., & House, D. (2010). *Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture*. To be published in *Proc. of LREC 2010*. Valetta, Malta.

Sjölander, K., & Beskow, J. (2000). *WaveSurfer - an open source speech tool*. In Yuan, B., Huang, T., & Tang, X. (Eds.), *Proceedings of ICSLP 2000, 6th Intl Conf on Spoken Language Processing* (pp. 464-467). Beijing.

Generating and annotating corpora of multimedia telecommunications of pediatric cancer patients and their families and friends

Thomas Bliesener
Institut für Kommunikationswissenschaft, Universität Duisburg-Essen
Universitätsstr. 12, D-45117 Essen
E-mail: thomas.bliesener@uni-duisburg-essen.de

Abstract

The generation and analysis of human multimodal telecommunication corpora imposes complex requirements. Since the process takes place by mutal transmission of multimedia mappings of the separated locations, separate recordings must be made at each location and synchronized lateron. Since operations on the media devices are part of the telecommunication, these actions as well as their failures and the resulting flaws must be registered and annotated in additional categories that do not fit into usual modalities. Even minimal behaviors like trembling of a mouseclick can have important consequences like disappearing video, so an annotation system that takes account of the inevitable fallibility of technological operations must be open for non-predefined categories.

Most corpora to which multimodal annotations and analyses are applied, consist of recordings of live behavior and communication of physically co-present persons in one geograpical place. However, since the rapid progress of broadband internet connections all over the globe, the technically mediated multimedia telecommunication between people in separate places, like by Skype, became a mass phenomenon. Contemporary messaging and conferencing programs combine two-point or even multipoint audio, video, textchat, application sharing, desktop sharing, file access, file transfer, and online gaming in one single session. Participants do not have direct perceptions of each other in any modality, but rely on realtime capturings and transmissions of each other's communicative behavior in order to preceed within their tele-communication process. Additionally, the mutual multimodal perceptions of the participants are enhanced by a large range of auditive and visual signals and symbols, added by the intermediate conferencing software and the local hardware at the endpoints. Now, if you produce audiovisual recordings of such a multimedia telecommunication, you get a capturing of the combined innercommunicative capturings, plus the addional technological signs and signals. Consequently, considerations on annotation must develop notions and systematics that take into account the double layers of imaging and their enhancement by upper layer signals – a requirement even more complex than multimodal annotations done so far.

In a study on the use of audiovisual telecommunication of pediatric cancer patients in a germ-reduced isolation ward at the University Clinic of Essen, accomplished from 2006 to 2009 under the direction of Prof. H. Walter Schmitz (Institut für Kommunikationswissenschaft, Essen) more than two hundred hours of sessions with MSN Messenger, ICQ and Skype were screencaptured with Techsmith Morae Recorder, including the audio inputs and outputs produced by the soundcards. Participants were informed that none of their local computer activities, but all of their computer mediated communications by messaging clients were automatically recorded and saved to the harddisk. They were given two explanations for the recordings: (a.) subsequent analyses of communicational problems that are caused by the features of the unusual technological setting or by the deficiencies of the technological conditions, (b.) demonstration of problems and of successful achievements in order to train future supporters and to convince potential sponsors of the benefit of such a humanitarian project. They were assured that they could prevent a recording anytime by a doubleclick on a desktop icon. Actually, no participant ever turned off the recording, and only one (female) participant withdrew her informed consent in the very moment when the video function started working successfully. Parts of the material are open to research under certain conditions.

The purpose of the research did not consist in the generation of anything different than improvements for future communications of the same kind. It did not aim at obtaining categories or rules for the creation of artificial agents (embodied or virtual ones), and its categories were not designed to stand a test in practical simulations (Kipp, 2008). So, a proof of adequacy by simulation is not at hand. Instead, the analytical descriptions were meant to seize the degree of richness of the particular communicative events. They were to enable a kind of quality check. Corrections and complementations were brought about by observations, talks and sometimes a common inspection of a recording. In case of shortcomings, the findings were to trigger activities for improving both, technology and support.

The material from first to last is *natural material*. Communication was not done for tests or tasks of the researchers. The observed and recorded situations and actions, their purposes and their qualities were determined only by the participants themselves in accordance with the requirements and possibilities of their daily lives and environments. It is true, many changes of technological conditions were introduced and compared. But for the patients, these activities were just a support of their own continuous strive for technological improvements for their pursuit of existential needs.

Due to technological restrictions and flaws of various origins and due to limitations of user skills, more than half

of the material does not contain complete two-way audio and two-way video, or at least not in a sufficient quality. This is, however, less a shortcoming but rather a value. It is true, a minor portion of the deteriorations could be identified as artefacts of the recording technology. E.g. an upgraded version of Skype automatically turned off microphone playback control, thereby preventing the recording of the local participant's voice; while both partners understoood each other well, the damaged recording is unsuitable for reconstructing their telecommunication. But the majority of the deficiencies in our material reflects problems in the telecommunication itself. This part of the material is most interesting for the investigation of errors, obstacles and remedies within the telecommunication process, the purpose for which we started the research.

Now, finding out the source of troubles within the videoconferences often requires a good deal of detective work, which can only partially draw on general technological background information (like known issues of software versions or of the clinic network). Rather, a number of detailed, not predefined behavioral clues within the recorded material are needed, plus a fabric of sophisticated hypotheses and inferences. The same requirement holds for the general distinction between trouble in the object and trouble in the recording.

To give an example of a Skype session: A routined movement of the mouse cursor to the button for sending snapshots to the remote partner happened to stick at the button, while the sending process did not start. The reason for this malfunction, in turn, shall be skipped here. A little later the partners finished their document based activities and wanted to go back to pure audiovisual conversation. They agreed to turn on their video. The remote video already appeared, but the activation of the local video ran into a desaster. Skype, including the audio connection, quite unexpectably crashed. In human face to face communication, turning away one's face or falling into silence would be of utmost importance. In human technology mediated communication, there may be a greater tolerance because of the known intricacies of digital life. But a total interruption is a serious incident nevertheless. In technology mediated communication, not only embodied expression, which is possibly somewhat coded and somewhat "recognizable", has a high communicative impact. Rather, an inconspicuous mouseover on the mediating machine can do just as much.

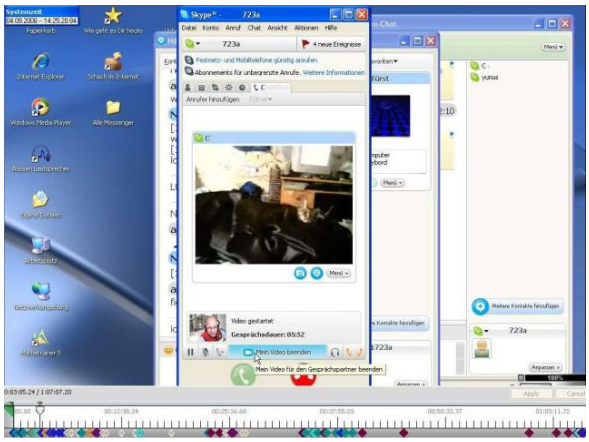


Figure 1. Reproductions of multimodal behavior together with other multimedia in synchronous telecommunication

For analyses and annotations of multimodal human interaction, the interposition of transmission media induces grave complications. In direct communication, each participant draws from perceptions and recognitions of voice, gesture, posture, movements of his co-present partner. The basis of his communication is a binary relation. Since behavior analysis takes the same perspective as a participant, it is grounded on the same binary relation. This starting point is exemplified in the following table, leaving out all subsequent differentiations of highly developped coding schemas like MUMIN (2010).

Production of behaviors	Reception of production of behaviors
vocal verbal paraverbal misc. vocal	identification, appraisal and evaluation of auditive behaviors
eye gaze face head hand posture position	identification, appraisal and evaluation of visual behaviors

Table 1. Approach to multimodal behavior itself

In telecommunication, a participant does not draw from perceptions of his partner, but from perceptions of technologically *modified reproductions* of his partner's voice, gesture, posture, movements. Even these multimodal behaviors themselves are not as usual, but right from their origin, they are *adapted* to the conditions of technological mediation. Additionally, a participant draws from hints on operations by which his partner *adjusts the transmission media* or adjusts himself to them, e.g. the partner zooms a camera or approaches his mouth to a microphone. Moreover he gets numerous displays of communicated material like text in a chat tool, emoticons, animoticons, avatars, and actions in game space. Finally, he gets lots of particular signals which may be triggered by his partner, like a "buzzz", or may be sent automatically by the system, like "If a third party joins your Skype conference, you'll lose the video connection with your present partner. Do you want to continue?" All of these must be identified or described or at least localized in the recording by an uncategorized time mark.

In our multimodal transcriptions, we include descriptions or marks for a multitude of behaviors that are operative on the media devices. Some appear in the same column of the transcript that registers events in the live video window, like gesture, posture und movements. But since they are particularly targetted at devices, we add a marker for "audio device operation", "video device operation" etc. Should the same operation become audible, it will be noted in the column for audio events including speech, but again with an addional marker for the kind of operation.

Local production of behavior	Remote representation of behavior	Reception of remote representation	Reception of inferred original production
vocal verbal paraverbal misc. vocal	reproduction of vocal verbal reproduction of paraverbal reproduction of misc. vocal	identification, appraisal and evaluation of auditive reproduction	inference, appraisal and evaluation of original auditive production
eye gaze face head hand body posture position	reproduction of eye reproduction of gaze reproduction of face reproduction of head reproduction of hand reproduction of body posture reproduction of position	identification, appraisal and evaluation of visual reproduction	inference, appraisal and evaluation of original auditive production
writing of text writing style painting using signal objects	reproduction of writing of text reproduction of writing style reproduction of painting reproduction of using signal objects		
input to media devices	substitutes and augmentations for all kinds of multimodal behaviors	identification, appraisal and evaluation of auditive and visual data	inference, appraisal and evaluation of original behavior of input to media device
input to media devices	user-triggered signals		
adaption to or adjustment of media devices	media generated changes in reproductions of multimodal behavior		
- - -	system generated signals system generated effects		

Table 2. In synchronous telecommunication, reproductions of multimodal behavior appear among multimedia contents

There is an additional severe complication in telecommunication, affecting behavior as well as research. In unmediated communication, most often production and display of a behavior are done as one event in one place. If you want to magnify yours partner's sight of your face, you may move your face towards him. In mediated communication, usually production and display are separated. By mouse action, you click some buttons in your videosoftware, meanwhile a hundred kilometers away, your face appears in the video window as zooming in slowly. If the conference does not include desktop sharing or a secondary context camera, your operation remains invisible to your partner. So, in order to reconstruct the process as a whole, we have to make *separate simultaneous recordings in all locations*. Each recording will be annotated separately. Of course, recordings and transcriptions have to be synchronized later, so that "bringing your face closer to your partner" reappears in the transcript in two or more columns.

However, post-synchronization of recordings from two different computers in different locations, which were simultaneously used for regular CPU hungry tasks, is not easy. Even if each computer was synchronized over the internet with the atomic clock of the national institute for scientific and technical services (PTB), there may happen *aberrations* due to network conditions and hardware performance. As a consequence, recordings with 20 frames per second may later be put together only with an inexactitude of several frames. For these cases, we developped procedures for manually adjusting them more precisely.

An analysis of successfully post-synchronized simultaneous recordings from two locations reveils soon that one popular phenomenon of multimodal interactions is more or less disturbed in synchronous multimedia telecommunication: rhythm. The inter-individual "rhythms of dialogue" (Jaffe, 2001) are impaired by signal delay caused by hardware limits and transmission times. In the decade before DSL and FTTH, voice over internet suffered severely from delayed slots in turntaking (Ruhleder, 2001). Today quite similar problems return in mobile communication. Additionally, the intra-individual

rhythms of speech and paraverbal behavior are often impaired by extensive multitasking with one or even multiple communication partners in a given session. If one focusses on a selection of feedback, turn management and sequencing behavior that in some instances are supposed to be a source of "rapport and pleasantness" (Allwood et al., 2008, 266), skyping and related practices do not seem to provide good times. On the other hand, even in sessions with limited coordination and doubtful rhythms, the participating children may display much laughter and may be committed to multitasking with body and soul. So maybe, we only have to inspect the material under different criteria and abstractions which then will reveal again patterns of involvement and closeness. We are still watching out for them.

The demonstration will not show a completed best practice, but is intended to exemplify the peculiarities and intricacies of this kind of material. A selection of short events and sequences from our recordings of Skype conferences with textchat, audio, video, file transfer, desktopsharing, and online gaming will be displayed. The corresponding demands for categories, transcription conventions and presentation tools will be discussed.

References

Allwood et al. (2007). The analysis of embodied communicative feedback in multimodal corpora: a prerequisite for behavior simulation. In *Language Resources & Evaluation*, 41, pp. 255--272.

Bliesener, T. (2004). Training of Synchronous Cooperative Tele-Learning: Experiments with Syntopical Monitoring. In *Proceedings of Ed-Media 2004, Lugano, June 21-26*, CD-rom: procbook2.pdf, pp. 2505--2511.

Döring, J., Schmitz, H. W., Schulte, O. A. (Eds.) (2003). *Connecting Perspectives. Videokonferenz: Beiträge zu ihrer Erforschung und Anwendung.* Aachen: Shaker Verlag. (Essener Studien zur Semiotik und Kommunikationsforschung).

Joffe, J. et al. (2001). *Rhythms of Dialogue in Infancy.* London: Blackwell.

Joisten, M. (2007): Multimediale Gespräche in Skype: Hybridisierung von Gebrauchsweisen in der interpersonalen Kommunikation. In S. Kimpeler et. al. (Eds.), *Die digitale Herausforderung. Zehn Jahre Forschung zur computervermittelten Kommunikation*. Wiesbaden: Verlag für Sozialwissenschaften, pp. 149--158.

Kipp et al. (2008), An annotation scheme for conversational gestures: how to economically capture timing and form. In *Language Resources & Evaluation*, 41, pp. 325--339

MUMIN (2010), www.cst.dk/mumin/resources.html

Ruhleder, K., Jordan, B. (2001). Co-Constructing Non-Mutual Realities: Delay-Generated Trouble in Distributed Interaction. In *Computer Supported Cooperative Work*, 2, pp. 113--138.