Programme of the Workshop on Multimodal Corpora

9.00 Welcome

SESSION "MULTIMODAL EXPRESSION OF EMOTION"

9.15 Annotation of Cooperation and Emotions in Map Task Dialogues. *F. Cavicchio and M. Poesio*

9.45 Double Level Analysis of the Multimodal Expressions of Emotions in Human-machine Interaction. *J.-M. Colletta, R. Kunene, A. Venouil, and A. Tcherkassof*

10.15 - 10.45 Coffee break

SESSION "MULTIMODALITY AND CONVERSATION"

10.45 Multimodality in Conversation Analysis: A Case of Greek TV Interviews. *M. Koutsombogera, L. Touribaba and H. Papageorgiou*

11.15 The MUSCLE Movie Database: A Multimodal Corpus with Rich Annotation for Dialogue and Saliency Detection. *D. Spachos, A. Zlantintsi, V. Moschou, P. Antonopoulos, E. Benetos, M. Kotti, K. Tzimouli, C. Kotropoulos, N. Nikolaidis, P. Maragos and I. Pitas*

SESSION "MULTIMODAL ANALYSIS OF ACTIVITIES"

11.45 A Multimodal Data Collection of Daily Activities in a Real Instrumented Apartment. *A. Cappelletti, B. Lepri, N. Mana, F. Pianesi and M. Zancanaro*

12.15 Unsupervised Clustering in Multimodal Multiparty Meeting Analysis. *Y. Matsusaka, Y. Katagiri, M. Ishizaki and M. Enomoto*

12.45 Discussion

13.00 - 14.30 Lunch

SESSION "INDIVIDUAL DIFFERENCES"

14.30 Multimodal Intercultural Information And Communication Technology – A Conceptual Framework For Designing And Evaluating Multimodal Intercultural ICT. *J. Allwood and E. Ahlsén*

15.00 Multitrack Annotation of Child Language and Gestures. *J.-M. Colletta, A. Venouil, R. Kunene, V. Kaufmann and J.-P. Simon*

15.30 The Persuasive Impact of Gesture and Gaze. *I. Poggi and L. Vincze*

16.00-16.30 Coffee break

16.30 On The Contextual Analysis Of Agreement Scores. *D. Reidsma, D. Heylen and R. Op den Akker*

SESSION “PROCESSING & INDEXING"

17.00 Dutch Multimodal Corpus For Speech Recognition. *A. G. Chitu and L. J.M. Rothkrantz*

17.30 Multimodal Data Collection in the Amass++ project. *S. Martens, J. Hendrik Becker, T. Tuytelaars and M.-F. Moens*

18.00 The Nottingham Multi-modal Corpus: a Demonstration. *D. Knight, S. Adolphs, P. Tennent and R. Carter*

18.15 Analysing Interaction: A comparison of 2D and 3D techniques. *S. A. Battersby, M. Lavelle, P. G.T. Healey and R. McCabe*

18.30 Discussion

19.00 End of workshop

(Followed by an informal dinner)

Organiser(s)

MARTIN, J.-C. (CNRS-LIMSI, France)
PAGGIO, P. (Univ. of Copenhagen, Denmark)
KIPP, M. (DFKI, Germany)
HEYLEN, D. (Univ. Twente, The Netherlands)

Programme Committee

Jan Alexandersson, D
Jens Allwood, SE
Elisabeth Ahlsén, SE
Elisabeth André, D
Gerard Bailly, F
Stéphanie Buisine, F
Susanne Burger, USA
Loredana Cerrato, SE
Piero Cosi, I
Morena Danieli, I
Nicolas Ech Chafai, F
John Glauert, UK
Kostas Karpouzis, G
Alfred Kranstedt, D
Peter Kuehnlein, NL
Daniel Loehr, USA
Maurizio Mancini, F
Costanza Navarretta, DK
Catherine Pelachaud, F
Fabio Pianesi, I
Isabella Poggi, I
Laurent Romary, D
Ielka van der Sluis, UK
Rainer Stiefelhagen, D
Peter Wittenburg, NL
Massimo Zancanaro, I

Table of Contents

Author Index

Introduction

A 'Multimodal Corpus' targets the recording and annotation of several communication modalities such as speech, hand gesture, facial expression, body posture, etc. Theoretical issues are also addressed, given their importance to the design of multimodal corpora. This workshop continues the successful series of similar workshops at LREC 00, 02, 04 and 06 also documented in a special issue of the Journal of Language Resources and Evaluation. There is an increasing interest in multimodal communication and multimodal corpora as visible by European Networks of Excellence and integrated projects such as HUMAINE, SIMILAR, CHIL, AMI, CALLAS. Furthermore, the success of recent conferences and workshops dedicated to multimodal communication (ICMI, IVA, Gesture, PIT, Nordic Symposia on Multimodal Communication, Embodied Language Processing) and the creation of the Journal of Multimodal User Interfaces also testifies to the growing interest in this area, and the general need for data on multimodal behaviours. The focus of this LREC'2008 workshop on multimodal corpora is on models of natural interaction and their contribution to the design of multimodal systems and applications.

Topics to be addressed include, but are not limited to:
• Multimodal corpus collection activities (e.g. direction-giving dialogues, emotional behaviour, human-avatar interaction, human-robot interaction, etc.)
• Relations between modalities in natural (human) interaction and in human-computer interaction
• Application of multimodal corpora to the design of multimodal and multimedia systems
• Fully or semi-automatic multimodal annotation, using e.g. motion capture and image processing, and its integration with manual annotations
• Corpus-based design of systems that involve human-like modalities either in input (Virtual Reality, motion capture, etc.) and output (virtual characters)
• Multimodal interaction in specific scenarios, e.g. group interaction in meetings
• Coding schemes for the annotation of multimodal corpora
• Evaluation and validation of multimodal annotations
• Methods, tools, and best practices for the acquisition, creation, management, access, distribution, and use of multimedia and multimodal corpora
• Interoperability between multimodal annotation tools (exchange formats, conversion tools, standardization)
• Metadata descriptions of multimodal corpora
• Automated multimodal fusion and/or generation (e.g., coordinated speech, gaze, gesture, facial expressions)
• Analysis methods tailored to multimodal corpora using e.g. statistical measures or data mining

We expect the output of this workshop to be: 1) deeper understanding of theoretical issues and research questions related to verbal and non-verbal communication that multimodal corpora should address, 2) larger consensus on how such corpora should be built in order to provide useful and usable answers to research questions, 3) shared knowledge of how the corpora are contributing to multimodal and multimedia system design, and 4) an updated view of state-of-the-art research on multimodal corpora.

ORGANISING COMMITTEE
MARTIN Jean-Claude, LIMSI-CNRS, France
PAGGIO Patrizia, Univ. of Copenhagen, Denmark
KIPP Michael, DFKI, Saarbrücken, Germany
HEYLEN Dirk, Univ. Twente, The Netherlands

Annotation of Cooperation and Emotions in Map Task Dialogues

Federica Cavicchio, Massimo Poesio
CIMeC, Università degli Studi di Trento
Palazzo Fedrigotti, Corso Bettini 31, 38068 Rovereto (Tn) Italy
E-mail: federica.cavicchio@unitn.it, massimo.poesio@unitn.it

Abstract

In this research we investigate the relationship between emotion and cooperation in map task dialogues. It is an area were still many unsolved questions are present. One of the main open issues is the labeling of "blended" emotions, their annotation and recognition. Usually there is a low agreement among raters in "giving name" to emotions. Moreover, emotion recognition is surprisingly higher in a condition of modality deprivation (only acoustic or only visual vs. bimodal). Because of this previous results we don't ask raters to directly annotate emotions, but to use a small set of features (as lips or eyebrows shapes) to annotate our corpus. The analyzed materials come from an audiovisual corpus of Map Task dialogues elicited with scripts. We point out the "emotive" tokens by simultaneous recordings of video clips and phsychophysiological indexes (ElectroCardioGram ECG, Galvanic Skin Conductance GSC, ElectroMyoGraphy EMG). After this, we select emotive tokens and we annotate each of them with our multimodal annotation scheme. Each annotation will lead to a cluster of signals identifying the emotion corresponding to a cooperative/non cooperative level; the last step involves agreement among coders and reliability of the emotion description. Future research will investigate with brain imaging methods the effect of putting emotions into words and the role of context in emotion recognition.

1. Map Task Revisited

Which is the relationship between emotions and cooperative behavior? Do positive emotions enhanced cooperation and negative emotions disrupt it? Which are the "recovery" conversational strategies from the emotional point of view? In dialogue mediated context how are negative or positive emotions displayed by the face? Is there any difference in cooperation and emotion displaying when you see your interlocutor with respect to when you don't see her? To answer questions like these we are collecting a multimodal corpus of interactions using a modified Map Task to elicit dialogues.

Map Task is a cooperative task involving two participants used for the first time by the HCRC group at Edinburg University (Anderson et al., 1991). In this task two speakers sit opposite one another and each of them has a map that the other cannot see. One speaker, designated the Instruction Giver, has a route marked on her map; the other speaker, the Instruction Follower, has no route. The speakers are told that their goal is to reproduce the Instruction Giver's route on the Instruction Follower's map. The maps are not identical and the speakers are told this explicitly at the beginning of their session. However, it is up to them to discover how the two maps differ.

In our Map Task the two participants (both native Italian speakers) are separated by a short barrier or a full screen. They both have a map with some objects. Some of them are in the same position and with the same name, but most of them are in different positions or have names that sound similar to each other. A further condition is added: the follower or the giver could be alternatively a confederate with the aim of getting the giver angry. It is said to the participants that the whole task should end in 15 minutes. The confederate at minutes 4, 9 and 13 acts the following script (Anderson et al., 2005):

- *"You driving me in the wrong direction, try to be more accurate!"*
- *"It's still wrong, this can't be your best, try harder! So, again, from where you stop"*
- *"You're obviously not good enough in giving instruction"*

During this dialogue giver or follower psychophysiological state is recorded and synchronized with video and audio recordings. These elicitation and collection methods allow us pointing out at which moment something is happening at the peripheral nervous system level. When an emotive state is felt participant heart rate and skin conductance are significantly different from the resting state and the task state. At the same time it is a quite impossible recognizing which is the felt emotion only on the basis of these data (Cacioppo et al., 1993). Thus, it is up to our coding scheme to label multimodal aspects of the emotive tokens, annotating face and body displays and the correlation with the corresponding cooperative behavior performed.

2. Method

The emotion annotation coding scheme used to analyze our map task chunks partially follows Craggs & Woods (2005) and Martin et al. (2006) annotation schemes of blended emotions. As for the emotions analyzed by those authors, our corpus emotions are expressed at different blending levels (i. e. blending of different emotion and emotive levels). In Craggs & Woods opinions' annotators have to label the given emotion with a main emotive term (e. g. anger, sadness, joy etc.) correcting the emotional state with a score ranging from 1 (low) to 5 (very high).

From a cognitive and neuroscience point of view, several studies have shown how emotional words and their connected concepts influence emotion judgments and, as a consequence, emotion labeling (for a review see Feldman Barrett et al., 2007). Moreover research on emotion recognition by face has found out that some emotions (e. g. anger or fear) are discriminated only by mouth or eyes/eyebrows configuration. Face seems to be evolved to transmit orthogonal signals, with a low correlation each other. Those signals are deconstruct by the human "filtering function" as optimized inputs (Smith et al., 2005). Moreover PCA analyses on emotive

face displays find out that basic emotions are well recognized by mouth vs. eyebrows/eyes configuration (Susskind et al., 2007). On the basis of these findings, we decide not to label emotions directly. We attribute valence and activation to verbal and nonverbal signals "deconstructing" them in simpler signs with implicit emotive dimensions. Thus, in our coding scheme a smile would be annotate as ")" and a large smile as "+)", meaning a higher valence and activation (see Fig. 1). Using those signals to annotate emotions could recall FACS annotations, but our method is simpler and requires less training for the annotators.

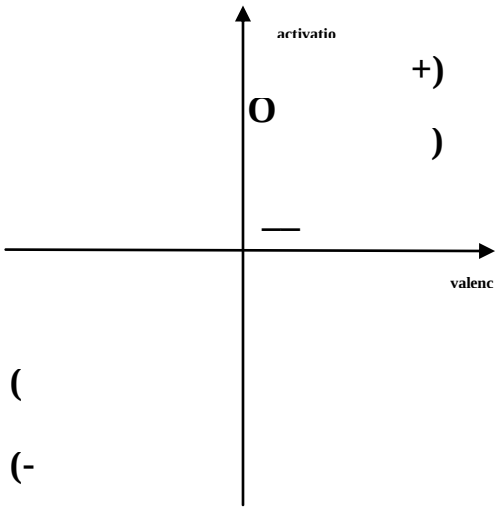

Figure 1: Example of mouth display data collected from annotation scheme into an emotional space

As our corpus is multimodal we analyze and annotate the different communication modality, as shown in Table 1.

Modality	**Expression type**
Facial displays	Eyebrows Eyes Gaze Mouth Head
Gestures	Hand gestures Body posture
Speech	Segmental Suprasegmental

Table 1: Modalities involved in the multimodal analysis

In the following we describe the modalities and the annotation features of our multimodal annotation scheme.

3. The Coding Scheme

Selected audio and video clips are orthographically transcribed. For orthographic transcription we adopted a subset of the conventions applied to LUNA project corpus transcription (see Rodriguez et al., 2007).

As regards cooperative behavior we listed different types of cooperative or not cooperative utterances disrupting or following Gricean cooperative maxims (see Table 2**,** Davies, 2006 adapted). We assign a score ranging from -2 to +2 that pointed out cooperative vs not cooperative behavior.

We also analyze turn management: we annotate turn taking of giver and follower, if one of them give the turn to the other (even when speaker is offering the speaking turn to the interlocutor), kept or hold of the conversation turn.

As regards facial movements we analyze the movements of upper and lower part of the face. Thus an analysis of emotive labial movements as well as eyebrow

No answer to question: cooperation level -2
No information add when required: cooperation level -2
Inappropriate reply (no giving info): cooperation level -1
Giving instruction: cooperation level 0
Question answering y/n: cooperation level +1
Repeating instruction: cooperation level +1
Question answering y/n + adding info: cooperation level +2
Checking the other understands (*ok? Are you there?*): cooperation level +2
Spontaneous info/description adding: cooperation level +2

Table 2: Cooperation types and attributed levels implemented in our coding scheme

movements and forehead wrinkles are implemented in our annotation system. The annotation is based on a little amount of signs similar to emoticons. We sign two levels of arousal using the plus and minus signs. In particular, as regard mouth movements:

- **Closed lips:** when the mouth is closed choose **closed** label.
- **Corners up:** e.g. when smiling**,); +)** very happy .
- **Corners down:** e.g. in a sad expression**, (, +(** very sad.
- **Protruded:** when the lips are rounded, O.
- **Lip biting:** when one of the lips is bite, usually lower lip.
- **1 corner up:** for asymmetric smiles.

As regard *Gesture,* the categories used to annotate hand movements mainly are taken from McNeill's work (McNeill, 2005) Hand gesture annotation presupposes the so-called gesture phrases identification as a markable. In fact to simplify the annotators work we do not try to capture the internal structure of a gesture phrase (i. e. preparation, stroke and retraction phases). In other words, annotators find the gestures she wants to annotate

marked, and can go on with emotive aspects tagging. The tagging of the shape of hand gestures is very simplified in comparison with the coding scheme used at the McNeill Lab. We only take into account the two dimensions *Handedness* and *Trajectory*, without analyzing orientation and shape of the various parts of the hand(s), and we define trajectory in a very simple manner, similar to what is commonly done for gaze movements. The semantic-pragmatic analysis consists of the categorization of the gesture type in semiotic terms, the second concerns the communicative functions of gestures. We also annotate the emotive aspects of gesture inspired by the annotation scheme used for the emoTv database (Martin et al., 2006). We are interested in annotating the semantic aspects of gesture as well as the emotive ones because some gestures (e. g. symbolics) are strictly linked to emotions (i. e. hitting on the desk with the hand or the punch or the hand palm on the forehead). Moreover we want to point out which gesture types are more used in cooperative and non cooperative behavior.
Coding scheme is implemented in AnViL, a software allowing us to analyze audio and video features (see Fig. 2).

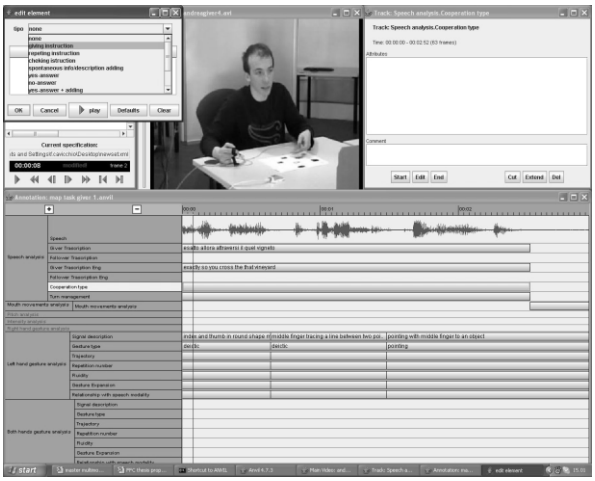

Figure 2: Coding scheme implemented in AnViL

4. Multimodal Relations

Facial displays and gestures can be synchronized with speech at different levels as at the phoneme, word, phrase or long utterance level. In this coding scheme, the smallest speech segment we expect to annotate is word. Thus we do not expect annotators to take syllables or phonemes into consideration because we want to avoid too long training.
Our multimodal tags build on the classification proposed in Poggi and Magno Caldognetto (1996, see Table 3).

5. Conclusions and Future Works

Cooperative behavior and its relationship with emotions is a topic of great interest in the field of dialogue annotation. Usually emotions achieve a low agreement among raters and surprisingly emotion recognition is higher in a condition of modality deprivation (only acoustic or only visual vs. bimodal). Cognitive and neuroscience research shows that emotion recognition is a process performed firstly by sight and processed by limbic system, but the awareness and consequently labeling of emotions is mediated by prefrontal cortex. Moreover a predefined set of emotion labels can influence the perception of facial expressions. Thus we decide to deconstruct each signal without attributing directly an emotive label. Even if we don't have final results we considerate promising the implementation in computational coding schemes of neuroscience evidences on transmitting and decoding facial expression of emotions. Further researches will carry out an fMRI experiment to investigate the influence of task context on labeling emotive terms.

Function	**Relationship between gesture/facial display and speech**
repetition	gesture/facial display bears exactly same meaning as words (this can be also a reinforcement if the gesture puts what has been said with speech in focus)
addition	the meaning of the gesture/facial display adds information to word meaning (this can lead to redundancy of information)
substitution	gesture/facial display replaces unsaid word(s). This is quite difficult to understand, in same cases we can say that the gestures stand on their own
contradiction	gesture/facial display meaning contradicts what has been said vocally, e.g. to denote sarcasm, irony.
no relationship	

Table 3: Gesture/modality and speech relationship

References

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, 34, pp. 351--366.

Anderson J.C., Linden W., Habra M.E. (2005). The importance of examining blood pressure reactivity and recovery in anger provocation research. *International Journal of Psychophysiology*, 57, pp. 159--163.

Cacioppo, J. T., Klein, D. J., Berntson, G. G., Hatfield, E. (1993). The psychophysiology of emotion. In R. Lewis & J. M. Haviland (Eds.), *The handbook of emotion*. New York: Guilford Press, pp. 119--142

Craggs, R., Wood, M. (2004). A Categorical Annotation Scheme for Emotion in the Linguistic Content of Dialogue. In *Affective Dialogue Systems*, Elsevier, pp. 89--100

Davies, B.L. (2006). Testing dialogue principles in task-oriented dialogues: An exploration of cooperation, collaboration, effort and risk. Leeds Working Papers in Linguistics and Phonetics, No.11.

Douglas-Cowie, E., Devillers, L., Martin, J.-C., Cowie, R., Savvidou, S., Abrilian, S., Cox, C. (2005). Multimodal Databases of Everyday Emotion: Facing up to Complexity. In *9th European Conference on Speech Communication and Technology (Interspeech'2005)* Lisbon, Portugal, September 4-8, pp. 813--816.

Feldman Barrett, L., Lindquist, K. A., Gendron, M. (2007). Language as Context for the Perception of Emotion. *Trends in Cognitive Sciences*, 11, 8, pp. 327--332.

Martin, J.-C., Caridakis, G., Devillers, L., Karpouzis, K., Abrilian, S. (2006). Manual Annotation and Automatic Image Processing of Multimodal Emotional Behaviors: Validating the Annotation of TV Interviews. In Fifth international conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy.

McNeill, D. (2005). *Gesture and thought*. Chicago: University of Chicago Press.

Poggi, I., Magno Caldognetto, E. (1996). A score for the analysis of gestures in multimodal communication. In: Messing, L. (Ed.), *Proceedings of the Workshop on the Integration of Gesture and Language in Speech. Applied Science and Engineering Laboratories*. Delaware: Newark and Wilmington, pp. 235--244.

Rodríguez, K., Stefan, K. J., Dipper, S., Götze, M., Poesio, M., Riccardi, G., Raymond, C., Wisniewska, J. (2007). Standoff Coordination for Multi-Tool Annotation in a Dialogue Corpus. In *Proceedings of the Linguistic Annotation Workshop at the ACL'07* (LAW-07), Prague, Czech Republic.

Smith, M. L., Cottrell, G. W., Gosselin, F., Schyns, P. G., (2005). Transmitting and Decoding Facial Expressions. *Psychological Science*, 16, 3, pp. 184--189.

Susskind J.M., Littlewort G., Bartlett M.S., Movellan J., Anderson A.K. (2007). Human and computer recognition of facial expressions of emotion. *Neuropsychologia*, 45, pp. 152--162

Double level analysis of the Multimodal Expressions of Emotions in Human-Machine Interaction

Jean-Marc Colletta [1], Ramona Kunene [1], Aurélie Venouil [1], Anna Tcherkassof [2]

[1]Lidilem, Université Stendhal, BP25, 38040 Grenoble Cedex 9, France
[2] LPS, Université Pierre Mendès France, BP47, 38040 Grenoble Cedex 9, France
e-mail: Jean-marc.colletta@u-grenoble3.fr, kuneneramona@yahoo.com, a.venouil@free.fr, anna.tcherkassof@upmf-grenoble.fr

Abstract

This paper presents the method and tools applied to the annotation of a corpus of multimodal spontaneous expressions of emotions, aimed at improving the detection and characterisation of emotions and mental states in human-machine interaction. The annotation of multimodal corpora remains a complex science as the preparation of the analysis tools have to be in line with the objectives and goals of the research. In human expressions and emotions the verbal and non verbal behaviour all play a crucial role to reveal the mental state of a speaker and as such voice, silences, hesitations from the verbal aspect, and every movement from the scratching of one's eye to the movement of toes from the non verbal aspect, have to be taken into consideration. The physical description of the bodily movements, although necessary, remains approximative when based on 2D and lacks the analytical aspects of human behaviour. In this paper we define a two-level procedure for the annotation of the bodily expressions of emotions and mental states, as well as our annotation grid for speech cues and body movements.

1. Introduction

This paper presents the annotation procedure of a corpus of multimodal spontaneous expressions of emotions and mental states in human-machine interaction.

The corpus collection was part of a study on the fusion of multimodal information (verbal, prosodic, facial, gesture, posture and physiology) to improve the detection and characterisation of expressions of emotions in human-machine interaction (Le Chenadec, Maffiolo & Chateau, 2007).

The overall objective was to develop computer systems which can « perceive and understand human behaviour and respond appropriately » (Le Chenadec, Maffiolo, Chateau & Colletta, 2006). In the optic to develop these affective computer systems which detect and characterise expressions of emotions and mental states, the data collected had to reflect the multimodal character of human behaviour.

The annotation considerations were to identify the mental and emotional states from a video corpus of 18 adults.

The next section presents an overall view of the data collection and the methodological aspects of this study. The following sections discuss annotation procedures and present the annotation scheme we created for this study.

2. Elicitation Methodology

The experimental setup has been well presented in Le Chenadec, Maffiolo, Chateau & Colletta, (2006) and Le Chenadec, Maffiolo & Château, (2007). Here we give a brief recapitulative.

The objectives of the corpus collection were threefold: the range of emotional and mental states had to be widest as possible, emotions and mental states had to be expressed freely and spontaneously, and expressions had to be multimodal through vocal, gesture, postural, facial, physiological behaviour.

This experiment was conducted in a laboratory test platform based on the Wizard-of-Oz methodology, in which an interaction between a human and a virtual character could be set up. In this experiment the interaction was on the repetition of a play. The instructions given to the users were to play three scenes of *Don Quixote de la Mancha*, written by M. de Cervantes (1605). The human subject was to play the part of Sancho Panza and had to give his cue to the virtual character as Don Quixote. Cues from the virtual actor were controlled by the experimenter in real time. The experimenter simulated an autonomous system.

In order to elicit spontaneous emotional expressions of users, different system bugs were designed: uncoordinated movements or stammering of the virtual actor and the request to the user was to repeat his cue, or if the system displayed "lost data", the request to the user was to repeat one of the three scenes of the act. From the users' perspective, some bugs were clearly related to a system's failure, other bugs were perceived to be a result of their mistakes of their cues. They were expected to express the emotional feelings and mental states they experienced as a result of being confronted with these bugs, which were designed to be funny, or boring, repetitive or deeply annoying.

The multimodal behaviour of each user was recorded with two digital cameras (head-only and upper body) and a microphone. Eighteen actors (nine females and nine males, aging from 25 to 50 years) took part in the experiment. Interview recordings lasted 1h15mn for each participant.

The data collected during the experiment were completed by the gathering of the user's viewpoint

immediately after the interaction with the virtual character (Le Chenadec, Maffiolo & Chateau, 2007). Each user was asked to comment what he/she felt during the interaction while viewing its recording, and to determine the starting and ending time where he/she experienced these feelings. A subsequent interview was conducted with a close relative of each user. The recording of the interaction was played back to this relative who was asked to comment on the behaviour of the user using the same method.

Finally, a categorisation experiment was conducted at the LPS laboratory, Université Pierre Mendès France, Grenoble. The same recordings were played back twice to 90 third-party observers, all students in social psychology. The first viewing allowed each subject to familiarise him/herself to the idiosyncratic behavioural characteristics of the user observed. During the second viewing, he/she was asked to stop the video each time he/she observed that the user felt something, i.e. experienced an emotional or a cognitive state. He/she then had to attribute an emotional or cognitive value to the observed behaviour and indicate his/ her starting and ending time.

The next section discusses the key factors applied to the annotation process of the data collected during the experiment.

3. Transcription Considerations

Currently, several researchers are interested in the multimodal complexity processes of oral communication. This issue has brought about increased interest to researchers aiming to transcribe and annotate different kind of multimodal corpora. Some researchers, as Abrilian (2005), work on the annotation of emotional corpora in order to examine the relation between multimodal behaviour and natural emotions. Other researchers working in the field of autism (*inter alia* Grynszpan, Martin & Oudin, 2003) or language development (Colletta, 2004; Colletta et al, this symposium) also takes into consideration these multimodal clues in their studies. Researchers in computer sciences take into account the multimodal clues in order to improve the ECAs – Embodied Conversational Agents – (Hartmann, Mancini & Pelachaud, 2002, 2005; Ech Chafai, Pelachaud & Pelé, 2006; Kipp, Neff & Albrecht, 2006; Kipp et al., 2007; Kopp et al., 2007; Vilhjalmsson et al., 2007).

It is without doubt that these methods and tools of annotation have paved the way for more interesting exploratory means to study multimodal corpora in detail. However, some theoretical and methodological difficulties still arise when one tries to annotate body movements. We will discuss these points in the following section 3.2.

The 18 recordings of the interactions between the subjects and the theatrical application, treated by the Lidilem laboratory, Université Stendhal, Grenoble, were specifically dedicated to the obtaining of the multimodal expressions of spontaneous emotional and mental states.

Two kinds of annotations were conducted: an annotation of each user's speech as well as other paraverbal phenomena – prosodic and voice considerations –, and an annotation of their corporal behaviour throughout the repetition of the play experiment.

3.1. Verbal and prosodic annotation

Linguistic and prosodic attributes often betray the emotional as well as the mental state of the speaker's mind. Each emotion has its words and verbal expressions, as research on the semantics of emotion show (Galati & Sini, 2000; Plantin, 2003; Tutin, Novakova, Grossmann & Cavalla, 2006). Stronger cues are supported by the voice features (Lacheret-Dujour & Beaugendre, 1999; Aubergé & Lemaître, 2000; Scherer, Bänziger & Greandjean, 2003; Keller et al., 2003; Shochi, Aubergé & Rilliard, 2006). In fact, all aspects of prosody may contribute to express emotional and mental states: pitch, intensity, speech rate, hesitations, grunts, various mouth and throat noises, etc.

Our verbal annotation was done on the software *PRAAT* developed by P. Boersma and D. Weenink [1]. As Table 1 shows (see: annexures), we did not code for pitch, intensity or rate as this data can be directly collected from the speech signal analysis. We coded for silent and filled pauses, linguistic errors, unexpected articulation of words, false starts, repetitions, laughs, coughs and sighs, all linguistic or posodic cues which may be an indication of reflection, embarrassment or various emotions.

3.2. Non-verbal annotation

The verbal transcriptions were aligned and imported to the software *ANVIL* developed by M. Kipp [2]. All recordings with their corresponding visual components were annotated accordingly in respect of the non-verbal performance of the subject.

As gesture researchers have already demonstrated in the past, all bodily movements may help express attitudes, emotional and mental states (Feyereisen & De Lannoy, 1985; Kendon, 1990, 2004; Feldman & Rimé, 1991; Descamps, 1993; Cosnier, 1994; Plantin, Doury & Traverso, 2000; Knapp & Hall, 2002). Attitudes and postures were correlated to mental states, pathologies and emotional disposition of the subjects; the gaze contributes to the expression of emotion and its appearance was correlated to the levels of activation or attention of the subject; the facial expressions exteriorise the whole range of emotions and feelings. Finally, among gestures, we can observe that some appear more frequently in stressful situations and are correlated to anxious states and certain affects: the gestures which are self centred, which include the

[1] Available from http://www.fon.hum.uva.nl/praat/

[2] Available from http://www.anvil-software.de/

gestures of self-contact (rub oneself on the face, scratching oneself, massaging oneself) and the gestures of manipulation of objects (playing with his/her keys, fiddling with his/her pen).

To annotate the non-verbal behavioural features of the subjects in this study, the coding scheme used was divided in 16 tracks (see Figure 1 and Table 2: annexures) representing all different parts of the human body. Our annotation grid was thus split into:

(i) self-contact gestures and auto-manipulations;
(ii) posture attitudes and changes (2 tracks);
(iii) head gestures (2 tracks);
(iv) gaze direction and changes (2 tracks);
(v) facial expressions (2 tracks);
(vi) torso movements;
(vii) shoulders movements;
(viii) arms location and movements;
(ix) hand gestures (2 tracks);
(x) lower body movements;
(xi) gestures performed by the actor while giving his clues to the animated character and part of his acting.

Each subject file had two subfiles; a video with both the face and body which allowed to annotate all the above mentioned body part, and a purely facial video to allow for precise, accurate and detailed coding of facial expressions.

From an etymological perspective (Pike, 1967), to obtain an annotation of the mental and emotional state behaviour of the speaker, an *etic* approach is necessary which will emphasise the physical aspects of the movement and allow for a microanalytical description. Researchers in gesture synthesis all agree on the necessity to rely on physical and accurate descriptions of the body movements. The transcription tools they propose all annotate for the body parts, as: gesture, gaze, head, torso, face, legs, lips and other behaviour (Vilhjalmsson et al., 2007). They also annotate for various location and movement parameters. For instance, to annotate for gesture expressivity, Hartmann, Mancini and Pelachaud (2005) distinguish between overall activation, spatial and temporal extent of the movement, fluidity (smooth *vs*. jerky), power (weak *vs*. strong) and repetition. To annotate for hand gestures, the researchers trying to unify a multimodal behaviour generation framework called the "Behavior Markup Language" (Kopp et al., 2006; Vilhjalmsson et al., 2007) mention the following parameters: wrist location, trajectory of movement, hand shape, hand orientation. Kipp, Neff & Albrecht (2006) propose to annotate for "handedness", trajectory, hand location (height, distance, and radial orientation), position of the arm (arm swivel) and hand to hand distance for a two hands gesture. When the annotation of hand gestures aims at studying the relationship between gesture and speech (see McNeill, 1992, 2005; Colletta, 2004), it also requires temporal information about the phases of the gesture phrase realisation, as first described by Kendon (1972, 1980) and integrated in gesture synthesis by Kipp, Neff & Albrecht (2006).

In our grid (Table 2) , the *etic* approach is displayed under all tracks except those which are subtitled "function", and it gives information on :

(i) the body part and its location (for an arm or a hand gesture),
(ii) direction of the movement,
(iii) characteristic of the movement (swaying, frowning, shrugging, etc.),
(iv) shape of the movement (for a hand gesture),
(v) speed of the movement, and
(vi) frequency of occurrence when the movement is repeated.

However the *etic* approach is not sufficient to present a comprehensive description of bodily behaviour. Kendon (1990), in line with other researchers, have pointed that in everyday life we "read" bodily behaviour of others through mentally precategorised concepts; such as laughing, smiling, ease, nodding, pointing, gesturing, miming, etc.

Each concept covers a range of behaviours, whether small or large, whose physical properties may vary in proportion. For instance, I can smile with a closed mouth or with an open mouth; I can express a subtle smile or a broad smile; I can express a mouth-only smile or be all smiles, etc. Yet all these various forms of smiles are examples of the same broad expressive category called "smile". As for a pointing gesture, I can point with a hand or a head or the chin; I can point to an object or a person present in the physical setting, or to a direction; I can point to an object or person with an extended hand or just with an extended index finger; I can point once to an object or person, or point repetitively to it, etc. There again, all these various forms of pointing share the same function and are exemplars of the category called "pointing gesture".

At this point, it is worth noting that the researchers who currently aim at unifying a multimodal behaviour generation framework for ECAs (Vilhjalmsson et al., 2007) « have proposed knowledge structures that describe the form and generation of multimodal communicative behaviour at different levels of abstraction ». The first level represents the interface between planning communicative intent and planning the multimodal realisation of this intent, and is mediated by the "Functional Markup Language" (FML). The second level represents the interface between planning the multimodal realisation of a communicative intent and the realisation of the planned behaviours, and is mediated by the "Behaviour Markup Language" (BML). Although the FML remains largely undefined in the authors work, the FML/BML distinction surprisingly resembles Kenneth Pike's distinction between the *emic/etic* levels of behaviour description.

In our view, a more *emic* approach (Pike, 1967) is thus essential to annotate the body movements that express the mental and emotional state behaviour of the speaker, and to complement the *etic* physical description of

these movements. In our grid (see Table 2: annexures), this approach is displayed under all tracks which are subtitled "function" and it serves as an indication of:

(i) a general behaviour or attitude (scratching, touching, handling, comfort posture…);
(ii) a significant head movement (head nod, head shake, head beat, deictic or pointing movement);
(iii) a gaze behaviour (waiting, reading, staring, scanning);
(iv) a significant facial expression (smile, laughter, biting, pursing, licking lips, pouting);
(v) a coverbal hand gesture (deictic or pointing movement, beat, iconic gesture, metaphoric gesture, interactive gesture.).

During the annotation process, every body movement was annotated for its *etic* or physical properties as well as for its *emic* properties or emotional/function properties.

4. Transcription and Validation

Coders selected for the annotation had previous experience in gesture and emotion studies. Additional training on annotation tool was included to familiarise them with the *ANVIL* software as well as with the video data. File sequences were initially transcribed manually on *Excel*, in which the coders would first examine the video files and have a global view of the frequency and nature of movements in order to prepare the relevant grid.

The non verbal transcription was then carried out in parallel by two coders. Each coder annotated independently from the other coder. In most cases, the validation of an annotation scheme is based on the comparison of the annotations done by two independent coders. This method is useful to test the validity of an annotation scheme, but it does not allow to check and to stabilise the analysis of a corpus at the end of an annotation procedure. Indeed, in our case, it is not a question of testing a body movement annotation grid, but it is rather a question of validating the annotation of a multimodal corpus before using the results of the annotation in a study on the fusion of multimodal information (Le Chenadec, Maffiolo & Château, 2007). As a consequence, a third coder was asked to finalise the annotation from choices made by both coders and decide in case of disagreement.

Having a two-stage process with the independent coding as well as the decision stage cannot ensure that this analysis procedure is a hundred percent conclusive. To annotate for emotions and mental states is to observe the whole body, including the problem of identifying the movements, which does not arise when we annotate for precise gestures (e.g., the coverbal hand gestures). On the other hand, another means of validation is to cross-check the information resulting from the annotation by the coders with other data sources. For this study on the fusion of multimodal information, the other available data source is (1) the collection of the user's viewpoint after the experiment, completed by interviews with their relatives, and (2) a categorisation experiment conducted with 90 third-party observers (see section 2 for more details). In the end, it will be most interesting to compare the transcriptions by the three coders to the analysis performed by the users and their relatives on one side, and by the 90 students, on the other side.

5. Final remarks

Our paper describes the method and the analysis tools applied as well as the annotating considerations we employed. Our aim is to enhance the understanding of the technical issues surrounding the annotation of a multimodal corpus. Annotating mental and emotional states of mind in adults requires a vigorous approach and attention to detail. The objectives of this research required the minute examination of: the voice, linguistic features, sounds or the absence of sounds as all these play a role in revealing the emotional and state of a speaker. In verbal annotation, we observed all the linguistic and prosodic cues as they offer us a window to the state of nervousness, anxiety, irritation, humour, etc.

The non verbal annotation also required a vigorous if not somewhat lengthy approach. If one seeks the understanding of gesture related to speech it would be much simpler to annotate for hand and head movements, and stick to communicative or representational gesture. In this study, the quest for emotions and human mental states showed that each and every part of the body from the head to the toes has a story to reveal. The grid used on *ANVIL* enabled us to annotate this rather complex set of movements as the human speaker is in constant motion, from scratching his head in anxiety to smiling in contentment.

Our analysis procedure aimed at using the double level (*etic/emic*) annotation, which we hope, will help to enhance in the designing of annotation tools. The missing puzzle remains in the cross-validation from several data sources.

6. Acknowledgements

This research was conducted and financed by the France Telecom R & D, Lannion. The authors thank Valérie Maffiolo and Gilles Le Chenadec for the designing of the experiment and for contributing to the creation of the annotation grid.

7. References

Abrilian, S. (2005). Annotation de corpus d'interviews télévisées pour la modélisation de relation entre comportements multimodaux et émotions naturelles. *6ème colloque des jeunes chercheurs en Sciences Cognitives* (*CJCSC'2005), Bordeaux, France.*

Aubergé, V., Lemaître, L. (2000). The Prosody of Smile. In *Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, sept. 5th-7th, 2000*, pp. 122--126.

Colletta, J.-M. (2004). Le développement de la parole chez l'enfant âgé de 6 à 11 ans. Corps, langage et cognition. Hayen, Mardaga.

Cosnier, J. (1994). *Psychologie des émotions et des sentiments*. Paris, Retz.

Descamps, M.-A. (1993). *Le langage du corps et la communication corporelle*. Paris, P.U.F.

Ech Chafai, N., Pelachaud, C., Pelé, D. (2006), Analysis of Gesture Expressivity Modulations from Cartoons Animations. In *LREC 2006 Workshop on "Multimodal Corpora", Genova, Italy, 27 May*.

Feldman, R., Rimé, B., Dir. (1991). *Fundamentals of non verbal behaviour*. Cambridge, Cambridge University Press.

Feyereisen, P., De Lannoy, J.-D. (1985). *Psychologie du geste*. Bruxelles, Pierre Mardaga.

Galati, D., Sini, B. (2000). Les structures sémantiques du lexique français des émotions. In. C. Plantin, M. Doury, V. Traverso, *Les émotions dans les interactions*. Presses Universitaires de Lyon.

Grynszpan, O., Martin, J.C., Oudin, N. (2003). On the annotation of gestures in multimodal autistic behaviour. In *Gesture Workshop 2003*, Genova, Italy, 15-17 April.

Hartmann, B., Mancini, M., Pelachaud, C. (2002). Formational Parameters and Adaptative Prototype Instanciation for MPEG-4 Compliant Gesture Synthesis. *Computer Animation Proceedings, Genève, June 2002*.

Hartmann, B., Mancini, M., Pelachaud, C. (2005). Implementing Expressive Gesture Synthesis for Embodied Conversational Agents. *Gesture Workshop, LNAI, Springer, May 2005*.

Keller, E., Bailly, G., Monaghan, A., Terken, J., Huckvale, M. (2003). *Improvements in speech synthesis*. Chichester, UK, John Wiley.

Kendon, A. (1972). Some relationships between body motion and speech. In A.W. Siegman et B. Pope (eds.), *Studies in dyadic communication*. Elmsford, NY, Pergamon Press, pp. 177--210.

Kendon, A. (1980). Gesticulation and speech, two aspects of the process of utterance. In M.R. Key (ed.), *The relationship of verbal and nonverbal communication*. The Hague, Mouton, pp. 207--227.

Kendon, A. (1990). *Conducting interaction. Patterns of behavior in focused encounters*. Cambridge, Cambridge University Press.

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

Kipp, M., Neff, M., Albrecht, I. (2006). An Annotation Scheme for Conversational Gestures: How to economically capture timing and form. In *Proceedings of the Workshop on Multimodal Corpora (LREC'06)*, pp. 24--27.

Kipp, M., Neff, M., Kipp, K.H., Albrecht, I. (2007). Towards Natural Gesture Synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis. In C. Pelachaud et al. (eds.), *Intelligent Virtual Agents 2007, Lecture Notes in Artificial Intelligence 4722*. Berlin, Springer-Verlag, pp. 15--28.

Knapp, M**.**, Hall, J. (2002). *Nonverbal communication in human interaction.* Harcourt College Publishers.

Kopp, S., Krenn, B., Marsella, S., Marshall, A.N., Pelachaud, C., Pirker, H., Thórisson, K.R., Vilhjálmsson, H. (2007). Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. In J. Gratch et al. (eds.), *Intelligent Virtual Agents 2006, Lecture Notes in Artificial Intelligence 4133*. Berlin, Springer-Verlag, pp. 205--217.

Lacheret-Dujour, A., Beaugendre, F. (1999). *La prosodie du français*. Paris, CNRS Editions.

Le Chenadec., G., Maffiolo V., Chateau N. (2007). Analysis of the multimodal behavior of users in HCI : the expert viewpoint of close relations. *4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, 28-30th June, Brno, Czech Republic.*

Le Chenadec, G.., Maffiolo, V., Chateau, N., Colletta, J.M. (2006). Creation of a Corpus of Multimodal Spontaneous Expressions of Emotions in Human-Interaction. In *LREC 2006, Genoa, Italy.*

McNeill, D. (1992). Hand and mind. What gestures reveal about thought. Chicago, University of Chicago Press.

McNeill, D. (2005). *Gesture and thought.* Chicago, University of Chicago Press.

Pike, K.L. (1967). *Language in relation to a unified theory of the structure of human behavior*. Janua Linguarum, series maior, 24. The Hague: Mouton.

Plantin, C. (2003). Structures verbales de l'émotion parlée et de la parole émue. In J.-M. Colletta, A. Tcherkassof, *Les émotions. Cognition, langage et développement*. Hayen, Mardaga, pp. 97--130.

Scherer, K.R., Bänziger, T. Grandjean, D. (2003). L'étude de l'expression vocale des émotions : mise en évidence de la dynamique des processus affectifs. In J.M. Colletta, A. Tcherkassof, *Les émotions. Cognition, langage et développement*. Hayen, Mardaga, pp. 39--58.

Shochi, T., Aubergé, V., Rilliard, A. (2006). How Prosodic Attitudes can be False Friends: Japanese vs. French social affects. *Proceedings of Speech Prosody 2006, Dresden*, pp. 692--696.

Tutin, A., Novakova, I., Grossmann, F., Cavalla, C. (2006). Esquisse de typologie des noms d'affect à partir de leurs propriétés combinatoires. *Langue Française*, 150, pp. 32--49.

Vilhjalmsson, H., Cantelmo, N., Cassell, J., Chafai, E.N., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A.N., Pelachaud, C., Ruttkay, Z., Thórisson, K.R., van Welbergen, H., van der Werf, R.J. (2007). The Behavior Markup Language: Recent Developments and Challenges, in C. Pelachaud et al. (eds.), *Intelligent Virtual Agents 2007, Lecture Notes in Artificial Intelligence 4722*, Berlin, Springer-Verlag, pp. 99--111.

Annexures

Type of annotation	**Name of phenomenon**	**Definition**
Prosody	Silent pause	pause in the middle of a speech segment
	Intelligible pause	silence voluntarily added in the middle of a speech segment
	Pause filler	"euh" ou "hum"
Linguistic	Commentary	commentary on the interaction
	Error	error in syllable or word pronounciation
	Unexpected articulation	pronounciation of an unsual final syllable with a silent "e.
	False start	a "*" attached to the word + annotate the complete word sequence
	Elision	presence of elision
	Recovery	reformulation of a portion of a speech segment
	Repetition	repetition of a portion of a speech segment
	Incomprehensible words	transcription of an impossible word or speech segment
Dialogue	Repétition	repetition of the identical
	Reformulation	repetition of response with other terms
Sounds	Sounds of the system	
	Speech cuts	the virtual actor cuts the live actor's speech
	cough, throat, mouth	cough, throat clearing, noise made by the mouth
	Laugh	
	Exhalation, breath, sigh	
	Inhalation	

Table 1 : Verbal and prosodic annotation grid

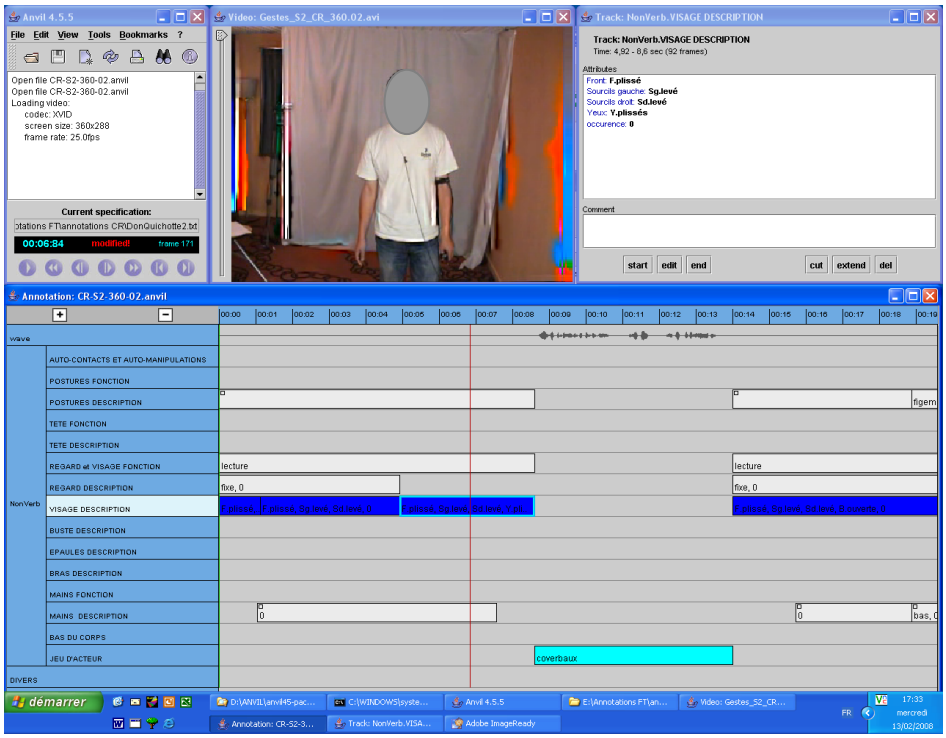

Figure 1: Anvil Screen

Type of annotation	Name of phenomenon
1- Self-contact gestures & auto-manipulations	Action: scratch/ touch/ twist/ handle
	Body part location: hair/temple/brow/glasses/ nose
2- Posture (function)	Comfort/ stretching
3- Posture (description)	Pattern : swaying/ complex movement/ freezing
	Leg movements: frontward/ backwards/ left/ right
	Speed: slow/ normal/ fast
4- Head (function)	Movement : nod/ shake/ beat/ deictic
5- Head (description)	Tilted high/ low
	Turn: left/ right
	Complex movement: front / backward
	Single movement: up / down
	Single movement: front / backward
	Single tilt: left/ right
	Single side-turn: left/right
	Speed: slow/ normal/ fast
6- Gaze (function)	Characterisation: waiting/ reading/ staring/ scanning
7- Gaze (description)	Direction: up/ down
	Direction: left/right
	Movement: sweeping/ rolling eyes
	Speed: slow/ normal/ fast
8- Face (function)	Smile, laughter/ biting/ pursing/ licking lips/pouting
9- Face (description)	Brows: frowning
	Left eyebrow: raising / frowning
	Right eyebrow: raising/ frowning
	Eyes: closing / opening/ wide opening/ rolling/ blinking/ winking
10- Torso (description)	Movement: forward/ backward
	Movement: left/right
	Unsteady movement
	Bend: forward/ backward
	Turn:left/ right
	Twist: left/ right
	Side: left/ right
	Position: protruded/ retracted
	Speed: slow/ normal/ fast
11- Shoulders (description)	Identification: left/ right/both
	Description: shrugging/ sagging
	Number: left/ right/ both
	Occurrence: 0 to 5
	Speed: slow/normal/fast
12- Arms (description)	Left-arm direction: going up/down, moving sideways, forwards, backwards, to the side, up, not moving
	Left-arm position: bent, half-bent, stretched out
	Right-arm direction: going up / down, moving sideways, forwards, backwards, to the side, up, not moving
	Right-arm position: bent, half-bent, stretched out
	Both arms action: crossing
	Occurrence: 0 to 5
	Speed: slow/ normal/ fast
13- Hands (function)	Deictic, beat, iconic, metaphoric, interactive
14- Hands (description)	Left hand action: rotation, opening, closing
	Left-hand direction: up/ down/left/ right/ forward/ backward
	Left-palm direction: Left-hand direction: up/ down/left/ right/ forward/ backward
	Right hand action: rotation, opening, closing
	Right-hand direction: up/ down/left/ right/ forward/ backward
	Right-palm direction: up/ down/left/ right/ forward/ backward
	Occurrence: 0 to 5
	Speed: slow/ normal/ fast
15- Lower body	Free comments
16- Acting	Mime, exaggerated gestures and expressions

Table 2: Coding scheme for the non verbal annotation grid.

Multimodality in Conversation Analysis: A case of Greek TV Interviews

Maria Koutsombogera†*, Lida Touribaba*, Harris Papageorgiou†
†Institute for Language and Speech Processing
Artemidos 6 & Epidavrou, 15125 Maroussi, Athens
*University of Athens, Department of Linguistics
University Campus, 15784 Ilissia, Athens
E-mail: {mkouts,lida,xaris}@ilsp.gr

Abstract

In this paper we present a study for the role of multimodal expressions in face-to-face TV interviews. Specifically, we examine the type of facial displays and gestures and their respective communicative functions in terms of feedback and turn management in an attempt to develop a deeper analytical understanding of the mechanisms underlying the multimodal aspects of human interaction in the context of media communication. Taking into account previous work on the analysis of non-verbal interaction, we discuss the distribution of the features of interest and we investigate the effect of the situational and conversational setting of each interview on the interactional behavior of the participants. We describe the tools and the coding scheme employed, we report on the results of the annotation and, finally, we conclude with comments on future work and exploitation of the resulting resource.

1. Introduction

Multimodal communication is a rapidly evolving field that has been addressed by complementary groups working on such matters from different perspectives, either theoretical or algorithmic. In this paper we present a cross-disciplinary research on the communicative role of multimodal expressions in TV face-to-face interviews occurring in various settings.

In general, TV discussions present a mixture of characteristics oscillating between institutional discourse, semi-institutional discourse and casual conversation (Heritage, 2005; Ilie, 2001). This media content spans a variety of discourse types such as information, entertainment and casual talk.

TV interviews are subject to certain restrictions such as time (duration of the show, interruptions for commercial breaks), agenda (topicalization) and technical features (camera direction and focus, editing) that further influence turn management in all its expressive dimensions (speech, gestures, facial displays).

The setting in which the interview takes place, the social and discursive roles of the speakers and the relevance of the topic to the individual participants are features that formulate not only the discourse structure, but the multimodal behavior of the speakers as well.

Our motivation is to identify and interpret gestural and facial features that critically contribute to the structure and content of a message. Specifically, we describe their interrelations as well as their distribution across the data in an attempt to find evidence about their potential systematic role.

Moreover, we explore the interactive behaviors in order to attest multimodal feedback and turn-taking with regards to different types of TV discussions. In this context, we take a first step towards the description and annotation of a multimodal corpus of Greek TV interviews, available for further development and exploitation.

2. Corpus Description

The corpus comprises 66 minutes of audiovisual material corresponding to interviews extracted from 3 different Greek TV shows. The interviews exhibit a one-to-one interactional pattern. The structure consists of question-answer sequences performed by an interviewer (the *host*) to an interviewee (the *guest*). No audience is present in any of the three discussions.

Apart from the commonly shared features, each interview has a unique nature outlined by the setting, the topic, the roles and personalities of the speakers, their interests and their commitments.

The first interview (I1) takes place in a TV studio between the host and a politician and provides information concerning current political issues. It can be regarded as a more *institutionalized* interaction, as it appears to be more standardized in role distribution and turn pre-allocation.

The second one (I2) is a pre-filmed informal discussion between the host and an entertainer. The interview setting is a classroom, where the guest gives an account of social and personal issues and is subsequently confronted with the host's reactions and suggestions. Due to the spontaneous and intimate character of the interaction, this type of interview is oriented towards *casual conversation*.

The third interview (I3) is a discussion of intellectual bias between the host and a writer taking place in an office. It displays a semi-institutional character, because apart from the evaluation of the writer's production, the discussion is not strictly information-focused and it promotes the personal and emotional involvement of both speakers, allowing spontaneous and unpredictable behavior to be expressed.

Furthermore, I1 is more host-controlled and therefore shows certain predictability in turn management, whereas I2 and I3 are more participant-shaped and present a relatively lower degree of predictability and weaker talk control.

As far as the conversational behavior of the hosts is concerned, in I1 the host assumes a strictly institutional role, in I2 the interviewer displays active personal involvement in the discourse in a way similar to informal conversations and in I3 he assumes a semi-institutional role, acting partly in his professional or expert role, and partly in his social role as an individual.
Although there are clearly distinctive features between the three types of interviews, we keep in mind that there might be certain deviations deriving from the fact that casual talk often exhibits an institutional character and, at the same time, institutional talk acquires less formal attributes.

3. Coding Scheme

For the labeling of the elements of interest, we adopted the MUMIN coding scheme v.3.3 (Allwood et al., 2005), which serves as a general instrument for the study of gestures and facial displays in interpersonal communication, focusing on the role of multimodal expressions for feedback, turn management and sequencing. In our study, we slightly modified the scheme by merging certain features and their respective terminal values. Moreover, for the description of multimodal relations we opted for those proposed in Poggi and Magno Caldognetto (1996), also adopted in MUMIN v.1.3. Finally, we added a level for body posture annotation; although in most TV settings the speaker is visible from the waist up, we noticed that the torso movement was quite often used by the speakers in order to reinforce their message.
Our research focuses on how communication is accomplished through multimodality. In this sense, we are not interested in annotating neutral and mechanical expressions or reactions to physical or external stimuli, e.g. noise, light etc.**,** that do not account for communicative functions.
The annotation scheme provides a generalized perspective for the description of features. We deal with the multimodal expressions as macro functional blocks; we do not intend to dissect and deconstruct them into the micro processes these expressions encapsulate. This means that the temporal structure of the expressions (preparation, stroke and retraction phases) is not described. This choice is determined by the purpose of the task (the communicative perspective of the material) and the nature of the data as well; the image broadcasted in TV shows depends on the movement of the camera (e.g. focus on one speaker at a time), obliging the annotator to deal with what he sees, even if that constitutes only a part of the actual expression; thus, the internal structure of an annotated broadcasted gesture does not coincide with the actual gesture. However, we anchor the start and end point of each expression and we ensure that the relevant annotation layers are synchronized with it. The duration of the expressions can be drawn from the links to the time stamps.

4. Annotation Process

We examine multimodality by discriminating between the description of the form of expressions and their respective communicative functions in a semantic and pragmatic context.
Each level of annotation is modality-specific and describes the expression types of speech, gestures, facial expressions and body posture for each speaker involved.
At a second level, the communicative functions that the aforementioned features represent were annotated according to the specified set of values.
The annotators' tasks involved the gradual annotation of the material in line with the coding scheme. Initially, the audio signal was extracted from the relevant video, orthographically transcribed and further enriched with information about pauses, non-speech sounds, etc. using the Transcriber tool[1]. The output was imported to ELAN[2], the tool that was used for the entire video annotation, and it was ensured that the speech transcript was kept synchronized with the video. Next, the annotators identified gestures and facial displays of interest marking their start and end points, assigned the respective tags and labeled the implied communicative functions (based on the interconnection between the facial display/gesture and the corresponding utterance). Finally, the data were revised in order to correct possible errors and to assess the consistency of the annotations throughout the corpus.

5. Results

In the interviews we studied, all participants frequently make use of multimodal expressions. However, the types and functions of facial and gestural expressions may vary, as they depend on the role that the speakers assume, the constraints imposed by the host and the discursive nature of the show from which the interview is taken.

5.1 Expression Types

The annotations of the distinct modalities reveal that the speakers employ simple or more complex expressions using their facial characteristics (gaze, eyebrows, nods), gestures (single or both hands, fingers, shoulders) and upper part of the body (leaning forward and backward) in order to reinforce their speech and express their emotions towards the uttered messages.
A closer look on the data shows that there are repeated patterns that are independent of the message content. For example, there are standard gestures in opening and closing utterances as well as in the hosts' prefatory statements (the gaze direction and hand orientation are towards the interlocutor, etc.). This sort of standardized multimodal behavior is aligned to the succession of turns across time and may lead to the formulation of possible conversational multimodal scenarios, e.g.:
- A question is formulated by the host and it is supported by certain multimodal expressions (e.g gaze towards interlocutor, torso leaned forward, eyebrows raising).
- The guest initially has an unfocused gaze as he contemplates on his answer or tries to find the appropriate words. When he finally takes the turn, he gazes towards the host and moves his hands or torso in order to denote

[1] http://trans.sourceforge.net/
[2] http://www.lat-mpi.eu/tools/elan/

that he has understood the question and is ready to answer. When he finds the exact words he expresses his certainty or opposition or any other feeling using facial displays. If he is willing to hold the turn he reinforces his speech with repeated hand gestures or eyebrows raising, while he is using mostly the eyes to elicit an acknowledgement or to ensure that the host keeps up with the conversation. Finally, he completes his turn either by gazing down, or by closing his mouth and staring at the host.

This kind of scenario describes a regular flow that is subject to modification in case of interruptions, overlapping talk, or strong objections and reactions that declare the speakers' emotional involvement.

Moreover, the timing and the extent of the turns of each interview type has a large effect in the production of non-verbal expressions. In case of I1, where the shifts are monitored, the guest has to express his views in a restricted time interval. Consequently, he cannot deviate from the agenda or be very spontaneous, and instead he is more reserved in his expressivity and makes a closed set of simple, accompanying gestures. Conversely, in I2 and I3 where the shifts and topics are more negotiated, the guests are entitled to elaborate on their views; they feel less restricted, more spontaneous, and thus more prone to produce a variety of facial and gestural expressions.

Semiotic types of non-verbal expressions seem to be independent of the setting; they are however related to the content and the discursive features of the interview. For example, I1 is an opinion interview, where the politician builds his argumentation based on concrete facts and supports it mainly by *non-deictic* expressions. At the same time, in more casual discussions such as I2 and I3 a large part of the guests' talk involves narration (guests' personal experiences), a discourse type that is complemented with *iconic* multimodal expressions. *Deictic* and *symbolic* types are quite equally distributed in the three interviews (cf. Table 1).

5.2 Communicative Functions

Our analysis focuses on the communicative type of the *multimodal relations* in order to attest the contribution of facial displays and gestures in turn management of TV interviews. There is a relatively high percentage of non-verbal expressions that complement the message by providing additional information that is not overtly expressed by speech only (*addition*) or by replacing unsaid words (*substitution*).The majority of *substitution* annotations is related to acknowledgements and is represented mostly by facial displays, usually corresponding to a head nod or a smile. The *repetition* type comes next, and it is used to denote that the non-verbal signal provides no extra information. Few relations were characterized as *neutral*, where the annotators believed that there is no significant relation among the distinct modalities. Finally, the *contradiction* relation is rarely used, namely in cases of irony. It is important to report that the *contradiction* type was found mainly in I1, possibly as a feature of argumentative speech, while it is rarely used in less institutionalized interviews such as I2 and I3.

Multimodal feedback in terms of *perception* and *acceptance*[3] is usually expressed through gaze and nods rather than gestures. Non-verbal expressions of *feedback give* evolve in the course of time, as the speaker denotes that he has perceived the message, he is interested or willing to contribute and, as the turn is elaborated, he shows signs of certainty about the content of his talk, possibly by forming an opinion.

Emotional/attitudinal feedback is closely related to the topic of discussion and the role that the participants assume. Positive emotional feedback is attributed to a large number of gestures and expressions of I2 and I3, where emotions/attitudes (e/a) such as joy and satisfaction are expressed more overtly. On the contrary, non-verbal declaration of negative e/a such as disappointment, anger etc. is manifested in I1. However, speakers in I1 shift to more positive feedback when personal and family matters are on the table.

	Value	I1	I2	I3
Semiotic Types	Deictic	5.2%	6.9%	**7.4%**
	Non Deictic	**88.1%**	73.1%	75.9%
	Iconic	4.6%	**15.9%**	13.8%
	Symbolic	2.1%	**4.1%**	2.9%
Multimodal Relations	Repetition	15.6%	**17.8%**	6.8%
	Addition	51.7%	54.6%	**71.4%**
	Contradiction	**1.7%**	0.4%	0.3%
	Substitution	19.8%	19.9%	**20.2%**
	Neutral	**11.2%**	7.3%	1.3%
Turn Management	Turn Take	**44.8%**	12.1%	16.4%
	Turn Yield	1.9%	8.2%	**9.2%**
	Turn Accept	2.7%	**8.8%**	7.4%
	Turn Hold	42.1%	**45.2%**	40.3%
	Turn Elicit	5.4%	**15.3%**	14.2%
	Turn Complete	3.1%	10.4%	**12.5%**
Feedback (Emotions/ Attitudes)	Positive[4]	4.4%	12.1%	**27.3%**
	Negative	**9.7%**	1.6%	1.9%

Table 1: Distribution of semiotic and communicative features over interview types. Highest values are highlighted in red.

Gestures and expressions that are evoked in turn management are different in type and frequency when they take place in a normal flow rather than overlapping talk. In I1 we rarely see expressions related to the unbiased completion of the turn and its subsequent yielding to the interlocutor. In most of the times, the speakers take the turn without explicitly being asked to do so by their interlocutors.

Overlapping speech is usually triggered by a pause, a hold or a repair, which are often accompanied by an unfocused

[3] The annotation values for Feedback pertain to 3 groups: *perception*, *acceptance* and *emotions/attitudes*.

[4] Percentage of positive (*happy, satisfied*) and negative (*angry, disappointed, disgusted*) e/a feedback calculated on all occurrences of e/a feedback. The remaining values either are not represented in the corpus (*sad, frightened*) or do not denote a clear positive or negative orientation (*surprised, certain, uncertain, interested, uninterested*).

gaze or a head side turn. The guest wants to hold the turn but sometimes he hesitates to answer or takes his time to focus, think or remember what to say. The host then takes advantage and he takes the turn. At this phase of the talk, it seems that speech and prosody features (e.g. higher intonation) are not enough; the speakers engage all their potentials to maintain their turn, including a variety of gestures and facial displays. This explains the fact that there is a high density of annotated multimodal expressions during overlapping speech. Usually, the speaker who manages to gain the turn is the one who makes the most gestures and facial displays. The distribution of communicative features across the 3 interviews can be shown in Table 1.

6. Conclusion

We presented an ongoing study on the analysis and interpretation of non-verbal modalities attested in three distinct types of Greek TV interviews, focusing on the functions of feedback and turn management. In order to provide more accurate and systematic descriptions of the multimodal features contributing to this kind of interaction we are planning to enrich the corpus with more interviews. The communicative function and role of speech features (disfluencies, prosody elements like pitch and loudness) should also be further explored.

Finally, we plan to exploit the aforementioned multimodal corpus in our multimedia processing framework. In the core of this platform lies an open, adaptable architecture that decides the way different metadata might be fused (Papageorgiou et al., 2005) in accordance with both the users' interests and digital equipment and the typology and semantic characteristics of the original audiovisual material. In this respect, conversation analysis metadata can be further exploited in order to accommodate multimedia retrieval & summarization applications (Georgantopoulos et al., 2006).

7. Acknowledgements

The authors wish to thank the reviewers for their helpful comments. The research described in this paper was supported by the research project "TV++" (A/V Digital Archive Management, funded in the framework of Measure 3.3 of the Operational Programme "Information Society" of the 3rd CSF) and by the Greek State Scholarship Foundation.

8. References

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P. (2005). The MUMIN Annotation Scheme for Feedback, Turn Management and Sequencing. In *Gothenburg papers in Theoretical Linguistics 92: Proceedings from The Second Nordic Conference on Multimodal Communication*, pp. 91-109.

Bertrand, R., Ferré, G., Blache, P., Ader, M., Espesser, R., Rauzy, S. (2007). Backchannels Revisited from a Multimodal Perspective. In *Proceedings of the Auditory-Visual Speech Processing Conference*.

Cerrato, L. (2004). A Coding Scheme for the Annotation of Feedback Phenomena in Conversational Speech. In *Proceedings of the LREC Workshop on Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interface*, .pp. 25-28.

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

Clayman, S., Heritage, J. (2002). *The News Interview: Journalists and Public Figures on the Air*. Cambridge: Cambridge University Press.

Douglas-Cowie, E. Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., Martin, J.-C., Devillers, L. Abrilian, S., Batliner, A. (2007). The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In *Proceedings of ACII 2007*. Berlin-Heidelberg: Springer, pp. 488-500.

Duncan, S. Jr. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology*, 23(2), 283-292.

Ekman, P. (1999). Emotional and Conversational Nonverbal Signals. In L.S. Messing and R. Campbell, *Gesture, Speech and Sign*, New York, Oxford University Press.

Georgantopoulos, B., Goedemé, T., Lounis, S., Papageorgiou, H., Tuytelaars, T., Van Gool, L. (2006). Cross-media summarization in a retrieval setting. In: *Proceedings of the LREC 2006 workshop on "Crossing media for improved information access"*, pp. 41-49.

Heritage, J. (2005). Conversation Analysis and Institutional Talk. In Robert Sanders and Kristine Fitch (eds), *Handbook of Language and Social Interaction*. Mahwah NJ: Lawrence Erlbaum, pp. 103-146.

Ilie, C. (2001). Semi-institutional Discourse: The Case of Talk Shows. *Journal of Pragmatics 33*, pp. 209-254.

MacNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, Chicago.

Martin, J.-C., Caridakis, G., Devillers, L., Karpouzis, K., Abrilian, S. (2007). Manual Annotation and Automatic Image Processing of Multimodal Emotional Behaviours: Validating the Annotation of TV Interviews. *Personal and Ubiquitous Computing*, Special issue on Multimodal Interfaces, Springer.

Papageorgiou, H., Prokopidis, P., Protopapas, A., Carayannis, G. (2005). Multimedia Indexing and Retrieval Using Natural Language, Speech and Image Processing Methods. In G. Stamou & S.Kollias (Eds.), *Multimedia Content and the Semantic Web: Methods, Standards and Tools*. Wiley, pp. 279 - 297.

Poggi, I., Magno Caldognetto, E. (1996). A Score for the Analysis of Gestures in Multimodal Communication. In *Proceedings of the Workshop on the Integration of Gesture and Language in Speech, Applied Science and Engineering Laboratories*. Newark and Wilmington, Del, pp. 235-244.

The MUSCLE movie database: A multimodal corpus with rich annotation for dialogue and saliency detection

D. Spachos, A. Zlatintsi*, V. Moschou, P. Antonopoulos, E. Benetos, M. Kotti, K. Tzimouli,

C. Kotropoulos, N. Nikolaidis, P. Maragos*, I. Pitas

Department of Informatics, Aristotle University of Thessaloniki, Box 451, Thessaloniki 541 24, Greece

E-mail: {dspachos, vmoshou, pantopo, empeneto, mkotti, kattzim, costas, nikolaid, pitas}@aiia.csd.auth.gr

* School of Electrical and Computer Engineering, National Technical University of Athens, Athens 157 73, Greece

E-mail: {nzlat,maragos}@cs.ntua.gr

Abstract

Semantic annotation of multimedia content is important for training, testing, and assessing content-based algorithms for indexing, organization, browsing, and retrieval. To this end, an annotated multimodal movie corpus, the so called MUSCLE movie database, has been collected to be used as a test bed for development and assessment of content-based multimedia processing, such as speaker clustering, speaker turn detection, visual speech activity detection, face detection, facial feature detection, face clustering, scene segmentation, saliency detection, and multimodal dialogue detection. All metadata are saved in xml format following the MPEG-7 ISO prototype to ensure data compatibility and reusability by different users and applications. The entire database can be downloaded through the web for research purposes. Furthermore, we describe a novel annotation tool called Anthropos7 Editor.

1. Introduction

The wide prevalence of personal computers, the decreasing cost of mass storage devices, and the advances in compression techniques have fuelled a vast increase in digital multimedia content, giving rise among others to online music and video stores, personal multimedia collections and video on demand. However, the convenience of multimedia libraries and the functionality of the aforementioned applications will be in doubt, unless efficient multimedia data management, necessary for organizing, navigating, browsing, searching, and viewing the multimedia content, is employed (Benetos, 2008). Multimedia standards such as MPEG-4 and MPEG-7 provide important functionality for manipulation and transmission of objects and the associated metadata, but the extraction of the semantic descriptions and the multimedia content is out of the standard scope (Chang, 2001).

In this paper, we present a large multimodal corpus that has been collected and annotated in order to test and assess different algorithms and hypotheses, such as actor clustering, visual speech detection, dialogue detection, or multimodal saliency detection. Rich annotation by multiple human annotators for concepts such as dialogue manifestations in audio and video, based on the level of background audio, presence of faces, presence of lip activity, is offered. Another concept that is defined in the database is saliency. The database covers 4 distinct modalities, namely audio, video, audiovisual, and text and offers annotated examples for the aforementioned concepts. We also describe a novel video annotation tool named Anthropos7 Editor, which offers capabilities for visual reviewing and editing of MPEG-7 data, following the MPEG-7 ISO format.

The outline of this paper is as follows. Section 2 lists some well known video and audio annotation tools and surveys ANVIL. It also provides a general overview of the Anthropos7 Editor with emphasis to data display and editing using Anthropos7 editor. Section 3 provides a description of the collected movie database. Finally, conclusions are drawn in Section 4.

2. Video annotation tools

A number of video annotation tools have been developed the past years. In addition to the tools reviewed in (Garg, 2004), we mention the following ones: IBM-MPEG-7 Annotation Tool, Ricoh – Movie Tool, ZGDV – VIDETO, COALA – LogCreator, and ELAN. Several factors influence the choice of the annotation tool. First, the tool must be able to support the annotation scheme. Second, it must be user friendly and, in many cases, compatible with other tools. Third, it is desired that the tool can transcribe both audio and video data. Finally, the tool must be suitable for several tasks, such as annotation of speakers and addressees as well as several types of dialogue acts (Garg, 2004). In the following, we survey ANVIL and describe the features of a novel annotation tool called Anthropos7 editor.

ANVIL is a free video annotation tool, used at research institutes world-wide. It offers frame-accurate, hierarchical multi-layered annotation driven by user-defined annotation schemes. The intuitive annotation board shows color-coded elements on multiple tracks in time-alignment. Special features include cross-level links, non-temporal objects and a project tool for managing multiple annotations. ANVIL can import data from the widely used, public domain phonetic tools PRAAT and XWaves, which allow precise and comfortable speech transcription. ANVIL's data files are xml-based. Special ASCII output can be used for import in statistical toolkits (like SPSS).

Anthropos7 Editor is an annotation tool for MPEG-7 advanced viewing and/or editing. It makes viewing and editing of MPEG-7 video content description an easy task.

Such a description can be related to time/duration of scenes and shots, frame-based information, such as the Regions of Interest (ROI) that encompass a specific actor in a frame, and high-level information regarding the video, such as the names of persons or actors appearing in the video. In order to visualize and manipulate time/duration-related information, Anthropos7 Editor uses the Timeline Area. Information based on a single frame, is visualized in the Video Area. Other static movie information, as well as duration and frame-based properties appear in the Static Information Area. These areas communicate with each other, automating various tasks and improving the way the user interacts with the Anthropos7 Editor. For example, the Static Information Area automatically shows the properties of the component the user interacts with; the Timeline area follows the playback of the Video Area. The user may also change the video position from the Timeline Area. Anthropos7 Editor uses overlays on top of the Video Area, e.g. it can visualize the ROI of each actor on every frame, if such information is present in the MPEG-7 file. The user can interact with these ROIs using the mouse. Every 2-D image region that encompasses an actor, or parts of actor's body defined in the Anthropos7 file can be overlaid on the corresponding video frame as a Polygon or a Box (rectangle) and the user can modify its position and its properties, such as the size of the box. A ROI (or parts of it) can be moved or deleted and new ROIs can be added. ROI edges can be also deleted or added. The application automatically tracks all these changes and saves them in the corresponding Anthropos7 file, an xml file in the MPEG-7 format. For more accurate editing, one can use the static ROI property window, which is opened as soon as the user clicks on a ROI. In the current version, ROIs are retrieved only according to the Anthropos7 description of the Actor Instance. No user defined schemas are supported. Apart from a drawn ROI, the name of the associated actor is also depicted on screen. This way, the end user can directly identify ROIs and actors, track face detection results and locate errors.

3. MUSCLE movie database specifications

The basic requirement for the movie database annotation is that the concepts (e.g. dialogue, saliency) must be described in each modality independently as well as in a cross-modal manner. This means that there must be audio-only and video-only descriptions, but audio-visual descriptions as well. This fact emerges from the research community needs to process the same data for different applications. Thus, several modalities along with the corresponding dialogue and saliency annotations are supported: audio-only, video-only, text-only, audio-visual. A more detailed description of these annotations is provided in subsections 3.1 and 3.2, respectively. The movie database and the xml annotation files can be downloaded for research purposes through the URL: http://poseidon.csd.auth.gr/EN/MUSCLE_moviedb.

3.1 Dialogue annotation

In total, 54 movie scenes of total duration 42 min and 41 sec have been extracted from 8 movies from different genres (Table 1). The audio language for all selected scenes is English. The duration of each scene is between 24-123 seconds and the scenes have been carefully selected to represent all possible cases. More details on the movie scenes are listed in Table 1. Each movie scene is separated in two different files: an audio file, which contains the audio of the scene and a video file, which contains the video of the scene without audio.

Movie title	Number of Dialogue scenes	Number of non-dialogue scenes	Scenes per Movie
Analyze That	4	2	6
Cold Mountain	5	1	6
Jackie Brown	3	3	6
Lord of the Rings I	5	3	8
Platoon	4	2	6
Secret Window	4	6	10
The Prestige	4	2	6
American Beauty	10	0	10
Total number of scenes	39	19	58

Table 1: MUSCLE movie database description

Different human annotators worked on the audio and video files. The dialogue type label was added to each one of the scenes (audio and video), one label per scene. The dialogue types for audio are as follows. CD (Clean Dialogue): Dialogues with low-level audio background; BD (Dialogue with background): Dialogue in the presence of a noisy background or music. A monologue is classified as either CM (Clean Monologue), i.e. monologue with low-level audio background or BM (Monologue with background), i.e. monologue in the presence of a noisy background or music. All scenes that are not labeled as CD or BD are considered to be non-dialogue (Non Dialogue - ND). The dialogue types for video are as follows. CD (Clean Dialogue): Two actors are present in the scene, their faces appear simultaneously or in an alternating pattern (A-B-A-B), and there is lip activity; BD (Dialogue with background): At least two actors are present, their faces appear simultaneously or in an alternating pattern in the scene and there is lip activity, while other actors, apart from the two that are engaged in the dialogue, appear. Large intervals where no dialogue occurs might be included in the scene. The monologue types for video are labeled as CM (Clean Monologue), i.e. one actor is present in the scene, his face is visible and there is lip activity or BM (Monologue with background), i.e. at least one actor is present, his face is visible and there is lip activity while other actors might appear and

large intervals where no dialogue occurs might be included in the scene. Similar to audio scenes, all video scenes that are not labeled as CD or BD, including monologues, are considered to be non-dialogue (Non Dialogue - ND).

The extracted annotation metadata for the audio files are speech activity data, namely speech intervals, defined from the start and the end time, for each actor in a scene. For the video files, lip activity data are extracted for each actor (2 actors in each scene maximum), defined through intervals specified by the start and end time and frame. The following three states are used to label each lip activity interval: 0 indicates that back of actor's head is visible; 1 indicates that actor's frontal face is visible, but no lip activity occurs; 2 is indicative of actor's frontal face visibility with lip activity. The structure of the annotation is described in xml format, not following the MPEG-7 ISO prototype.

Afterwards, shot cut information, human face detection, and face tracking information are extracted for all scenes. Shot cut information is extracted using the Shot Boundary module of the DIVA3D software package. The module provides shot boundary detection and shot information management capabilities. The extracted information was subsequently processed by a human annotator that corrected the errors. Human face detection and face tracking information is extracted for each frame using the DIVA3D tracking module. The module allows the user to perform either only automatic human face detection, or to combine the face detection process with face tracking. The face of each actor participating in a dialogue or monologue is assigned a bounding box in each frame of the scene. Face tracking results were edited when needed by human annotators using the Anthropos7 Editor. The extracted data are saved in an xml MPEG-7 compliant manner.

Finally, the two xml files (audio, video) are merged into one xml file for each scene following the MPEG-7 format. The annotations for the two modalities are synchronized since they make use of the same timeline, thus providing joint audio-visual annotation information. Furthermore, the annotation data include the captions for the dialogues and monologues in the scene. It should be noted for the time-being dialogue annotation and captions do not exist for the films The Prestige, and American Beauty.

3.2 Saliency annotation

Saliency annotation is being produced based on manual detection of an audio or visual event that "pops-out", i.e. which has the unique condition or quality of standing out relative to its environment. Attention in audio signals is focused on abrupt changes, transitions and abnormalities in the stream of audio events, like speech, music, environmental noises in real life or sound effects in movies. The salient features that attract more attention can be detected more clearly. The same observations are valid in case of video signals, where outstanding colors (compared to the background color), abrupt scene changes or movements, or sudden events attract the viewer's attention (Rapantzikos, 2007).

Three movie clips of total duration ~ 27 min have been selected from 3 different movies of different genres ("300", "Cold Mountain" and "Lord of the Rings 1"). The clips have been selected after careful consideration; to represent all possible cases of saliency, i.e. visual, audio and audiovisual saliency, as well as smooth alternations between action/non action parts, and dialogue/non dialogue parts to be included. The audio content includes speech in various conditions; speech in form of dialogue, speech with background sound which can be music, noise, other speech or environmental sound. The music content can be found in various conditions too, music with background noise, speech or effects. The background sounds in the clips include environmental sounds such as animals (dog barking, birds singing), autos, knockings, sword sounds etc. and sound effects. The visual content includes a variety of different elements, i.e. abrupt scene changes, computer made light effects and other editing effects.

All movie clips are annotated by two different annotators. No strict duration for the annotation elements is specified, yet an audio event is a bounded region in time that is characterized by a variation or transitional state to one or more sound-producing sources (Evangelopoulos, 2008). An event considered salient is annotated separately, as a means to assign a separate saliency factor. The saliency factor for an audio sequence depends on the impact the sound makes in different scenes and its importance for the annotator. No semantic or linguistic consideration of the content is taken for speech saliency, which is only based on the intensity and strength. Visual saliency concerns pop-out events (pop-out color and pop-out motion) and how salient they are considered by the annotator. Abrupt changes and sudden events can also be regarded as salient. Silence on the other hand, meaning that no significant sound is occurring in the segment is not annotated at all.

The annotators, having already predefined all the above, agree on definitions of the audio and visual events but since each one of them can have an individual opinion about what is salient, likable, interesting or outstanding for the senses, they are free to decide the saliency factor of each event based on their own likes and dislikes. Consequently, annotations from different annotators show some analogy; however since the annotators have different likes and dislikes there are variations on the saliency factor. Such disparities are notable at the annotation of generic saliency where the annotator marks only the parts that bear a saliency factor.

Anvil has been used for saliency annotation. A rich annotation scheme has been defined in order to get all possible saliency factors. The three main saliency categories of the annotation scheme are visual saliency, audio saliency and generic saliency.

Audio saliency is annotated using only the auditory sense; visual saliency only the visual sense while generic saliency is annotated using both modalities simultaneously.

Audio saliency includes a description of the audio type found in a scene. The categories that have been chosen to best fit all possible kinds of sounds in movies are: voice/dialogue, music, noise, sound effect, environmental sound, machine sound, background sound, unclassified sound and mix sound. The annotator has the

opportunity to choose more than one of the above sound types to describe every event, since in a movie up to 5 sounds or more can be detected simultaneously. Thereafter, a factor of high, mid, low or none is assigned for the saliency. Speech saliency is measured by the intensity and loudness of the voice (and defined as extra strong, strong, normal, or reduced). Audio and speech saliency features are presented in Table 2.

Audio Saliency	
Audio type	Voice/Dialogue, Music, Noise, Environmental sound, Machine sound, Background Sound, Unclassified sound, Mix sound
Saliency Factor	None, Low, Mid, High
Speech Saliency	
Actor Id	(actor's numeric label)
Visibility	Visible, Non visible, Voice-Over visible, Voice-Over non visible
Saliency Factor	None, Reduced, Normal, Strong, Extra Strong

Table 2: Audio and speech saliency features

Visual saliency includes a description of the object's motion in every scene. Changes of cast and pop-out events are annotated too. Pop-out events, as stated before, can either refer to color or motion (compared to their environment). Visual saliency is measured as high, mid, low or none. In Table 3, all visual saliency features are presented in detail.

Visual Saliency	
Motion	Start-Stop, Stop-Start, Impulsive event, Static, Moving, Other
Changes of cast	(binary decision)
Pop-out event	(binary decision)
Saliency Factor	None, Low, Mid, High

Table 3: Visual saliency features

Generic saliency is a low-level description of saliency, where the description features are: audio saliency, visual saliency and audiovisual saliency, i.e. when both modalities contribute equally to saliency. Saliency can be measured as high, mid or low. Generic saliency features can be seen in Table 4.

Generic Saliency	
Saliency Type	Visual, Audio, Audio Visual
Saliency factor	None, Low, Mid, High

Table 4: Generic Saliency Features

The above selected audiovisual features have already been proven useful and promising in ongoing experiments aiming at comparing human vs. automatic annotations as well as in testing human evaluations of video summaries. The performance comparison of the audiovisual saliency event detector against the manual annotation on the selected clips showed good agreement. The output of this procedure was a saliency indicator function $I_{sal}(n)$ where n is the temporal frame index. The salient regions were computed automatically by selecting a threshold on the median filtered audiovisual saliency curve. Median filters of different length frames were used. Especially for the longer median filter, the correct frame classification (as salient or non-salient) was up to 80% (Evangelopoulos, 2008).

4. Conclusions

In this paper, MUSCLE movie database was described. It is a multimodal annotated movie database. The fact that MUSCLE movie database encompasses 4 modalities, namely audio-only, video-only, text-only, and audiovisual makes it an efficient test bed for the audio and video research communities. Well known annotation tools are surveyed including a novel tool, named Anthropos7 Editor. Future work will focus on the assessment of agreement/disagreement between annotators for the concepts of dialogue and saliency.

5. Acknowledgment

This work is supported in part by European Commission 6th Framework Program with grant number FP6-507752 (MUSCLE Network of Excellence Project).

6. References

Benetos, E., Siatras, S., Kotropoulos, C., Nikolaidis, N., Pitas, I. (2008). Movie Analysis with Emphasis to Dialogue and Action Scene Detection", in P. Maragos, A. Potamianos, & P. Gros (Eds.), *Multimodal Processing and Interaction: Audio, Video, Text*. N.Y.: Springer.

Chang, S. –F., Sikora, T., Puri, A. (2001). Overview of the MPEG-7 standard. *IEEE Trans. Circuits and Systems for Video Technology*, 11(6): 688–695.

Garg, S., Martinovski, B., Robinson, S., Stephan, J., Tetreault, J., Traum, D. R. (2004). Evaluation of Transcription and Annotation Tools for a Multi-modal, Multi-party Dialogue Corpus. In *Proc. 4th Int. Conf. Language Resources and Evaluation*, pp. 2163-2166.

IBM: MPEG-7 Annotation Tool. www.alphaworks.ibm.com/ tech/videoannex

Ricoh: MovieTool. www.ricoh.co.jp/src/multimedia/MovieTool/

ZGDV, VIDETO: Video Description Tool. www.rostock. zgdv.de/ZGDV/Abteilungen/zr2/Produkte/videto/

EPFL, COALA: Content-Oriented Audiovisual Library Access – Log Creator. http://coala.epfl.ch/demos/demosFrameset. htm

ELAN: EUDICO Linguistic Annotator. www.let.kun.nl/ sign-lang/echo/ELAN/ELAN_intro.html

ANVIL: The Video Annotation Research Tool. www.anvil-software.de

Rapantzikos, K., Evangelopoulos, G. Maragos, P., Avrithis, Y. (2007). An Audio-visual Saliency Model for Movie Summarization. In *Proc. IEEE Workshop Multimedia Signal Processing*, pp 320-323.

Evangelopoulos, G. Rapantzikos, K., Potamianos, A., Maragos, P., Zlatintsi, A., Avrithis, Y. (2008). Movie Summarization Based on Audio-Visual Saliency Detection, *IEEE Int. Conf. Image Processing*, submitted.

A multimodal data collection of daily activities in a real instrumented apartment

A. Cappelletti, B. Lepri, N. Mana, F. Pianesi, and M. Zancanaro
FBK-irst
via Sommarive, 18 – Povo (Trento), Italy
E-mail: {cappelle,lepri,mana,pianesi,zancana}@fbk.eu

Abstract

This paper presents technical setup and methodology used for the data collection in progress within the NETCARITY project, as of middle of February, 2008. The goal of this work is to collect a large amount of high quality acoustic and visual data, concerning people doing common activities of daily living. The final expected structured and annotated database of activities will be helpful to develop systems that, starting from audio-visual cues, automatically analyze the daily behaviour of humans and recognize different kinds of daily living activities (distinguished in single activities, parallel activities, and single activities with some background noises).

1. Introduction

European society is strongly ageing. In 2005 people aged over 65 was 13% of the population [Czaja and Hiltz, 2005] and such figure is expected to increase. It has been estimated that by 2020 one out of four Europeans will be over 60 years old, and one out of five over 65 [Mikkonen et al, 2002]. Consequently there will be more and more aged people, in need of social, home and long-term care services.

Technology can play a crucial role in enhancing in the elderly people (and in their families and associated caring personnel) the feeling of confidence required for ageing-in-place, by assuring the basic support of everyday activities and health critical situations management.

On this way it is located the NETCARITY project[1], aiming to propose a new integrated paradigm to support independence and engagement in elderly people living alone at home. One of the final objectives of this project is the development of a light technological infrastructure to be integrated in the homes of old people at reduced costs. Such technologies should allow both assurance of basic support for everyday activities and detection of critical health situations.

As the project is targeting real everyday needs in real life contexts, one of its first steps is to carry out a data collection in order to obtain significant examples of activity of daily living (ADL) to be studied and modelled for the project purposes.

ADLs monitoring has become an important goal and a valued technological objective mainly for three reasons. Firstly, ADLs monitoring is important for healthcare [Katz, 1983]. Trained caregivers, clinicians and therapists spend much time measuring and tracking ADLs accomplishment in order to asses the functional status of a person and her/his degree of autonomy, detect problems in performing those activities, and plane care interventions. However, current methods for recognizing and monitoring these activities consist of time and resource consuming manual tasks, relying on paid observers (e.g. a nurse, monitoring periodically an elderly patient) or on self-reporting (e.g. patients having to complete an activity report everyday). Automated aids that can improve the caregiver work practice would be of value.

Secondly, ADLs are common activities that people (not only old one) perform daily. Therefore, these activities become interesting also outside the elder-care field. In fact, a large variety of tasks, such as security monitoring or training, which are currently considered expensive human jobs, become amenable to automated support if the computer can recognize human activities.

Finally, home ADLs recognition is a highly interesting and challenging scientific task, aiming to the number of activities people get involved in, and the different ways they can be performed.

For all these reasons, jointly to the project objectives, we are currently working on a multimodal data collection, aiming to have a structured and annotated database of high-quality, realistic and synchronized acoustic and visual data of common daily activities performed in a home setting. This dataset may be very worth for researchers working on automatic activity recognition. Especially for people who use machine learning techniques and need large data corpora for training multimodal activity recognition algorithms.

At present there are already existing data collections coming from a number of smart homes in Europe and in USA, used to collect data on common daily activities. Among these:

- AwareHome project of the GeorgiaTech in Atlanta. In this project, the Georgia Tech Broadband Institute's Residential Laboratory is used as living laboratory for ubiquitous computing in home life [Kidd et al., 1999];
- Philips' HomeLab in Eindhoven, used as showroom and usability laboratory. Subjects live in the lab for several days while researchers may observe and study people in a naturalistic home environment in order to develop better products [Aarts and Eggen, 2002];

[1] NETCARITY (A Networked multisensor system for elderly people: health care, safety and security in home environment") is an Integrated Project, supported by the European Community under the Sixth Framework Programme (Information Society Technologies, Ambient Assisted Living, IST-2006-045508). For more details see http://www.netcarity.org/

- PlaceLab residential facility, maintained by the House_n research group at the MIT Department of Architecture. It is equipped with hundreds of sensing components and it is used as a multi-disciplinary observational facility for the scientific study of people and their interaction patterns with new technologies and home environments [Intille et al, 2006];

However, Philips' HomeLab and AwareHome project were not collecting multimodal data, useful for devising and testing activity recognition systems. On the contrary, the House_n research group was doing it, but it was not including audio and visual features.

There are other multimodal (audio-visual) corpora recently collected, not on common daily activities but on meetings, with people sat around a table. Among these, the MM4 corpus [McCowan et al., 2004] and the VACE corpus [Chen et al., 2005] include low-level cues of human behavior, such as speech, gesture, posture, and gaze, in order to interpret high level meeting events. Similar purposes have also been pursued by the AMI project, collecting a large multimodal corpus [Rienks et al., 2006].

We also collected two multimodal corpora on meeting scenarios, namely the "Mission Survival Corpus 1" [Pianesi et al., 2006], and the "Mission Survival Corpus 2" [Mana et al., 2007].

The paper is organized as follows: Section 2 presents the technical set-up used on the audio and video acquisition sides. Section 3 describes the architecture of the data acquisition system. Section 4 presents the recording procedure, while in Section 5 the expected final result is illustrated. Finally, Section 6 summarizes the present work, formulates some considerations and draws some future steps.

2. Technical Setup

To allow gathering of multimodal data in a real context we have instrumented two rooms of an apartment (specifically a living-room and a kitchen) with audio and video sensors (see Figure 1).

A third room is used to store computers and capture boards. All computers and webcam are connected via an Ethernet LAN. In addition, a wireless network let communication between control machine and PDA.

2.1 Audio sensors

Three groups of T-shape microphone arrays are installed into each room for a total of 24 audio sensors. An array (see dagger sign on Figure 2) is composed by 4 omni-directional microphones, mounted at 2 meters tall on wall. Microphones are connected to A/D converter that samples audio input with a frequency of 48 kHz and a resolution of 16bits. Converters are connected through optical cable to a 24 channels acquisition board, installed on the capturing workstation that provides also an internal synchronization clock to assure alignment between channels.

Figure 1: Camera view in the living room and kitchen

2.2 Video sensors

The two apartment rooms are covered by a total of three webcameras (see oval sign on Figure 2) with Pan-Tilt-Zoom (PTZ) functionalities[2]. They are mounted on the ceiling and offer a large (40º - 150º) field of view due to lens capability. Each camera has an IP address and dispatches "Motion-Jpeg" images over Ethernet Network at a variable frequency from 10 to 20 frames per second, depending on light conditions. To provide power supplying we use a Power Over Ethernet (POE) switch.

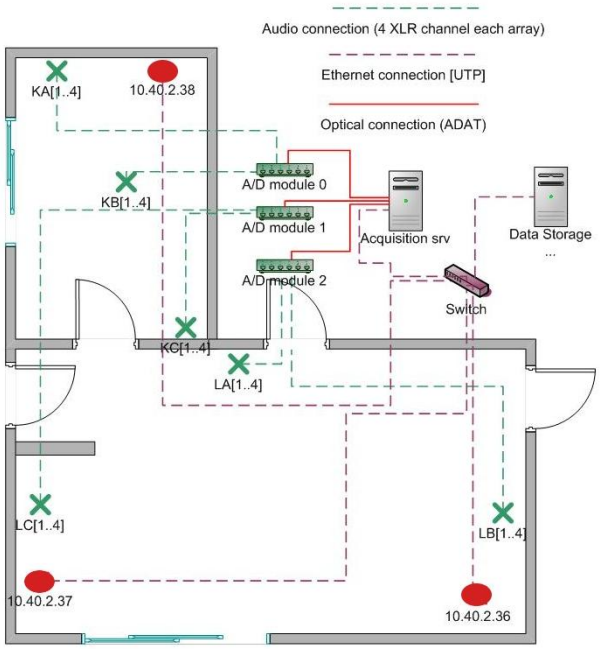

Figure 2: Sensors location in the apartment

[2] These features are not used at the moment.

Image resolution is 640x480 pixels, while jpeg compression level is set to 100 to reduce artifacts on captured images.

3. Data Acquisition System

The architecture of an acquisition system has a significant impact on the quality of collected data, as well as the data nature plays a crucial role into the architectural choice. Data we are collecting are audio and video image streams that are time-based information.

One of the main problems when multimodal data are collected is to guaranty time alignment between streams. That could not be trivial in real environment, especially when capturing systems are distributed, or timers are not precise. Some standards, such as MPEG7 [Martinez 2004], allow to handle multimedia data but we decided to develop our specific protocol to make infrastructure light and easy adaptable to requirements but absolute time remains the core indexer for all the data.

The procedure we are using follows the approach to acquire all streams synchronized referring to a unique time clock with enough resolution. For this reason we developed an ad hoc software application composed by libraries that access to each acquisition hardware. An operator, located in the third room, can initialize and manage all experiment through a GUI by controlling acquisition of streams, sending instructions to user, and making annotations. UI runs on the main machine, which uses a high resolution clock as reference timer[3]. Saving process is a I/O bound and can freeze capturing threads that lead to lack of data. All collected data are dispatched on distributed application over a local network. However, given network resources are limited, it is important to provide enough bandwidth to avoid saturation or data loss.

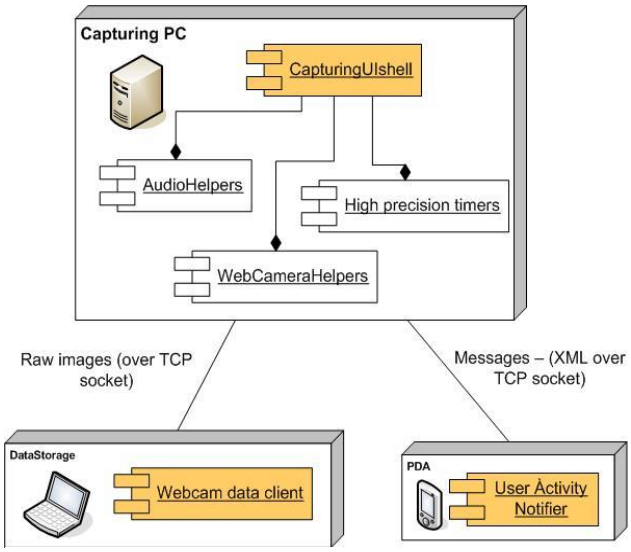

Figure 3: Deployment of acquisition system

Figure 3 shows the deployment of infrastructure. To handle experiments the operator uses a *CapturingUIShell*, (see Figure 4) running on the capturing PC.

From this shell the operator starts and ends experiments (see 1 on the figure). A counter indicates the experiment duration (2). All activities that a subject will perform during a session are listed in a randomized order and then numbered (3). Each one is marked with a different colour (4) according to its specific category (orange for the "single activities", light blue for the "noised activities" and green for the "parallel" ones – see Section 5.1).

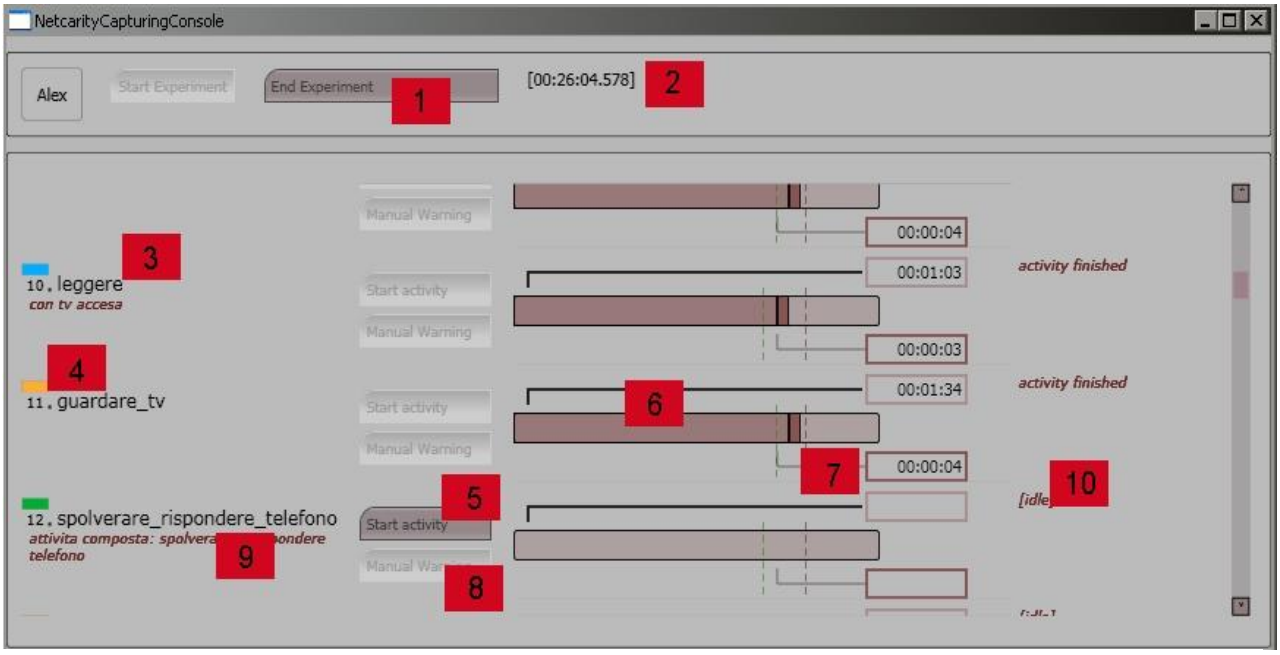

Figure 4: Capturing UI shell

[3] See http://msdn.microsoft.com/msdnmag/issues/04/03/HighResolutionTimer/).

The operator announces to the subject the next activity to be performed by pressing the "Start activity" button (5).
A progress bar keeps track of the effective time of the activity (6) and its closure time (7), i.e. the time that the subject takes to end the action after having received the automatic audio warning message in correspondence of the expected duration (see Table 1 and Table 2). If the subject exceeds the upper bound of closure time, the operator can manually warn the closure again (8). Also other properties as activity description (9) and status (running, finished or idle - 10) are shown on the interface.
GUI links *AudioHelpers* and *WebCameraHelpers* libraries that are responsible of stream acquisition. They fetch data from dedicated hardware (audio-boards) or directly from sensor (webcams). *HighPrecisionTimer* provides a unique time service that all the system adopts to mark events. Given video from webcams does not have a fixed data rate, we annotate the timestamp of each frame. In this way we know when a specific information has been captured. The same idea is applied also for labels and event markers: when something happens, we register both information of time and content as a single item.
On the contrary, given audio data has a predefined sample frequency, we need to annotate just begin/end times.
All streams are then asynchronously saved on a different machine. Raw images are packed with their timestamp and sent via TCP connection to a remote DataStorage node, where clients are responsible to save incoming stream on disks. Later all distributed data are merged in a single repository by the operator.
Subjects are equipped with a PDA that runs an application, connected to the CapturingUIShell, which instructs them on the sequence of activities to do. The operator sends activity instructions to the mobile application. When subject reacts, a response is returned back to manager that records time, activity label, status of user and other surrounding information.
The whole infrastructure runs on Windows XP and it is built by NET framework. This offers capability to handle different acquisition devices, to allow interoperability with low level drivers and future sensor upgrading.

4. Recording Procedure

4.1 Instructions, task description and preparatory steps

Subjects involved in the data collection are mostly people of the administrative staff, but we are going to involve also old people, already collaborating within the project for the user study.
Subjects are firstly instructed on what to do during the session. They are informed that they are audio and video recorded while doing a sequence of common daily home activities, randomly generated from a limited set of activities, repeated more times (see Section 5.1). In doing these activities they are free to move between kitchen and living-room and to choose where and how to perform the task. Given the same activities are asked more times, subjects are invited to possibly perform them each time in a different way (e.g. in case of "reading" activity: once reading a book on the coach and a second time sat at a table or, alternatively, reading a book, rather than a magazine or a newspaper, etc).
Subjects are also informed that during some activities a second person (called "actor') enters in the scene, doing something else (e.g. the actor is answering to a phone call while the subject is watching TV). Subject and actor do not interact. The actor role is to be a sort of "noise generator" (see third action category in Section 5.1).
After having received instructions, subjects are asked to sign a consent form to make the collected audio and video data usable for research purposes.
Before starting a session recording, a visual representation (or model) of each session participant (subject and actor), as well as the recording ambient without any person (background), have to be acquired. These models will be supplied to the person tracker (Lanz, 2006) that will be used later in order to detect and track subject motions during their activities.

4.2 Recording procedure

As seen in Section 3, a software application is used in order to guide subjects in executing a sequence of activities and manage all incoming data: through a GUI the experimenter sends predefined warning audio messages to users via wireless network during experiments. Subjects have a PDA, where the client is running, and wear standard headphones to avoid vocal commands could be captured by microphones. For each activity subjects mark beginning and end of that activity on client in according with received audio warnings. In particular, each activity has a pre-defined duration (60 or 90 minutes). When subjects receive the message announcing the next activity to be done and they are ready to start it, they touch the PDA screen and in correspondence the beginning of the activity is marked. After a certain time (corresponding to the pre-defined duration), subjects receive via headphones a warning message inviting them to finish that activity. Only when they have ended the activity, subjects touch again the PDA screen (and in this way the activity end is marked). Activities' metadata and timestamps are then annotated on server.

5. Final Result: A Structured and Semi-automatic Annotated Database

5.1 Collected Activities

We are collecting some daily activities that people usually perform when at home. We have grouped these activities in three main categories: (1) single activities; (2) parallel activities (i.e. performed concurrently); and (3) single activities performed with some background noises.
The first category includes six basic activities. In particular: (a) phone answering; (b) cleaning-dusting; (c) TV watching; (d) ironing; (e) reading; (f) eating-drinking. A detailed description of these basic activities is depicted in Table 1.

Activity	Description	Location	Subject position	Activity duration (sec)
eating-drinking	eat a snack (chips/biscuits/fruit/ yogurt) and/or drink some water (taking the water from a bottle or a carafe)	kitchen or living room	stood or sat on a chair/armchair /couch	90
reading	read a book/newspaper/magazine	kitchen or living room	stood or sat on a chair/armchair /couch	60
ironing	iron a handkerchief or a napkin on a table or an ironing board	living room	stood	90
TV watching	do zapping or watch a TV program	living room	stood or sat on chair/armchair /couch	90
cleaning-dusting	clean or dust, by using a dust mop and a squirt gun or a feather duster	kitchen or living room	stood	60
phone answering	answer to a call on the mobile phone	kitchen or living room	stood or sat on chair/armchair /couch	90

Table 1: List of single activities

Activity durations are fixed (60 or 90 minutes). As seen in Section 3 and Section 4.2, after that time subjects receive an audio warning message inviting them to end the activity in progress. Given that subjects usually do not suddenly interrupt the activity, the actual duration of the recorded activity is longer than the fixed one (as evident from square 7 in Figure 4). Furthermore, still from Table 1, it is evident that subjects are free to choose where to perform these activities (kitchen and living room) and how (stood, sat, walking around, etc…).

Our choice of these daily activities is mainly motivated by the following reasons: first of all, they are common activities that all people (also elderly one) make during their daily living at home. Secondly, the audio-visual cues extractable from the recorded data are significant features for the recognition of these activities (e.g. on the acoustic side, the sound of the phone ringing is crucial for recognizing the "answering to a phone call" activity; at the same way, on the visual side, the head orientation for the recognition of TV watching activity). However, detection and recognition of some activities may be also quite challenging because some activities are very similar from the viewpoint of the audio-visual features (e.g. how to distinguish "eating a snack" and "watching TV" activities, when in both cases subject is sat on the coach?). Finally, the selected activities allow to use both the available rooms in the apartment (kitchen and living room) and to make data more variable.

The second activity category includes three parallel activities performed concurrently by subjects. In particular, these activities are: a) cleaning-dusting and phone answering; b) ironing and TV watching; c) eating-drinking and TV watching.

The choice of focusing our attention also on "parallel activities" has been guided by the consideration that people often perform activities concurrently in their daily living. However, there are few works in activity recognition field devoted to model and recognize the co-temporal relationships among multiple activities performed by the same subject [Wu et al., 2007]. Therefore, from this point of view, this kind of collected data could be helpful.

Finally, the third activity category includes three different activities performed by subjects with some background noises generated by a second person (as seen in Section 4.2). Specifically: a) reading with a TV watching activity as background noise; b) eating-drinking with a TV watching activity as background noise; and c) TV watching with a phone call as background noise.

This last set of activities may be very useful to test the robustness of multimodal activity recognition systems. A robust activity recognition system should be able to distinguish parallel activities (e.g. a subject is eating while he/she is watching TV) and single activities performed while there are some background noises in the apartment (e.g. a subject is eating while another subject is watching TV).

activity	description	location	subject position	activity duration (sec)
reading (bkg_noise=phone call)	basic activity + phone calling in background (subject is ignoring the call; another person is answering)	kitchen or living room	sat around a table or on a armchair	60
eating-drinking (bkg_noise=TV watching)	basic activity + TV noise in background (note: TV is ignored by subject)	kitchen or living room	sat around a table or on a armchair	90
TV watching (bkg_noise=phone call)	basic activity + phone calling in background (subject is ignoring the call; another person is answering)	living room	sat on chair/armchair	90
cleaning & phone answering	clean or dust and in the same time answer to a phone call	kitchen or living room	stood	60
ironing & TV watching	watch TV while ironing	living room	stood	90
eating & TV watching	watch TV while eating/drinking	living room	stood or sat on chair/armchair	90

Table 2: List of parallel activities and single activities with background noises

A detailed description of the three parallel activities and the three single activities with background noises is depicted in the Table 2.

In addition, we are going to collect also one hundred examples of selected acoustic events (e.g. entry phone ringing, door knocking, cooking alarm) and about fifty examples of complex activities as tea making and coffee making. These activities are performed by the subject following a fixed script: (a) the subject enters in the kitchen; (b) he/she puts some water in the teapot/Italian coffee pot; (c) he/she reads a newspapers or a magazine on the kitchen table while he/she is waiting for the teapot/Italian coffee pot whistle; (d) then he/she turns off the hotplate and puts the water/coffee in a cup; (e) finally, the subject gets out from the kitchen bringing the cup.

These instances of tea/coffee making may be useful as data-set for training and testing learning algorithms able to recognize subjects' intentions and plans [Pollack et al, 2003].

5.2 Expected Outcome

At the end of the data collection we will have a multimodal structured database, having synchronized audio and video streams. As summarized in Table 3, this database will encompass activities performed by 50 subjects. For each subject and activity the database will have four instances/examples.

In short, the database will be set up by about an hour of recorded data for each subject. The total audio-video recordings will be longer than fifty hours.

Subjects	**50**
Examples per subject and activity	**4**
Recorded data per subject	**~ 1 h – 1h 20'**
Total estimated recordings	**> 50 h**

Table 3: Expected collected data

6. Conclusion and Future Work

In this paper we presented technical setup and methodology of the NETCARITY data collection.

The goal of this work is to collect a large amount of high quality data, concerning people doing common activities of daily living. As the data collection is in progress, we cannot provide any detailed descriptions of its content but we can formulate some preliminary considerations on technical issues.

To satisfy evolutions of multimodal feature extraction systems, rate and dimension of audio/video information must be at maximum possibilities nowadays hardware can provide. Given this constraint, the first consideration is that, as made evident by the experience we are doing, collecting such raw data requires a lot of resources.

Secondly, the capturing process is I/O bound: that means the bottlenecks are network infrastructure and access to data storage (all open/write/close operations). In particular, we have tested that such acquisition architecture produces 3MB/sec for audio channels and 20.63MB/sec for video streams. The nature of video image structure is enough to drastically reduce capability of I/O Bus. Saving process must store 60 (20 frames x 3 cameras) relative small images (170kByte) per second; this means that operative system must perform 60 "open/write/close" calls each seconds. Such operations are strongly time consuming and can lead quite fast to a lack of data or a freeze of the system. To avoid this bottleneck we are using a client workstation with SATA disks and a 1Gbit network connection.

On the other side, having been able to synchronized audio and video streams will let us to save time in doing any post-processing (otherwise necessary in order to cut and synchronize collected audio and video files). It requires time information must be trusted, in a sufficiently precise and fast way to retrieve.

Finally, the architecture of the acquisition system, where subjects mark directly start and end times of the activities, jointly to structured files (XML) including all information about the order of the performed activities and the corresponding times, let us to have a semi-automatic annotated database where we know which activity is carried on, when it starts and ends, and consequently which are the corresponding audio and video cues.

At the end of data collection, this annotated and structured database will consist of audio and video recordings of 12 activities, repeated 4 times during each recording session, for each subject (50 subjects), for a total length of more than 50 hours.

The next step will be to extract audio and visual cues from the recorded data. Finally, we plan to devise learning algorithms, based on audio-visual features, able to classify single and parallel activities performed by a subject in home setting.

More in general, this database may be helpful for whoever want to develop systems that, starting from audio-visual cues, automatically analyze the daily behavior of the subjects and recognize different kinds of daily living activities (single activities, parallel activities, single activities with some background noises).

7. Acknowledgements

The data collection described in this paper is supported by the European Union within the NETCARITY Project, under contract number IST2005-045508. The authors would like to thank all the subjects that participated in the experiments, as well as all colleagues collaborating in carrying on the data collection.

8. References

Aarts, E.H.L., and Eggen, B. (eds.) (2002) Ambient Intelligence in HomeLab. Eindhoven: Neroc.

Chen, L., Rose, R.T., Parrill, F., Han, X., Tu, J., Huang, Z., Harper, M., Quek, F., McNeill, D.,Tuttle, R., Huang, T (2005): VACE multimodal meeting corpus. Proc. of Multimodal Interaction and Related Machine Learning Algorithms.

Czaja, S.J and Hiltz, S. R (2005).: Digital aids for an aging society. In Communications of the ACM, 48(10).

Katz, S. (1983) Assessing Self-Maintenance: Activities of Daily Living, Mobility, and Instrumental Activities of Daily Living. Journal of American Geriatrics Society. vol. 31, no 12, pp.712-726.

Kidd, C.D., Orr, R., Abowd, G.D., Atkeson, C.G., Essa, I.A., MacIntyre, B., Mynatt, E., Starner, T.E., Newstetter, W. (1999) The Aware Home: A Living Laboratory for Ubiquitous Computing Research. In the Proceedings of the Second International Workshop on Cooperative Buildings - CoBuild'99. Position paper.

Intille, S.S., Larson, K., Munguia Tapia, E., Beaudin, J, Kaushik, P., Nawyn, J., and Rockinson, R. (2006) Using a live-in laboratory for ubiquitous computing research. In K.P. Fishkin, B. Schiele, P. Nixon, and A. Quigley (eds.) Proceedings of PERVASIVE 2006, vol. LNCS 3968,. Berlin Heidelberg: Springer-Verlag, pp. 349-365.

Lanz, O.: Approximate Bayesian Multibody Tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, September 2006 (Vol. 28, No. 9), pp. 1436-1449.

Mana, N., Lepri, B., Chippendale, P., Cappelletti, A., Pianesi, F., Svaizer, P., and Zancanaro, M. (2007) Multimodal Corpus of Multi-Party Meetings for Automatic Social Behavior Analysis and Personality Traits Detection. In Proceeding of Workshop on Tagging, Mining and Retrieval of Human-Related Activity Information, at ICMI07, International Conference on Multimodal Interfaces, Nagoya, Japan.

Martínez, J M: MPEG-7 Overview (version 10) (2004). Coding of moving picture and audio. International Organization for standardization (ISO/IEC JTC1/SC29/WG11 N6828), Palma de Mallorca.

McCowan, D, Gatica-Perez, S. , Bengio, Y., Moore, D,. and Bourlard, H. (2004): Towards Computer Understanding of Human Interactions. In: Ambient Intelligence, E. Aarts, R. Collier, E. van Loenen & B. de Ruyter (eds.), Lecture Notes in Computer Science, Springer-Verlag Heidelberg, pp. 235-251.

Mikkonen M., Väyrynen S., Ikonen V., Heikkilä M.O. (2002): User and Concept Studies in Developing Mobile Communication Services for the Elderly. in Personal and Ubiquitous Computing, 6(2):113-124..

Pianesi, F., Zancanaro, M., Lepri, B., Cappelletti, A. (in press): Multimodal Annotated Corpora of Consensus Decision Making Meetings. To appear in The Journal of Language Resources and Evaluation.

Pollack, M.E., Brown, L., Colbry, D., McCarthy, C.E., Orosz, C., Peintner, B., Ramakrishnan, S., and Tsamardinos, I. (2003) Autominder: An Intelligent Cognitive Orthotic System for People with Memory Impairment. *Robotics and Autonomous Systems*, 44, pp. 273-282.

Rienks, R., Zhang, D., Gatica-Perez, D., Post, W. (2006): Detection and Application of Influence Rankings in Small Group Meetings. In Proceedings of ICMI'06. Banff, CA.

Wu, H, Lian, C, and Hsu, J.Y. (2007). Joint Recognition of Multiple Concurrent Activities using Factorial Conditional Random Fields. In C. Geib and D. Pynadath (eds.) AAAI Workshop on Plan, Activity, and Intent Recognition. Technical Report WS-07-09. The AAAI Press, Menlo Park, California..

Unsupervised Clustering in Multimodal Multiparty Meeting Analysis

Yosuke Matsusaka[*], **Yasuhiro Katagiri**[†], **Masato Ishizaki**[‡], **Mika Enomoto**[§]

[*]National Institute of Advanced Industrial Science and Technology, 1-1-1 Umezono, Tsukuba, Ibaraki, Japan, yosuke.matsusaka@aist.go.jp

[†]Future University Hakodate, 116-2 Kamedanakano, Hakodate, Hokkaido, Japan, katagiri@fun.ac.jp

[‡]The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan, ishizaki@iii.u-tokyo.ac.jp

[§]Tokyo University of Technology 1404 Katakura, Hachioji, Tokyo, Japan menomoto@media.teu.ac.jp

Abstract

Integration of manual annotation and image processing is one of the key issues in developing multimodal corpora. We report, in this paper, an attempt to apply unsupervised clustering techniques to extract a set of meaningful bodily gesture categories for listener responses in multiparty consensus-building discussion meetings. We argue that, by combining statistical and qualitative analysis, these categories provide us with a systematic method to develop a "coding scheme" for multimodal corpora, incorporating both behavioral and functional regularities in nonverbal expressions.

1. Introduction

Nonverbal signals, such as gazing, head nodding, facial expressions and bodily gestures, play significant functions in organizing human interactions. Their significance is even more emphasized in multiparty settings, since many of the interaction organization behaviors, e.g., turn-taking and participation role assignment, are realized by nonverbal means. Several projects have been collecting multimodal corpora (Carletta et al., 2006; Chen et al., 2006) for multiparty dialogues in order to develop techniques for meeting event recognitions from nonverbal as well as verbal signals (e.g. (Stiefelhagen et al., 2002; Ba and Odobez, 2006)).

From the point of view of the development of multimodal corpora, the task of annotating nonverbal signals exchanged in conversations poses both theoretical and practical challenges. In many of the projects, manual annotation and automatic signal processing are both utilized in corpus building. Their usage pattern, however, is either division of labor, different methods for different types of signals (Pianesi et al., 2006), or validation, manual annotation of ideal values for signal processing (Martin et al., 2006).

We have been collecting a corpus of multiparty conversations to develop a comprehensive model of conversational structures in consensus-building discussion meetings. One of the foci of the study is to investigate the role of participant nonverbal signals exchanged in shaping the content of agreement. For that purpose, we have decided to incorporate unsupervised clustering techniques to combine statistical and qualitative analyses in corpus building. We report, in this paper, our methodologies and tools for the process of assisted annotation.

2. Challenges in multimodal meeting analysis

Different from conventional unimodal analyses of spoken dialogues, which have to handle only a limited channels of information, e.g., speech and transcripts, for a pair of sparticipants, multimodal meeting analyses demand a wider variety of channels of information, e.g., gaze direction, nodding, facial expressions, gestures, and so on, for a number of participants. This extension of signal types and participant number creates theoretical and practical challenges.

Firstly, for most of the nonverbal behaviors, we still lack clear and explicit definitions, particularly for functional categorizations. In order to maintain consistency both across annotations and across annotators, we need to prepare a "coding scheme," a set of clear and explicit definitions for the types of behaviors corresponding to annotation labels. Unlike spoken dialogue cases, which already have well developed coding schemes, such as ToBI, which we can rely on when we produce annotations, we have to develop a coding scheme by ourselves. In multimodal behavior analysis, a coding scheme development amounts to a theory development, and multimodal analysis tends to become an experimental process of defining the coding scheme, applying it to the data, and assessing the quality of its outcome, all of which together forms a unit of cycle and possibly leads to a revision of the coding scheme.

Secondly, the amount of annotations we need to handle is significantly larger than the amount for spoken dialogues. The increase is caused, in part, naturally by the increase in the number of information channels in multimodal analyses. But, the number of annotations for each of the channels itself gets larger. For example, when we manually transcribe a speech channel, the number of annotations is on the order of 1 annotation/sec. or 1 word/sec. We don't have to monitor each and every frame of speech data, because speech stream doesn't change too quickly. We can thus speed up the annotation process by skipping. In contrast, when we manually annotate gaze directions, we have to monitor each and every frame without skipping video recordings, because gaze is known to have very high change frequency. Thus, the number of annotations we need to produce will jump up to the number of video frames in the data, e.g., 30 annotations/sec. Consequently, manual annotation is extremely labor intensive and often takes very long time.

Figure 1: Photo of the recording device: AIST-MARC and recording scene

3. Nonverbal signals in consensus-building discussion meetings

Face-to-face conversation is a most effective means for a group of people to obtain an agreement, e.g., a joint action plan for a work group, or a purchase contract in a commercial transaction. In a multiparty conversation between three or more participants, they negotiate by producing, understanding and expressing approval or disapproval toward a number of proposals, until they get to an agreement. Typical exchanges consist of an assertion or proposal produced by one participant, followed by a variety of listener responses, such as assessments, discussions and counterproposals from other participants. These responses often take the form of nonverbal expressions, as well as of explicit linguistic utterances.

Backchannels, noddings and gazing could be counted as listener responses expressing a positive assessment or a support toward a proposal. Lack of backchannels, gaze aversions and talking to other participants, on the other hand, could be counted as expressing a negative assessment or a disapproval. A listener can also indicate readiness and intent on taking a turn by her bodily movements and gazing. The speaker, observing these listener responses, will, in turn, adjust his utterance production accordingly.

Given the experimental nature of nonverbal signal annotations, it is almost mandatory to rely on some form of automation to obtain a comprehensive picture of this intricate interaction process. Several researchers have focused on creating a tool to assist efficient hand annotation (e.g. (Kipp, 2004)). There is also various research which has introduced machine learning and automatic recognition techniques to speed up the multimodal corpus building processes (e.g. (Martin et al., 2006)). We believe, however, that these approaches still do not provide us with a sufficient environment for supporting efficient and reliable multimodal annotation processes. As we discussed above, manual annotation in multimodal analysis is labor intensive even with the use of efficient tools. We cannot avoid the problem of manual annotation even when we apply automatic recognition techniques, because certain amounts of training data have to be prepared by hand in advance, which are required to build an automatic recognizer by using machine learning algorithms. Usually, the required amounts of those training data are significant.

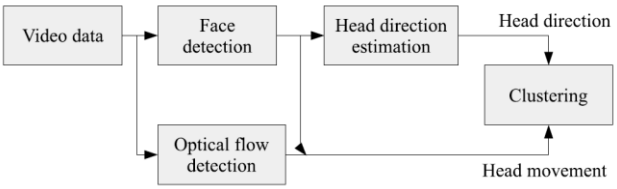

Figure 2: Image processing algorithm

In this paper, we propose two methods to try to solve these problems. We first introduce our image processing tool to enable automatic extraction of parametric features from video. Secondly, we introduce an unsupervised clustering algorithm to enable symbol based analysis of multimodal human behaviors. By using these techniques together we can start our analysis without any pre-defined coding scheme, which, we argue, facilitates easier and quicker connections with higher level qualitative analysis, and an efficient development of a coding scheme.

4. Meeting analysis

4.1. Data

We have been collecting Japanese conversation data on a multiparty design task. A multiparty design task is a type of collaborative problem solving task, in which participants are asked to come up with a design plan. Different from ordinary problem solving tasks, the design goal is only partially specified, and participants need to jointly decide on evaluative criteria for the design goal during the course of the discussion.

Data collection was based on the following settings:

Number of participants: A group of six people.

Conversation setting: Face to face, sitting around a round table.

Task setting: Write up a proposal for cell-phone service features in the near future.

Role setting: No pre-determined roles were imposed. None of the members had professional knowledge about the topic.

Information media: A sheet of paper for each participant to write down the ideas. No computers.

We have used AIST-MARC system (Asano and Ogata, 2006), shown in Figure 1, and 6 supportive cameras, to record the conversation. A sample meeting capture scene is also shown in Figure 1.

Participants of the data collection were recruited from among graduate students who major in information sciences. The data we examine in this paper consist of a 30 minute conversation conducted by 5 males and 1 female.Even though we did not assign any roles, a chairperson and a clerk were spontaneously elected by the participants at the beginning of the session.

4.2. Video based motion extraction algorithm

In order to classify nonverbal responses to be used to build a multimodal corpus, we first apply the image processing algorithm shown in Figure 2 to extract head directions and motions.

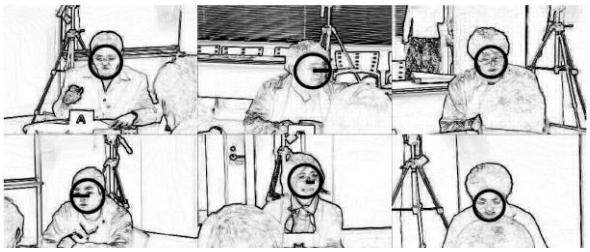

Figure 3: Sample scene with image processing results. The circles represent detected face areas, and the lines in the circles represent head directions.

For head directions, we first detect face positions by using haar features and cascade a classifier-based algorithm (Viola and Jones, 2004). Then, based on the texture of the detected area, we apply a head direction estimation algorithm (Matsusaka, 2005).

Optical flow is calculated from differences between the images of the current and the previous frames, which represents the moving amplitude and direction of each image pixels between the time frames. We use a gradient method to calculate the optical flow from the video stream. The motion of the whole head is defined as the mean amplitude and direction over the detected face area.

Figure 3 shows a sample scene and the results of applying face direction estimation algorithm.

4.3. Unsupervised clustering algorithm

Human behaviors in conversation should exhibit some regular patterns. Those patterns can be observed in the form of clusters in the space of head direction and motion parameters calculated in Section 4.2. By applying unsupervised clustering over the above parameters, we expect to get a clustered patterns of behaviors in conversation organization.

In this paper, we use k-means algorithm for unsupervised clustering. Since we don't have any preliminary knowledge on the number of clusters to be obtained, we investigate the following two methods.

Method 1 classifies the amplitude of the optical flow into 3 classes and, excluding the class with the smallest amplitude, classifies the direction of the face into 3 classes.

Method 2 decomposes the optical flow into horizontal and vertical amplitude, and classifies these data with 2 dimensional features into 4 classes.

4.4. Quantitative analysis

Table 1 shows the central value and the size of each cluster generated by method 1. The central value represents the mean of the (normalized) feature of the data in each cluster. Cluster #1 is classified as having the smallest amplitude in step 1. It corresponds to an almost motionless state since the central value of the cluster is approximately 0. This cluster occupies 80% of the data; that is, participants displayed movement at only 20% of the time during the course of the conversation. Clusters #2∼#4 represent forward, left, and right directions of the face, respectively. The distributions of these face directions do not fluctuate greatly, though the forward direction occupied a relatively higher percentage (8.8% vs. 6.3% and 4.9%).

Table 1: Size and (normalized) central value of clusters generated by method 1

#	Central value	Size	(Percentage)
1st step			
1	0.01	254543	(80.0%)
(2)	0.13	58674	(18.5%)
(3)	0.36	4781	(1.5%)
2nd step			
2	0.53	27862	(8.8%)
3	0.17	19909	(6.3%)
4	0.85	15684	(4.9%)

Table 2: Size and (normalized) central value of clusters generated by method 2

#	Central value (Vertical)	Central value (Horizontal)	Size	(Percentage)
1	0.01	0.01	267304	(84.1%)
2	0.07	0.25	9356	(2.9%)
3	0.29	0.07	13539	(4.3%)
4	0.12	0.04	27799	(8.7%)

The central value and the size of each cluster generated by method 2 is shown in Table 2. The central values of cluster #1 are almost 0 for both the vertical and horizontal directions, and, thus, corresponds to a motionless state. Cluster #2 exhibits mainly horizontal movement. It corresponds presumably to "looking-at-some-participant" behavior. Cluster #3, on the other hand, exhibits strong vertical movements, and presumably represents "looking-up" or "looking-down" behavior. Cluster #4 also exhibits vertical movement but its amplitude is smaller than that of cluster #3. It presumably represents a nodding. The percentages of cluster #2-#4 are about 3%, 4%, and 9%, respectively.

4.5. Qualitative analysis based on unsupervised annotations

Juxtaposition of the unsupervised annotations with speech transcripts gives us a way to get to an in-depth understanding of the interaction organization processes in discussion meetings.

4.5.1. Conversational engagement

Figure 4 shows an excerpt from the data. Each row in the figure represents speech contents and nonverbal behaviors for each of the six participants (A-F) from 170 to 195 sec. Nonverbal behavior annotations were obtained by method 2 in 4.4. The main speaker shifts from D to C in this scene.

We can observe from the figure:

- Speech utterances, including verbal backchannels, are frequent except from the clerk A.
- Main speaker utterances are accompanied by nonverbal behaviors.
- Nonverbal responses given by B, E, and F appear even without accompanying verbal utterances.
- After D yields the main speaker role to C, D doen't produce nonverbal behaviors.

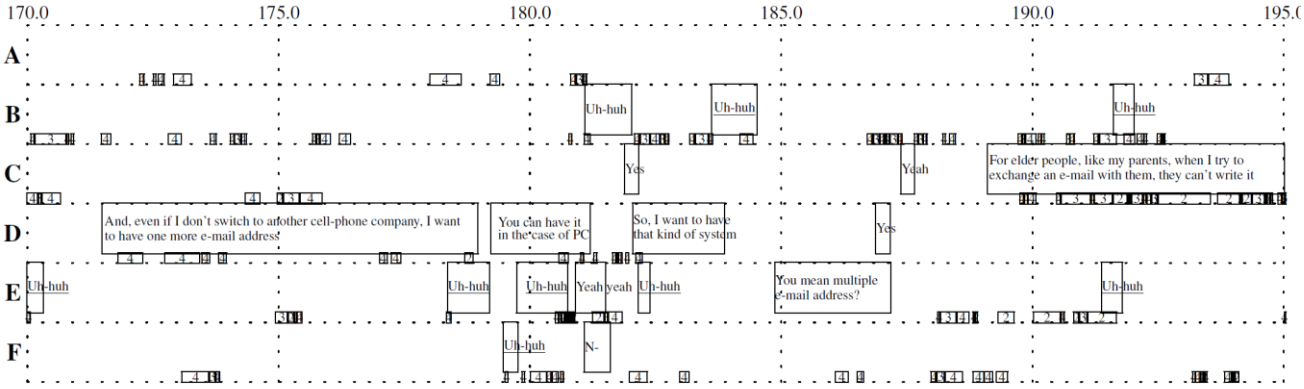

Figure 4: Nonverbal responses and conversational engagement. The underlined utterances indicate backchannels.

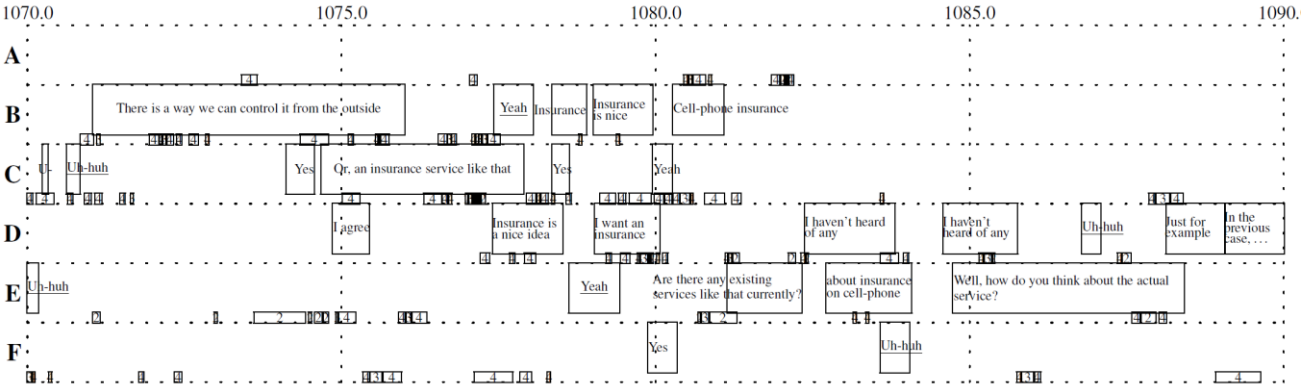

Figure 5: Listener's response after proposal.

These observations are coincident with our expectations that the occurrence of nonverbal behaviors are backed by participant's interest in the topic and high level of conversation engagement. The video inspection revealed that in this scene, the two main speakers D/C and the chairperson E played dominant roles as speaker and addressee, and exhibited frequent nonverbal responses. B and F were active side-participants, and they were producing both verbal and nonverbal responses. It was also observed that once D has released his turn, he looked down on the table and displayed a disinterested attitude toward the conversation.

4.5.2. Listener response to proposals

Figure 5 shows another excerpt, in which a number of listener responses followed a proposal. C produced a proposal on his idea about insurance service in the period between 1074~1078 sec. A lot of nonverbal responses follow from B, D and F, whereas E produces almost no nonverbal responses. The video inspection revealed that B, D and F expressed either strong support or follow up to C's proposal, whereas E didn't get the merit of the idea and directed to C a refinement question: 'Are there any existing services like that currently?'. These observations are also coincident with our expectations that occurrences (or the lack thereof) of nonverbal responses reflect participant's positive (or negative) attitudes.

The examinations of these excerpts demonstrate that unsupervised annotations provide characterizations of nonverbal behaviors which are functionally coherent with our interpretation of conversation organization processes, and suggest that the technique can be used to obtain a starting point in the development of nonverbal behavior coding schemes.

5. Conclusions

We presented in this paper an initial attempt to apply unsupervised clustering techniques in multimodal corpora construction to extract from video images categories of bodily movements which are significant in organizing multiparty interactions. By combining statistical and qualitative analysis, it was possible to obtain a clear picture of the interrelationships between types of movements, gaze shifts and head movements, and interaction organization functions of listener responses, degree of engagement and support.

For future plans, we are preparing to extend our clustering algorithm to handle features in time series, in order to realize more precise analysis of complex phenomena. And also, based on these algorithms, we are preparing to build an interactive GUI to query the corpora instantly.

Acknowledgement

The work reported in this paper was partially supported by Japan Society for the Promotion of Science Grants-in-aid for Scientific Research (B) 18300052 and (A) 18200007.

6. References

Futoshi Asano and Jun Ogata. 2006. Detection and separation of speech events in meeting recordings. In

Proc. Interspeech, pages 2586–2589.
S. Ba and J.-M. Odobez. 2006. A study on visual focus of attention recognition from head pose in a meeting room. In Third Int. Workshop on Machine Learning for Multimodal Interaction, pages 75–87.
Carletta J, Ashby S, Bourban S, et al. 2006. The ami meeting corpus: A pre-announcement. In Machine Learning for Multimodal Interaction, pages 28–39.
Chen L, Rose RT, Qiao Y, et al. 2006. Vace multimodal meeting corpus. In Machine Learning for Multimodal Interaction, pages 40–51.
Michael Kipp. 2004. Gesture Generation by Imitation: From Human Behavior to Computer Character A nimation. Dissertation. com, Boca Raton, FL.
Martin JC, Caridakis G, L.Devillers, et al. 2006. Manual annotation and automatic image processing of multimodal emotional behaviours: Validating the annotation of TV interviews. In Proc. LREC2006, pages 1127–1132.
Yosuke Matsusaka. 2005. Recognition of 3 party conversation using prosody and gaze. In Proc. Interspeech, pages 1205–1208.
Fabio Pianesi, Massimo Zancanaro, and Chiara Leonardi. 2006. Multimodal annotated corpora of consensus decision making meetings. In LREC2006 Workshop on Multimodal Corpora, pages 6–9.
Rainer Stiefelhagen, Jie Yang, and Alex Waibel. 2002. Modeling focus of attention for meeting indexing based on multiple cues. IEEE Transactions on Neural Networks, 13(4):923–938.
Paul Viola and Michael J. Jones. 2004. Robust real-time face detection. International Journal of Computer Vision, 57(2):137– 154.

Multimodal Intercultural Information and Communication Technology – A conceptual framework for designing and evaluating Multimodal Intercultural ICT

Jens Allwood, Elisabeth Ahlsén
SSKKII Center for Cognitive Science
University of Gothenburg
Box 200, SE 405 30 Göteborg, Sweden
E-mail: jens@ling.gu.se, eliza@ling.gu.se

Abstract

The paper presents a conceptual framework for designing and evaluating multimodal, intercultural ICT, especially when it uses embodied artificial communicators as front-ends for databases, digital assistants, tutors in pedagogical programs or players in games etc. Such a framework is of increasing interest, since the use of ICT across cultural boundaries in combination with the use of ICT by persons with low literacy skills is also rapidly increasing. This development presents new challenges for intercultural ICT. A desideratum for interculturally sensitive artificial communicators is a generic, exportable system for interactive communication with a number of parameters that can be set to capture intercultural variation in communication. This means a system for a Generic, Multimodal, Intercultural Communicator (a GMIC). Some factors of importance to take into consideration in the development and evaluation of a GMIC are: activity dependence, generic exportability, multimodal robustness, a flexible repertoire of expressive behaviors, an ability to handle cultural variation concerning content, function, perception, understanding and interpretation as well as concerning cultural differences in interactive features and other kinds of context dependence.

1. Purpose

This paper presents a conceptual framework for designing and evaluating multimodal intercultural ICT (Information and Communication Technology). The following content is included:

- Why interesting?
- Definition of multimodal intercultural ICT (MMIICT)
- Activity dependence of MMIICT
- Generic exportability and multimodal robustness
- Some expressive parameters needed
- Content and function
- Perception, understanding and interpretation
- Interactive features
- Other kinds of context dependence

2. Why interesting?

The use of ICT to support communication and information transfer across national, ethnic, cultural boundaries is becoming more and more common. Intercultural ICT, in this sense, can be present in intercultural use of e-mail, chat, digital news broadcasts, blogs, games, intercultural education and multimodal websites. Especially interesting here is the use of multimodal agents, avatars and robots to communicate and give information across cultural boundaries. The use of such devices as front-ends of databases, in games and chat fora (Life World etc) is quickly increasing.

It is likely that this use will increase even more as people with low literacy skills become users of ICT, since this will be the most natural way for them to communicate. In this situation, it will become more and more interesting to have avatars and other artificial communicators who possess natural (human like) communication skills.

3. Definition of Multimodal Intercultural ICT

By "Multimodal Intercultural ICT", we mean ICT which employs a multimodal GUI (i.e a GUI which uses two or more of the visual, auditive, tactile, olfactory and gustatory sensory modalities and/or two or more of the Peircean modes of representation (index, icon and symbol) (cf. Peirce, 1931). Our focus will be on dynamic, interactive ICT employing avatars or other artificial communicators, across national, ethnic, cultural boundaries, where we characterize an "avatar" as a VR representation of a user and an "artificial communicator" as any communicative agent with a multimodal or multirepresentational front-end (cf. above). An avatar will in this way be a special case of an "artificial communicator".

4. Activity dependence of ICT

Both in design and evaluation, it is important to relate ICT to the activity it is supposed to be part of. Thus, there are different requirements for an "artificial communicator" that has been constructed as a front-end to a database (e.g. for a multinational company to present its products), a personal digital assistant, a friendly tutor teaching small children to read and write or an avatar which is to represent a player in a game like War Craft.

Everywhere the activity, its purpose, its typical roles, its typical instruments, aids, procedures and environment determine what are useful characteristics of an "artificial communicator" and in general of the ICT employed. Both in designing a specification and in designing an evaluation schema, it is therefore important to build in systematic ways of taking activity dependence into account. (Cf. Allwood, 2001).

5. Generic exportability and multimodal robustness

A second desideratum for interculturally sensitive artificial communicators is to base them on a generic system for interactive communication with a number of parameters that can be set to capture intercultural variation in communication. For an interesting suggestion in this direction, cf. Kenny et al., (2007) and Jan et al., (2007).

A Generic Multimodal Intercultural Communicator GMIC): A GMIC would mean that one generic system in principle could be used to allow similar contents or functions to be localized in culturally sensitive, but slightly different ways. It is necessary here to say similar, since the contents (e.g. news casts) or functions (e.g. giving advice) could themselves be affected by cultural variation (Allwood, 1999). Below, we will provide a suggestion for some of the parameters that could characterize such a system.

A third desideratum for the system is multimodal robustness in the sense that the system should be able to handle difficulties in text understanding, difficulties in speech recognition and difficulties in picture/gesture recognition in a sensible way. The system should not halt or respond by "unknown input" or "syntax error" each time routines for recognition or understanding break down. The GMIC should provide routines for how, given a particular activity, such problems can be handled, e.g. by being able to record user contributions, even if they are not recognized or understood and then playing them back to the user as a repetition with question intonation, or by giving minimal feedback through head movements or minimal vocal contributions (which has the function of encouraging the user to continue).

6. Some intercultural parameters of a GMIC

6.1 Cultural variation in expressive behavior

Some expressive behavior exhibits large scale cultural variation (cf. Lustig and Koester, 2006). A GMIC needs to have parameters for

- head movements (nods, shakes, backward jerks, left turn, right turn, forward movement, backward movement)
- facial gestures (smiles, frowns, wrinkles)

- eye movements
- eye brow movements
- posture shifts
- arm and head movements
- shoulder movements
- intonation in speech
- intensity, pitch and duration in speech

In all of these parameters (cf. Allwood et al., 2006) there are several (stereotypical) cultural differences, e.g. head movements for "yes" vary between typical European-style nods and the Indian sideways wagging. Similarly, head movements for "no" vary between head shakes and the backward jerk with an eye-brow raise (sometimes called "head toss"), which is common from the Balkans through the Middle East to India (Morris, 1977, Allwood, 2002).

6.2 Cultural variation in content and function

National, ethnic cultures vary in what expressions, content and functions are seen as allowable and appropriate in different contexts. Should we always smile to strangers? Should women smile to men? Should voices always be subdued and modulated? How permissible are white lies? What is worse, a lying system or an insulting system?

Below are some content areas, where studies have shown cultural variation (cf. Lustig and Koester, 2006).

- Emotions. What emotions are acceptable and appropriate in different activities? E.g. is it permissible for two colleagues at work to quarrel and show aggression or is this something that should be avoided at all costs?
- Attitudes. What attitudes, e.g. regarding politeness and respect, are appropriate? Should titles and formal pronouns, rather than first names and informal pronouns be used?
- Everyday topics. What topics are regarded as neutral and possible to address, even for strangers, e.g. politics, the weather, job, income etc.?
- Common speech acts. Greetings and farewells. Are greetings and farewells always in place or should they be reserved only for some occasions?

6.3 Intercultural variation in perception, understanding and interpretation

Cultural variation in perception, understanding and interpretation is often connected with variation in expression and function. If a male person A does not know that males of group B think that in a normal conversation it is appropriate to stand 10 cm apart, and sometimes touch, their male interlocutors, he might misinterpret what a member of group B does when he steps closer and now and then touches him (A). For an interesting computational model of proximity in conversation, cf. Jan et al. (2007). In general, all differences in occurring expressive behavior are sensitive to expectations concerning appropriate contents and functions and can therefore be misinterpreted. Since many of the expectations are emotional habits on a low level of awareness and control, they might in many cases, more or less automatically, affect perception and understanding (cf. Hofstede, 1997). Thus, a GMIC also needs to have a set of parameters for expectations (e.g. values) and other factors that influence perception, understanding and interpretation.

6.4 Interactive features

Besides parameters for expressive behavior, content, function, and interpretation, other parameters need to be set up to cover variation in interactive features between people with differing cultural backgrounds. Such parameters concern

- Turntaking: How do we signal that speaker change is about to occur? Is it OK to interrupt other speakers? When should interruptions occur? How long should the transition time be from one speaker to the next speaker? Is it OK to do nothing or be silent for a while in a conversation? What should we do to keep a turn? How do we signal that we don't want the turn, but rather want the other speaker to continue? (Sacks, Schegloff and Jefferson, 1974; Allwood 1999).
- Feedback: How do speakers indicate, display and signal to each other that they can/cannot perceive, understand or accept what their interlocutor is communicating (cf. Allwood 2002). Is this done primarily be auditory means (small words like *mhm*, *m*, *yeah* and *no*) or by visual means (head nods, head shakes, posture shifts etc.)? What emotions and attitudes are primarily used? Is very positive feedback preferred or is there a preference for more cautious feedback? (See Kopp, Allwood, Ahlsén and Stocksmeier, 2008.)
- Sequencing: What opening, continuing and closing communication sequences are preferred in the culture, e.g. What is the preferred way of answering telephone calls in different activities (opening sequence)? What is the preferred way of ending telephone calls (closing sequence)? When and how should you greet friends and unknown persons when you meet them (opening sequence)? (See also Allwood et al., 2006.)

6.5 Other kinds of context dependence

Combined with social activity, there are other contextual features which influence communication, such features might, for example, be connected with the deictic features of a language (in English, e.g. words like *I, you, here, now* or tense endings), which in many languages (but not all) are dependent on features of the immediate speech situation. Other factors that might be influential are beliefs, expectations and values that apply to several activities, e.g. ways of showing or not showing respect for another gender, older people or powerful people.

6.6 Types of data needed

To get relevant data for the different parameters, a combination of methods is needed. For data on expressive behavior and interactive features, we need corpora of transcribed and annotated video recordings making possible cross-cultural comparisons and studies of direct data on intercultural communication. For data on differences in perception, understanding and interpretation, it would be desirable to combine corpus data with data from experiments or self-confrontation interviews. Finally, data on content, function and context can be obtained by combining corpus studies of recorded material with interviews and participant observation.

7. Concluding remarks

In this paper, we have given a first outline of a framework which attempts to highlight some of the parameters to be taken into account in design, by providing a sort of guidelines for what could be included in a system for multimodal intercultural ICT and in evaluation by providing a check-list of what should be included in such a system.

8. References

Allwood, J. (1999). Are There Swedish Patterns of Communication? In H. Tamura (ed.) Cultural *Acceptance of CSCW in Japan & Nordic Countries*. Kyoto Institute of Technology, pp. 90-120.

Allwood, J. (2001). Capturing Differences between Social Activities in Spoken Langauge. In I. Kenesei, & R.M. Harnish (Eds.) *Perspectives in Semantics, Pragmatics and Discourse*. Amsterdam: John Benjamins, pp. 301-319.

Allwood, J. (2002). Bodily Communication – Dimensions of Expression and Content. In B. Granström, D. House & I. Karlsson (Eds.) *Multimodality in Language and Speech Systems*. Dordrecht: Kluwer Academic Publishers, pp. 7-26.

Allwood, J., Cerrato, L., Jokinen, K., Paggio, P. & Navaretta, C. (2006). The MUMIN Annotation Scheme for Feedback, Turn Management and Sequencing. IN Proceedings from the Second Nordic conference on Multimodal Communication. Gothemburg Papers in Theoretical Linguistics, 92. University of Gothenburg, Department of Linguistics.

Hofstede, G.1997. *Cultures and Organizationa: Software of the Mind*. New York: McGraw-Hill.

Jan, D., Herrera, D., Martinovski, B., Novick, D. & Traum, D. (2007). A computational Model of Culture-specific Conversational Behavior. In *Proceedings of Intelligent Virtual Agents Conference*, pp. 45-56.

Jan, D. & Traum, D. (2007). Dynamic Movement and Positioning of Embodied Agents in Multiparty Conversations. In *Preccedings of ACL 2007 Workshop on Embodied Language Processing*, pp. 59-66.

Kenny, P., Harholt, A., Gratsch, J, Swartout, W., Traum, D, Marsella, S. & Piepol, D. (2007). Building Interactive Virtual Humans for Training Environments. In *Proceedings of I/ITSEC*.

Kopp, S., Allwood, J., Ahlsén, E. & Stocksmeier, T. (2008). Modeling Embodied Feedback with Virtual Humans. In. I. Wachsmuth & G. Knoblich (Eds.) Modeling Communication with Robots and Virtual Humans. LNAI 4930. Berlin: Springer, pp. 18-37.

Lustig, M. & Koester, J. (2006*). Intercultural Competence: Interpersonal Communication across Cultures*. New York: Longman.

Morris, D. (1977). *Manwatching*. London: Jonathan Cape.

Peirce, C. S. (1931*). Collected Papers of Charles Sanders Peirce, 1931-1958*, 8 vols. Edited by C. Hartshorne, P. Weiss and A. Burks. Camrbidge, MA: Harvard University Press.

Sacks, H, Schegloff, E.A. & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50, 696-735.

Multitrack annotation of child language and gestures

Jean-Marc Colletta, Aurélie Venouil, Ramona Kunene, Virginie Kaufmann and Jean-Pascal Simon

Lidilem – IUFM and Université Stendhal

BP25 – 38040 Grenoble Cedex 9, France

E-mail: jean-marc.colletta@u-grenoble3.fr, a.venouil@free.fr, kuneneramona@yahoo.com, virginie.kaufmann@gmail.com, jean-pascal.simon@grenoble.iufm.fr

Abstract

This paper presents the method and tools applied to the annotation of a corpus of children's oral and multimodal discourse. The multimodal reality of speech has been long established and is now studied extensively. Linguists and psycholinguists who focus on language acquisition also begin to study child language with a multimodal perspective. In both cases, the annotation of multimodal corpora remains a crucial issue as the preparation of the analysis tools has to be in line with the objectives and goals of the research. In this paper we present an annotation scheme aimed at studying linguistic and gesture development between childhood and adulthood, with emphasis to the relationship between speech and gesture and the way it develops. We also present a validation method for gesture annotation.

1. Introduction

This paper deals with an interlinguistic and intercultural perspective of child's speech development in its multimodal and semiotic aspects. It is grounded on the multimodal reality of speech as established by gesture researchers, as well as on the evidence of the complementary semiotic nature of speech signs and gesture signs. Research on gesture as well as cognitive science has shown data which reveal that the listener, or speaker, integrates auditory and visual information from linguistic, prosodic and gesture sources into a single message (McNeill, 1992, 2005; Beattie, 2003; Goldin-Meadow, 2006).

In relation to child language development, several researchers have revealed evidence that a gesture-speech system begins to operate from 16–18 months of age (Capirci, Iverson, Pizzuto & Volterra, 1996; Butcher & Goldin-Meadow, 2000; Volterra, Caselli, Capirci & Pizzuto, 2005; Ozcaliskan & Goldin-Meadow, 2006). Furthermore, there is additional evidence that coverbal gesture - hand or head gestures as well as facial expressions linked to speech - develop as well as vary as the child grows older (Colletta, 2004; Colletta & Pellenq, 2007; Sekine, 2007). However, how does this speech-gesture system develop in children older than 5 years? Does the relationship between gesture and speech become modified under the influence of new linguistic acquisitions and new communicative behaviour? Do new coverbal gestures appear through late speech development? When and how does culture influence this co-development of gesture and speech?

Four research teams from France, Italy and the United States, involving linguists and psychologists and previous experience in multimodal and discourse development, joined forces in order to tackle these questions (French ANR Multimodality Project NT05-1_42974, 2005-2008). Our aim for this workshop is to present the methodological procedures and annotation scheme used in our study in the collaboration as stated above .

Currently, several researchers are interested in the multimodal complexity processes of oral communication. This issue has brought about increased interest to researchers aiming to transcribe and annotate different kinds of multimodal corpora, for instance, researchers in computer sciences take into account the multimodal clues in order to improve the Embodied Conversational Agents (cf. Ech Chafai, Pelachaud & Pelé, 2006; Kipp, Neff & Albrecht, 2006; Kopp et al., 2006; Vilhjalmsson et al., 2007). Other researchers, as Abrilian (2005), work on the annotation of emotional corpora in order to examine the relationship between multimodal behaviour and natural emotions. Other researchers working in the field of autism (*inter alia* Grynszpan, Martin & Oudin, 2003) or language development (Colletta, 2004) also take into consideration these multimodal clues in their studies. It is without doubt that these methods and tools of annotation have paved the way for more exploratory means to study multimodal corpora in detail.

Our data collection is based on a protocol aimed at collecting spoken narratives from American, French, Italian and Zulu children and adults under controlled

experimental conditions. A 2 minute extract of an animated "Tom & Jerry" cartoon is shown to each subject. He/she is then asked to recount the story he/she has just seen to the experimentator. Each interaction is filmed with a camcorder. From each language group, 60 spoken narratives (performed by 20 children aged 5 years, 20 children aged 10 years, and 20 adults) were collected.

The collected data are analysed using the software ELAN (EUDICO Linguistic Annotator)[1]. Two main levels of transcription are selected for annotation: a verbal level and a gesture level (see table 1: annexures). We will briefly present the first level and we will elaborate on the second level as well as on the validation process.

2. Annotation of the verbal level

The main aim of our work is the narrative abilities and the way they develop in children. As children grow older, their linguistic performance in narratives changes; as they include longer, more complex sentences as well as changes in the use of tense, determiners and connectors (Fayol, 2000; Hickmann, 2003; Jisa, 2004). The pragmatic and discourse performance also changes, as the children include changes on the processing of ground information: background *versus* foreground, more freedom in the processing of the event frame (Fayol, 1997), and various speech acts such as narrating, explaining, commenting on the narrative or on the narration (McNeill, 1992 ; Colletta, 2004). The verbal level of our annotation scheme thus includes not only an orthographical transcription, but also a syntactic analysis and a discourse analysis (see figure 1: annexures).

2.1 Speech transcription and syntactic analysis

The transcription of the speakers' words appears on two tracks: one track for the experimentator and one for the child or the adult. The transcription is orthographical and presents the entirety of the remarks of the speakers.

In order to study age related changes in the subject's narrative performance, we first segment the speech into speech turns. To annotate and cut down the speech turns is important to see from what age the child is able to achieve a monologic narrative task in one-go, without assistance from the adult (on the development of monologic discourse, see Jisa & Kern, 1998; Hickmann, 2003; Colletta, 2004; Jisa, 2004). We then segment the speech into clauses and words. The number of clauses or the number of words contained in an account provides a good indication of its informational quantity, which is likely to grow with age. We also classify the clauses of the corpus in order to see whether there is or there isn't a change towards complex syntax in the course of age development. We annotate the words to identify clues of subordination such as conjunctions, relative pronouns or prepositions. Our coding scheme relies on Berman & Slobin's work, (1994), and on Diessel's analysis of the children's syntactic units, in Diessel, (2004). The annotation of words also serves to identify connectives and anaphoric expressions (pronouns, nouns, determiners, etc.) which play an important role in discourse cohesion (de Weck, 1991; Hickmann, 2003).

2.2 Discourse analysis

Before the annotation grid was completed, the extract of the Tom & Jerry cartoon was segmented into macro and micro-episodes. During the annotation process, each clause with narrative content is categorised as processing one of these macro and micro-episodes in order to have an estimate of the degree of accuracy of the retelling of the story by each subject as well as to study his/ her processing of the event frame (Fayol, 1997). Each clause is also categorised as expressing the part or whole of a speech act (narrating, explaining, interpreting or commenting) and as expressing foreground *versus* background of the story. It is a question of studying how age and culture affect pragmatic and discourse dimensions of the narrative activity, as also seen in Hickmann (2003).

The mean duration for the annotation of the verbal level, which includes the transcription of the words of the speakers, syntactic analysis, discourse analysis and validation of all annotations, is 6 hours per file.

3. Annotation of the gesture level

In general, the annotation schemes developed by researchers in computer sciences mainly focus on the description of corporal movements and the form of gestures. It is a question of capturing, as finely as possible, the corporal movements, or to allow for an automatic synthesis. (Kipp, Neff & Albrecht, 2006; Le Chenadec, Maffiolo, Château & Colletta, 2006; Kipp, Neff, Kipp & Albrecht, 2007; Le Chenadec., Maffiolo & Chateau, 2007). Our objective is very different as the annotation has to allow us to study the relationship between gesture and speech. As a consequence, only the corporal movements maintaining a relation to speech - coverbal gesture - interest us. This relationship as well as the function filled by the gesture has a lot of significance for us.

The gesture annotation is carried out in parallel by two independent coders 1 and 2, who annotate on five stages (see figure 2: annexures). Why five stages? In our developmental perspective, the five following parameters

[1] Available from http://www.mpi.nl/tools/ . Also see Brugman and Russel (2004).

prove to be interesting; To begin with; the number of coverbal gestures, which as one would expect, increases with age as we see longer, more detailed and more complex narratives (cf. Colletta, 2004). Another key parameter is the function of gesture. If the hypothesis of a gesture-word system is valid, then we ought to observe age related changes in gesture, with more gestures of the abstract and gestures marking discourse cohesion in the older children's and the adults' performance. The third important parameter is the gesture-speech relationship, which should evolve in parallel with linguistic acquisition and provide evidence of the evolution of language performance towards a more elaborated pragmatic and discursive use (McNeill, 1992; Colletta and Pellenq, 2007). The fourth parameter which is likely to vary with the age of the subjects is the manner which gestures and speech occur on the temporal level (synchrony and anticipation). The fith parameter is gesture form, which in addition to representational accuracy (for representational gestures) in the older children and the adults, should gain more precision in use. (see our three criteria below).

Other than the developmental perspective, every one of these five parameters is likely to vary with the language and culture of the subjects. The study of the interactions between age on one side, and language and culture on the other side, should lead us to a better understanding of the role played by linguistic, cognitive and social factors in multimodal language acquisition.

3.1 Identification of the gestures

In Kendon's work (1972, 1980, 2004), a pointing gesture, a representational gesture or any other hand gesture (an excursion of the body during speech) is called a "gesture phrase" and it possesses several phases including the "preparatory stage, the stroke, i.e., the meaningful part of the gesture phrase, the retraction or return and the repositioning for a new gesture phrase". Yet, some gestures are nothing else but strokes: a head gesture or a facial expression, for instance, are meaningful right from the start till the end of the movement and have no preparatory nor any retraction phases. As a consequence, our premise is that the "gesture stroke" is any coverbal gesture phrase or isolated gesture stroke that needs to be annotated.

To identify the gesture, each coder takes into account the three following criteria (based on Adam Kendon's proposals in Kendon, 2006):

(i) If the movement is easy to perceive, of good amplitude or marked well by its speed,

(ii) If location is in frontal space of locutor, for the interlocutor.

(iii) If there is a precise hand shape or a well marked trajectory.

Once a gesture has been identified, the coder annotates its phases using the following values (based on Kendon, 2004):

<Stroke> = the meaningful height of the excursion of the gesture phrase of a hand gesture, or a movement of the head, shoulders or chest, or a facial display.

<Prep> = the movement which precedes a hand gesture stroke, which takes the hand(s) from its (their) initial position (at place of rest) to where the gesture begins. Contrary to hands, the position of head, the bust or shoulders is fixed. These movements can therefore not be "prepared" as hand movements and consequently can only be annotated as "strokes".

<Hold> = the maintaining of the hand(s) in its (their) position at the end of a hand gesture stroke, before the returning phase or a chained gesture.

<**Chain**> = the movement which brings the hand(s) from its (their) initial position at the end of a hand gesture stroke to the place where a new stroke begins, without returning to a rest position between the two strokes.

<**Return**> = the movement which brings back the hand(s) from its (their) position at the end of a hand gesture stroke to a rest position, identical or not to the preceding one (called "recovery" in Kendon, 2004).

3.2 Attributing function to gesture

The coder then attributes a function to each gesture stroke. In literature about gesture function, there generally appears to be agreement amongst gesture researchers, although they do not always agree on terminology. According to several researchers, Scherer (1984), McNeill (1992), Cosnier (1993), Calbris (1997), Kendon, 2004), 4 main functions are always mentioned:

(i) gestures that help identify (pointing gestures) or represent concrete and abstract referents;

(ii) gestures that express social attitudes, mental states and emotions and that help perform speech acts and comment on own speech as well as other's;

(iii) gestures that mark speech and discourse, including cohesion gesture;

(iv) gestures that help to synchronise own-behaviour with interlocutor's in social interaction.

Our gesture annotation scheme mostly relies on Kendon's classification and covers the whole range of these functions. The coder selects between:

<Deictic> = hand or head gesture pointing to an object present in the communication setting, or to the interlocutor, or to oneself or a part of the body, or indicating the direction in which the referent is found from the actual coordinates of the physical setting. Not all pointing gestures have a deictic function as deictic pointing gesture strictly implies the presence of the referent or its location from the actual physical setting. Thus, gestures which locate a virtual character, object or action (like in sign languages of deaf communities) are to be annotated under <representational>.

<**Representational**> = hand or facial gesture, associated or not to other parts of the body, which represents an object or a property of this object, a place, a trajectory, an action, a character or an attitude (ex: 2 hands drawing the form of the referent; hand or head moving in some direction to represent the trajectory of the referent; 2 hands or body mimicking an action), or which symbolises, by metaphor or metonymy, an abstract idea (ex: hand or head gesture pointing to a spot that locates a virtual character or object; hand or head movement towards the left or the right to symbolise the past or the future; gesture metaphors for abstract concepts).

<Performative> = gesture which allows the gestural realisation of a non assertive speech act (ex: head nod as a "yes" answer, head shake as a "no" answer), or which reinforces or modifies the illocutionary value of a non assertive speech act (ex: vigorous head nod accompanying a "yes" answer).

<Framing> = gesture occurring during assertive speech acts (during the telling of an event, or commenting an aspect of the story, or explaining) and which expresses an emotional or mental state of the speaker (ex: face showing amusement to express the comical side of an event; shrugging or facial expression of doubt to express incertitude of what is being asserted).

<**Discursive**> = gesture which aids in structuring speech and discourse by the accentuation or highlighting of certain linguistic units (ex: beat gesture accompanying a certain word; repeated beats accompanying stressed syllables), or which marks discourse cohesion by linking clauses or discourse units (ex: pointing gesture with an anaphoric function, e.g. pointing to a spot to refer to a character or an object previously referred to and assigned to this spot; brief hand gesture or beat accompanying a connective).

<Interactive> = gesture accompanied by gaze towards the interlocutor to express that the speaker requires or verifies his attention, or shows that he has reached the end of his speech turn or his narrative, or towards the speaker to show his own attention (ex: nodding head while interlocutor speaks).

<Word Searching> = Hand gesture or facial expression which indicates that the speaker is searching for a word or expression (ex: frowning, staring above, tapping fingers while searching for words).

3.3 Definition of the relation of gesture to corresponding speech

The third stage consists in giving a definition of the relation of the gesture to corresponding speech.

<Reinforces> = the information brought by the gesture is identical to the linguistic information it is in relation with (ex: head nod accompanying a yes answer; face expressing ignorance while saying "I don't know"). This annotation does not concern the representational gestures, because we consider that information brought by the representational gesture, due to its imagistic properties, always says more than the linguistic information, as per McNeill (1992) or Kendon (2004). See <Integrates>.

<Complements> = the information provided by the gesture brings a necessary complement to the incomplete linguistic information provided by the verbal message: the gesture disambiguates the message, as in the case of deixis (ex: pointing gesture accompanying a location adverb like « here », « there »; pointing gesture aiming at identifying an object not explicitly named).

<Supplements> = the information brought by the gesture adds a supplementary signification to the linguistic information, like in the case of framing gestures and certain performative gestures (ex: vigorous shaking of head accompanying a no answer; face showing amusement signs to express a comical side of an event; shrugging or showing a mimic of doubt to express incertitude of what has been asserted).

<Integrates> = the information provided by the gesture does not add supplementary information to the verbal message, but makes it more precise, thanks to the imagistic properties of gesture. For instance, drawing a trajectory provides information on the location of the characters or objects we refer to, drawing the shape of an object may at the same time give information on its dimensions.

<Contradicts> = the information provided by the gesture is not only different from the linguistic information in which it is linked but contradicts it, as in the case of certain framing and performative gestures.

<Substitutes> = the information provided by the gesture replaces linguistic information, as in the case of certain performative and interactive gestures (ex: the speaker nods as a yes answer, shakes head as a no answer, shrugs to express his ignorance of the information required).

3.4 Indication of the temporal placement of the

gesture in relation to the corresponding speech

The fourth stage indicates the temporal placement of the gesture stroke in relation to the corresponding speech:

<Synchroneous> = the stroke begins at the same time as the corresponding speech segment, whether it is a syllable, a word or a group of words.

<Anticipates> = the stroke begins before the corresponding speech segment: the speaker starts his gesture while delivering linguistic information prior to the one corresponding to it.

<Follows> = the stroke begins after the corresponding speech segment: the speaker begins his gesture after having finished speaking, or while delivering a linguistic information posterior to the one corresponding to it.

3.5 Gesture form

Kipp, Neff & Albrecht (2006) mention two distinct ways to describe gesture form: "gesture form is captured by either a free-form written account or by gestural categories which describe one prototypical form of the gesture". In our work, as we focus on gesture function and gesture-speech relation, we rely on basic linguistic descriptions of the body movements.

The coder gives a brief linguistic description of each annotated gesture stroke, sticking to its most salient points:

- body part of movement: head, chest, shoulders, 2 hands, left hand, right hand, index, eyebrows, mouth, etc.

- if there is a trajectory: direction of the movement (towards the top, bottom, left, right, front, back, etc.)

- if there is a hand shape: the form of the hand (flat, cutting, closed in a punch-like form, curved, palm up, palm down, fingers pinched, fingers in a circle, etc.)

- the gesture itself: head nod, beat, circular gesture, rapid or not, repeated or not, etc.

4. Validation of the gestures' annotation

In most cases, the validation of the gestural annotation is based on the comparison of the annotations done by two independent coders, and even more rarely, on re-creating gestures by an animated agent (Kipp, Neff and Albrecht, on 2006). These methods are useful to test the validity of an annotation scheme, but they do not allow to check and to stabilise the analysis of a corpus at the end of an annotation procedure. Indeed, in our case, it is not only a question of testing a gestural annotation grid, but it is also a question of validating the annotation of a multimodal corpus (gestures+speech) before using the results of the annotation in statistical analyses.

As a consequence, the last step of the analysis covers two objectives:

- firstly, to finalise the gestural annotation from choices made by both coders and decide in case of disagreement;

- secondly, calculate the interreliability of agreement between all the coders.

The validation phase only applies to the first three parameters (identification of a gesture unit, function and relation to speech), as our goal is to check whether they vary as a function of age and culture. It does not apply to the fith parameter because gesture form is written in free form and therefore the coders can see the same gesture differently, which will be usefull in a more detailed and qualitative analysis. Nor does it apply to the fourth parameter (temporal placement), which will be useful too in such an analysis.

In order to achieve the validation task, a third coder independent of the first two, controls all the annotations. She first adds a supplementary track and annotates "agreement" when she agrees with both coders on the presence of a gesture, or when at least two coders on three agree ont the presence of a gesture, and "disagreement" on the contrary. She then adds two additional tracks to annotate using the same method of "agreement" *versus* "disagreement" for gesture function and gesture relation to speech. She furthermore proceeds to create three new tracks which have the definite annotation of gestures, gesture functions and gesture-speech relation which will help in quantitative analysis.

This last analysis step allows a measure of interreliability amongst the coders and is useful to enhance the process of validation of the annotation. We then calculate:

- Interreliability for the identification of gestures: number of agreement / number of gesture strokes per file.

- Interreliability for the identification of gesture function: number of agreement / number of gesture strokes per file.

- Interreliability for the identification of gesture-speech relation: number of agreement / number of gesture strokes per file.

The mean duration for the annotation of the gesture level, including the validation and final annotation, is 12 hours per file. The duration time varies a lot and is certainly dependant on the subject's communication behaviour, as some gesture far more than others.

5. Final remarks

The project described in this presentation requires the use of a transcription tool and the annotation of both verbal and gesture data. To fulfil our aim, we chose to use the annotation software *ELAN*, a multi-track software with the alignment of transcription of audio and video sources. A multilevel annotation makes it possible to study the gesture-word relations in a very concise manner. It makes it possible to identify, count and describe concrete *versus* abstract representational gestures, marking of connectives, syntactic subordination, the anaphoric recoveries, hesitation phenomena, etc. as well as to study narrative behaviour from a multimodal perspective.

Yet, some technical issues need to be enhanced: the gesture annotation can be more precise if one dissociates the body parts: head, face, hand(s), the whole body. This would avoid the fact that for the same complex gesture involving several body parts, several coders code different aspects of the same behaviour. Moreover, this is painstaking, particularly for adult gestures, where the same gesture can perform two, even three functions simultaneously, which means that the values given in the drop-down menus should, in the future, include this pluri-function feature.

Presently, the analysis in progress will make it possible to appreciate the use of our validation procedure of the gesture annotations… a crucial issue …

6. References

Abrilian, S. (2005). Annotation de corpus d'interviews télévisées pour la modélisation de relation entre comportements multimodaux et émotions naturelles. *6ème colloque des jeunes chercheurs en Sciences Cognitives* (*CJCSC'2005), Bordeaux, France.*

Beattie, G. (2003). *Visible Thought: The New Psychology Of Body Language*. Routledge, London.

Berman, R.A., Slobin, D.I. (1994). *Relating events in narrative : A crosslinguistic developmental study*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Butcher, C., Goldin-Meadow, S. (2000). Gesture and the transition from one- to two-word speech : When hand and mouth come together. In D. McNeill (Ed.), *Language and gesture*. Cambridge, Cambridge University Press, pp. 235--257 .

Brugman, H., Russel, A. (2004). Annotating Multi-media / Multi-modal resources with ELAN. In *4th International Conference on Language Resources and Language Evolution (LREC2004), Lisbon, 26-28 may*.

Calbris, G. (1997). Multicanalité de la communication et multifonctionnalité du geste. In J. Perrot, *Polyphonie pour Yvan Fonagy*, Paris, L'Harmattan.

Calbris, G. (2003). *L'expression gestuelle de la pensée d'un homme politique*. Paris, Editions du CNRS.

Capirci, O., Iverson, J.M., Pizzuto, E., Volterra, V. (1996). Gesture and words during the transition to two-word speech. *Journal of Child Language*, 23, pp. 645--673.

Colletta, J.-M. (2004). *Le développement de la parole chez l'enfant âgé de 6 à 11 ans. Corps, langage et cognition*. Hayen, Mardaga.

Colletta, J.-M., Pellenq, C. (2007). Les coverbaux de l'explication chez l'enfant âgé de 3 à 11 ans. *Actes du 2e Congrès de l'ISGS : Interacting bodies, corps en interaction, Lyon, 15-18 juin 2005, CDRom Proceedings*.

Cosnier, J. (1993). Etude de la mimogestualité. In R. Pléty, *Ethologie des communications humaines : aide-mémoire méthodologique*. Lyon, ARCI et Presses Universitaires de Lyon, pp. 103--115.

Diessel, H. (2004). *The acquisition of complex sentences*. Cambridge, Cambridge University Press.

Ech Chafai, N., Pelachaud, C., Pelé, D. (2006). Analysis of gesture expressivity modulations from cartoons animations. In *LREC 2006 Workshop on "Multimodal Corpora", Genova, Italy, 27 May*.

Fayol, M. (1997). *Des idées au texte. Psychologie cognitive de la production verbale, orale et écrite*. Paris, P.U.F.

Fayol, M. (2000). Comprendre et produire des textes écrits : l'exemple du récit. In M. Kail et M. Fayol, *L'acquisition du langage, T.2 : Le langage en développement. Au-delà de trois ans*. Paris, P.U.F., pp. 183--213.

Goldin-Meadow, S. (2006). Talking and thinking with our hands. *Current Directions in Psychological Science*, 15, pp. 34--39.

Grynszpan, O., Martin, J.C., Oudin, N. (2003). On the annotation of gestures in multimodal autistic behaviour. In *Gesture Workshop 2003*, *Genova, Italy, 15-17 April*.

Hickmann, M. (2003). *Children's discourse : person, space and time across languages*. Cambridge, Cambridge University Press.

Jisa, H. (2004). Growing into academic French. In R. Berman (ed.), *Language Development across Childhood and Adolescence, Trends in Language Acquisition Research, vol.3*. Amsterdam/Philadelphia, John Benjamins Publishing Company, pp. 135--161.

Jisa, H., Kern, S. (1998). Relative clauses in French children's narrative texts. *Journal of Child Language*, 25, pp. 623--652.

Kendon, A. (1972). Some relationships between body motion and speech. In A.W. Siegman et B. Pope (eds.), *Studies in dyadic communication*. Elmsford, NY, Pergamon Press, pp. 177--210.

Kendon, A. (1980). Gesticulation and speech, two aspects of the process of utterance. In M.R. Key (ed.), *The*

relationship of verbal and nonverbal communication. The Hague, Mouton, pp. 207--227.

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

Kendon, A. (2006). Reflections on the Development of Gestural Competence. *Conference, ANR Multimodality Project, Université Stendhal, Grenoble, july 2006*.

Kipp, M., Neff, M., Albrecht, I. (2006). An Annotation Scheme for Conversational Gestures: How to economically capture timing and form. In *Proceedings of the Workshop on Multimodal Corpora (LREC'06)*, pp. 24--27.

Kipp, M., Neff, M., Kipp, K.H., Albrecht, I. (2007). Towards Natural Gesture Synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis. In C. Pelachaud et al. (eds.), *Intelligent Virtual Agents 2007, Lecture Notes in Artificial Intelligence 4722*. Berlin, Springer-Verlag, pp. 15--28.

Kopp, S., Krenn, B., Marsella, S., Marshall, A.N., Pelachaud, C., Pirker, H., Thórisson, K.R., Vilhjálmsson, H. (2007). Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. In J. Gratch et al. (eds.), *Intelligent Virtual Agents 2006, Lecture Notes in Artificial Intelligence 4133*. Berlin, Springer-Verlag, pp. 205--217.

Le Chenadec., G., Maffiolo V. & Chateau N. (2007). Analysis of the multimodal behavior of users in HCI : the expert viewpoint of close relations. *4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, 28-30th June, Brno, Czech Republic.*

Le Chenadec, G.., Maffiolo, V., Chateau, N. & Colletta, J.M. (2006). Creation of a Corpus of Multimodal Spontaneous Expressions of Emotions in Human-Interaction. In *LREC 2006, Genoa, Italy.*

McNeill, D. (1992). *Hand and mind. What gestures reveal about thought*. Chicago, University of Chicago Press.

McNeill, D. (2005). *Gesture and thought.* Chicago, University of Chicago Press.

Özcaliskan, S., Goldin-Meadow, S. (2006). How gesture helps children construct language. In E. V. Clark & B. F. Kelly (Eds.), *Constructions in Acquisition*. Palo Alto, CA, CSLI Publications, pp. 31--58.

Scherer, K.R. (1984). Les fonctions des signes non verbaux dans la conversation. In J. Cosnier et A. Brossard, *La communication non verbale*. Neuchâtel, Delachaux et Niestlé, pp. 71--100.

Sekine, K. (2007). Age-related change of frames of reference in children: Gesture and speech in route description. *Actes du 2e Congrès de l'ISGS : Interacting bodies, corps en interaction, Lyon, 15-18 juin 2005, CDRom Proceedings*.

Vilhjalmsson, H., Cantelmo, N., Cassell, J., Chafai, E.N., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A.N., Pelachaud, C., Ruttkay, Z., Thórisson, K.R., van Welbergen, H., van der Werf, R.J. (2007). The Behavior Markup Language: Recent Developments and Challenges, in C. Pelachaud et al. (eds.), *Intelligent Virtual Agents 2007, Lecture Notes in Artificial Intelligence 4722*, Berlin, Springer-Verlag, pp. 99--111.

Volterra, V., Caselli, M.C., Capirci,O., Pizzuto, E. (2005). Gesture and the emergence and development of language. In M. Tomasello and D. Slobin, (Eds.) *Beyond Nature-Nurture. Essays in Honor of Elizabeth Bates*. Mahwah, N.J., Lawrence Erlbaum Associates, pp. 3--40.

Weck (de), G. (1991). *La cohésion dans les textes d'enfants. Etude du développement des processus anaphoriques*. Neuchâtel, Delachaux et Niestlé.

Annexures

Child		child's speech
Speech turns		*segmentation of child's speech in speech turns*
Clauses		*segmentation of child's speech in clauses*
	Types of clauses	Independent / Main / Name compl. / Verb compl. Sentence compl. / Adjective compl. / Adverb compl. Focalised name compl./ Factitive / Infinitive nominal sentence
	Words	*segmentation of child's speech in words*
	Synt.complex.clues	Preposition Relative pronoun Subordinating conjunction Coordinating conjunction
	Disc.coherence.clues	Name / Verb / Adjective / Determiner Adverb / Connective / Relative pronoun Pronoun / Zéro anaphora
Gest.phase		Prep. Stroke Hold Return Chain
Gest.function		Deictic Representational Performative Framing Discursive Interactive Word searching

Semant.relation		Reinforces Complements Integrates Supplements Contradicts Substitutes
Synchron. relation		Anticipates Synchroneous Follows
Gest.form		*description of the gesture features*
Narrative	child's speech segmented in clauses	
Macro-unit	A B C D E F G	In the nest From nest to bed The hatching "Imprinting" Damage How to calm the baby bird Back to the nest
Micro-unit	A1 A2 A3 …	The mother knits The mother looks at the egg The mother knits …
Pragmatic acts	Narrates Comments Explains	
Narrative level	Foreground Background	

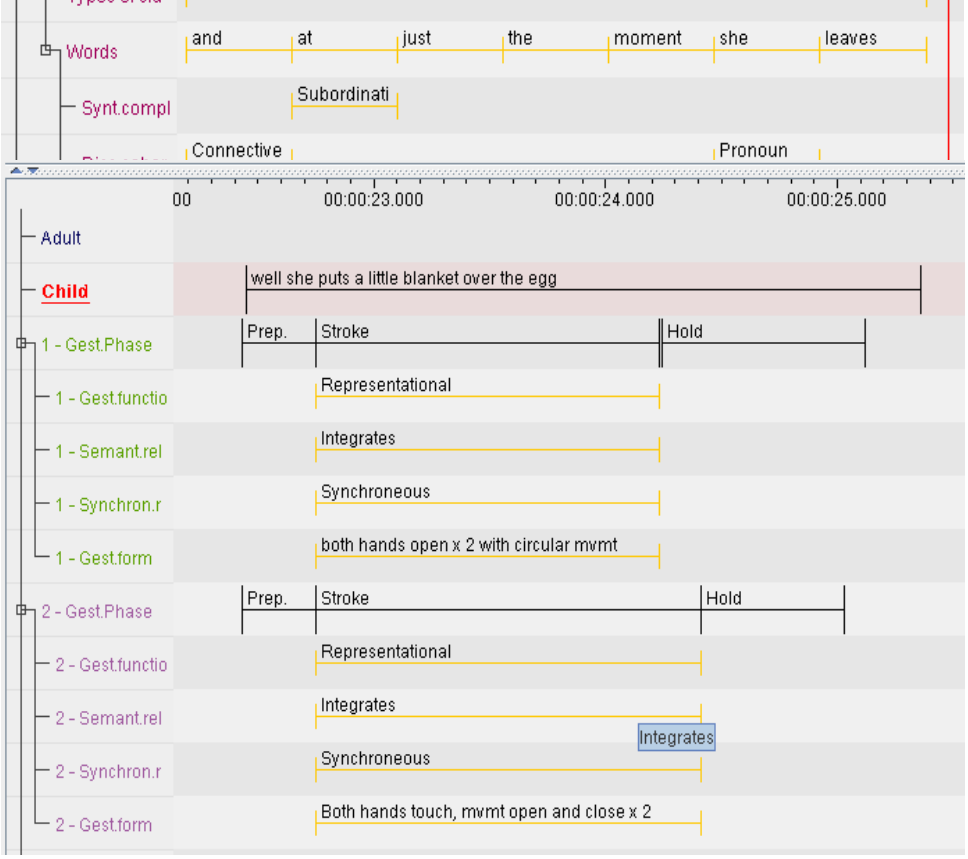

Figure 2: Annotation of the gesture level (extract from ELAN annotation file)

The persuasive import of gesture and gaze

Isabella Poggi[1], Laura Vincze[2]
1 Dpt. di Scienze dell'Educazione – Università Roma Tre
Via del Castro Pretorio 20, 00185 Roma
2 Dpt. di Linguistica – Università di Pisa
Via Santa Maria 85, 56126 Pisa
poggi@uniroma3.it ; l.vincze@ling.unipi.it

Abstract

The paper investigates the use of gesture and gaze in political discourse, and presents an annotation scheme for the analysis of their persuasive import. A model in terms of goals and beliefs is illustrated, according to which persuasion is a case of social influence pursued through communication in which the persuader aims to influence the persuadee to pursue some goal while leaving him free to adopt it or not, and arguing how that goal is in the persuadee's interest. Two studies are reported on electoral debates of three politicians in Italy and France (Achille Occhetto, Romano Prodi and Ségolène Royal), and an annotation scheme is presented through which the gesture and gaze items produced in some fragments of political discourse were analysed as to their signal and their literal and indirect meanings, and classified in terms of the persuasive strategies they pursue, *logos*, *ethos* or *pathos*. The results of the two studies are presented, showing that the pattern of persuasive strategies found in the meanings of gesture and gaze of each politician is coherent with either the persuasive structure of the specific fragment analysed or with the politician's general political strategy.

1. The Rhetorical body

The importance of bodily behaviour in persuasive discourse has been acknowledged back since the ancient Roman treatises of Rhetoric, by Cicero (55 B.C.) and Quintilian (95), as an indispensable part of "Actio" (discourse delivery), in that gesture, gaze and head movements fulfil various communicative functions, often of use in the economy of persuasive discourse. By gestures and other body movements we can summon, promise, exhort, incite, approve, express apology or supplication, display emotions (regret, anger, indignation, adoration), depict or point at objects. Also recent literature (Atkinson, 1984) overviews various aspects of the body's relevance in political communication, like the use of pauses and intonation (Bull, 1986), of facial expression and other bodily behaviours (Frey, 2000; Bucy & Bradley, 2004), and of gesture (Calbris, 2003; Streeck, in press; Poggi & Pelachaud, 2008).

As witnessed by these last works, the persuasive import of body behaviours is due to the meanings they convey. But how can one detect and compute the meanings borne by gestures, facial expression, or pauses? An annotation scheme is required that can take into account all the semantic contents of these signals, while also singling out those that are specifically relevant to the persuasive goals of a discourse.

The importance of co-verbal gesture in conveying information that is effective in persuasion has been shown at length in the studies above; but also facial behaviour and, within it, gaze, is relevant in this connection. Both in fact may be used, with persuasive functions, not only as an accompaniment of speech, but also while one is not holding the turn but is playing the role of the silent interlocutor. For example, in Italian political talk shows, while a politician is talking often the cameras record the facial expressions of his opponents, which are sometimes very communicative and may have a counter-persuasive role.

In this work we present an annotation scheme for the transcription and analysis of gesture and gaze in persuasive political discourse, and we argue for how this scheme allows to compute and analyse the quantity and quality of persuasive gesture and gaze in a discourse, and how this is coherent with the persuasive structure of a politician's discourse and his or her political strategy.

In sections 2. we overview a model of persuasion based on the notions of goal and belief, and a hypothesis on how to assess the persuasive import of gesture and gaze in persuasive multimodal discourse. In 3. and 4. we present an annotation scheme to describe and classify gesture and gaze in persuasive discourse, and through this we analyse gesture and gaze items in some fragments of political discourse, finally showing, in section 5., how this allows to find out different patterns of their persuasive use in different politicians.

2. A model of Persuasion and the structure of a persuasive discourse

According to a model of mind, social interaction and communication in terms of goals and beliefs (Castelfranchi & Parisi, 1980; Conte & Castelfranchi, 1995; Poggi, 2007), persuasion is an act aimed at social influence. Social influence, as defined by Conte & Castelfranchi (1995), is the fact that an Agent A causes an increase or decrease in the likeliness for another Agent B to pursue some goal GA. In order to have B more likely pursue a goal GA, A must raise the value that GA may have for B, and does so through having B believe that pursuing GA is a means for B to achieve some other goal GB that B already has, and considers valuable (Poggi, 2005). In some cases, even having someone feel some emotion is a way to influence him, since emotions are states with a high motivating power – they trigger goals (Miceli et al., 2006). Given this definition of social influence, there are many ways to influence others, ranging from education to threat, promise, manipulation, and the use of strength. But among these, persuasion is an action aimed at social influence that shares some features

with a particular kind of speech act: advice. In fact, *suadeo* in Latin means "I give advice". And like advice (Poggi & Castelfranchi, 1990), persuasion is characterised by the following:

1) A pursues a goal of social influence through communication, that is, not only he tries to induce GA in B, but also makes clear to B he wants to do so
2) A leaves B free of either pursuing or not the goal GA proposed by A, in this differing from threat, for example; and finally,
3) A aims to convince B that GA is in the interest of B. In fact, to persuade B to have GA as a goal of his, A must convince B, that is, induce B to believe with a high degree of certainty, that GA is worth pursuing – it is a goal of high value – since it is a sub-goal to some goal GB that B has.

In order to persuade B, A can make use (Poggi, 2005) of the three strategies already highlighted by Aristotle (360 B.C.): *logos* (in our terms, the logical arguments that support the desirability of GA and the link between GA and GB); *pathos* (the extent to which A, while mentioning the pursuit of goal GA, can induce in B emotions or the goal of feeling or not feeling them); and *ethos* (A's intellectual credibility – his having the skills necessary for goal choice and planning, that we may call "ethos-competence", and his moral reliability – the fact that he does not want to hurt, to cheat, or to act in his own concern – that we call "ethos-benevolence").

In order to persuade others we produce communicative acts by exploiting different modalities – written texts, graphic advertisement, words, intonation, gestures, gaze, facial expression, posture, body movements: we thus make multimodal persuasive discourses, that is, complex communicative plans for achieving communicative goals. Each discourse can be analysed as a hierarchy of goals: a communicative plan in which each single communicative act (either verbal or non verbal) aims at a specific goal. And each goal may also aim at one or more supergoals: further goals for which the first goal is a means. E.g., if I say "Are you going home?" my literal goal is to ask you if you are going home, but through this I may aim at the supergoal of asking for a lift. So, two or more communicative acts may have a common super-goal: saying "I am here with this face" plus saying "this is the face of an honest person" may aim at the supergoal of implying "I am an honest person". A discourse (both a unimodal and a multimodal one) is a sequence of communicative acts that all share a common supergoal. For example, in a pre-election discourse, all the sentences, gestures, face and body movements aim at one and the same common supergoal: "I want you to vote for me". They do so by making up a persuasive multimodal discourse, in which each signal with its direct and indirect meanings, that is, through its literal and intermediate supergoals, pursues a *logos*, *ethos* or *pathos* strategy. Thus, all signals in a persuasive discourse are planned as aiming at the global persuasive message, even if not all of them, of course, are planned at the same level of awareness. While verbal signals are generally planned in a conscious way, gestures, facial expressions, gaze and body posture may be planned and produced at a lower level of awareness. But this does not imply that they do not make part of the global communicative plan, nor that the Speaker does not have a (more or less aware) goal of communicating the meanings they bear. This is witnessed by the fact that, apart from cases of ambivalence or deception, the whole multimodal message is generally coherent with its global meaning (Poggi, 2007), that is "distributed" across modalities.

3. Persuasion in gestures

In a previous work, Poggi and Pelachaud (2008) investigated the impact of gestures in persuasive discourse. An annotation scheme was constructed to assess the persuasive import of gestures, divided into 7 columns. Here we present a clearer and more systematic version of this scheme, only formally different from that one, in that it is divided into 9 columns (see Table 1). In the columns we write, respectively:

1) the number of the gesture under analysis and its time in the video;
2) the speech parallel to the gesture under analysis;
3) a description of the gesture in terms of its parameters (Poggi, 2007): handshape, location, orientation and movement, and for the movement the parameters of expressivity were described (Hartmann et al., 2002): temporal extent, spatial extent, fluidity, power and repetition;
4) the literal meaning of the gesture. A gesture, as any communicative signal, by definition means something, that is, it corresponds to some meaning; this meaning can be codified, as in a lexicon, or created on the spot but in any case comprehensible by others, and then shared; and it may be paraphrased in words. (For examples of the signal-meaning pairs in gestures, see Poggi, 2007). This verbal paraphrase is written in col. 4);
5) a classification of the meaning written down in col.4, according to the semantic taxonomy proposed by Poggi (2007), that distinguishes meanings as providing information on the World (events, their actors and objects, and the time and space relations between them), the Sender's Identity (sex, age, socio-cultural roots, personality), or the Sender's Mind (beliefs, goals and emotions);
6) on the basis of the semantic classification in Column 5), the gesture is classified as to its persuasive function and the persuasive strategy pursued: whether it conveys information bearing on *logos*, *pathos*, *ethos benevolence*, or *ethos competence*;
7) 8) and 9). Columns 7), 8) and 9) contain, for possible indirect meanings of the gesture, the same analysis of cols. 4), 5) and 6).

The gestures analysed by Poggi & Pelachaud (2008) were taken from the political discourses delivered by Achille Occhetto and Romano Prodi, both candidates of the Centre-leftists against Silvio Berlusconi, during the Italian elections in 1994 and 2006. Some fragments were analysed as to their global meaning and their persuasive structure, and the gestures performed during discourse were annotated by two independent coders, previously trained in the annotation of multimodal data.

From this analysis, it resulted that there are no gestures whose meanings are meanings we can utterly define "persuasive"; rather, some gestures, or sometimes simply typically contained in the cognitive structure of persuasive discourse. In fact, what types of information can we call "persuasive", and where do they dwell in the cognitive structure of a persuasive discourse? In other words: how can one fill columns 6 and 9 of Table 1?

According to the model presented, some types of information that are typically conveyed in persuasion, and that make a discourse a persuasive discourse, are those linked to the scenario of persuasion: a goal proposed by a Sender, its being important and good for the Addressee's goals, and the certainty of this mean-end relationship, but also the Addressee's emotions and its trust in the Sender: in other words, the meanings relevant to persuasion are the following:

1. *Importance*. If something is important, to obtain it will be a high value goal that you want to pursue. And gestures that convey the meaning "important" mention the high value of a proposed goal, to convince the Addressee to pursue it. This meaning is typically borne by gestures that convey performatives of incitation or request for attention, or other gestures like Kendon's (2004) *"finger bunch"*, that convey a notion of importance as their very meaning; but expressing "importance" is also the goal of *beats*, since every beat stresses a part of a sentence or discourse, thus communicating "this is the important part of the discourse I want you to pay attention to". Finally, this can also be the goal of irregularity or discontinuity in the gesture movement: an effective way to capture attention.

2. *Certainty*. To persuade you I must convince you, that is, cause you to have beliefs with a high degree of certainty, about what goals to pursue (their value, importance) and how to pursue them (means-end relationship). To induce certainty in you, I may need to show self-confident and certain about what I am saying.

This is why gestures that convey high certainty, like Kendons' (2004) *"ring"*, may be persuasive.

3. *Evaluation*. To express a positive evaluation of some object or event implies that it is a useful means to some goal; thus, to bring about that event or to obtain that object becomes desirable, a goal to be pursued. In the marketplace, to convince someone to buy a food, the grocer's "*cheeck screw"* (rotating the tip of the index finger on cheek to mean "good", "tasty"), would be a good example of persuasive gesture.

4. *Sender's benevolence*. In persuasion not only the evaluation of the means to achieve goals, but also the evaluation of the Persuader is important: the Sender's *ethos*. If I am benevolent to you – I take care of your goals – you can trust me, so if I tell you a goal is worthwhile you should pursue it. A gesture driven by the ethos strategy of showing one's moral reliability is the gesture, quite frequent in political communication, of *putting one's hand on one's breast* to mean "I am noble, I am fair" (Serenari, 2003).

5. *Sender's competence*. Trust implies not only benevolence but also competence. If I am an expert in the field I am talking about, if I am intelligent, efficient, you might join with me and pursue the goals I propose. Here is an example. The Italian Politician Silvio Berlusconi, in talking of quite technical things concerning taxes, uses his *right hand curve open, with palm to left, rotating rightward twice*, meaning that he is passing over these technicalities, possibly difficult for the audience; but at the same time the relaxed appearance of his movement lets you infer that he is smart because he is talking of such difficult things easily, and unconstrained. This provides an image of competence.

6. *Emotion*. Emotions trigger goals. So A can express an emotion to affect B by contagion and thus induce him to pursue or not to pursue some goal. In talking about his country, for example, Romano Prodi, moving his forearm with short and jerky movements of high power and velocity, conveys the pride of being Italian to induce the goal of voting for him.

These are the meanings that, when found in a discourse, give it a persuasive import. Among these types of information, Emotion (n.6) typically makes part of a *pathos* strategy; the Sender's *benevolence* and *competence* (n.5 and 4), but also *certainty* (n. 2), are clearly *ethos* information; while the elements of *importance* and *evaluation* (n. 1 and 3) are generally conveyed through a *logos* strategy. Nonetheless, these categories can merge with each other: for example, expressing an emotion about some possible action or goal may imply it is an important goal for me, and should be so for you. In this case, at a first level there is a *pathos* strategy – the goal of inducing an emotion, but this pathos is aimed at demonstrating the importance of the proposed goal, thus conveying a *logos* strategy at the indirect level.

Let us see how these elements come out from Table 1. At line 1, Prodi quotes an ironic objection to his political action in order to counter-object to it. While saying "*Si è detto recentemente con ironia*" ("recently people ironically said"), *his hands, with palms up a bit oblique, open outward*: an iconic gesture referring to something open, public; a way to open a new topic in your discourse, like when the curtain opens on the stage: a metadiscursive gesture, but with no indirect meaning and no persuasive import. Then (line 2), while saying *"ma guarda Prodi fa il discorso con la CGIL e con la confindustria"* ("Oh look, Prodi is talking to both trade unions and factory owners"), *he puts his left hand on his hip*, and at the same time, with his *chest erected, he shakes his shoulders* (first left shoulder forward and right backward, then the reverse). His *hand on hip* bears the meaning of someone taking the stance of a judge, the *erected chest* shows self-confidence, almost, a self attribution of superiority, and *shoulders shaking* show that he is gloating for the other being judged and ridiculed. This whole movement is a way to mimic those saying the quoted sentence, while making fun of them. Actually, he is somehow meta-ironizing: he is being ironic about others' irony, by ridiculing their attitude of superiority through exaggeration. Irony in fact is often brought about through hyperbole (Attardo *et al.* 2003). This gesture has a persuasive import in that

ridiculing brings about the Addressees' emotion of amusement, thus exploiting a pathos strategy in order to intends to lead the audience to prefer him. Then he says (line 3): *"s**ì** f**a**ccio il disc**o**rso con la cigi**elle** e la confind**u**stria"* ("Yes I am talking to trade unions and factory owners"), again with *left hand on hip*, but with *bust bowing* five times rhythmically, simultaneously with the stressed syllables in the concomitant sentence. The *bust bow*, like an ample nod, means: "I acknowledge that what you say is true", while the *hand on hip* claims self-confidence. But acknowledging that an accusation or a criticism is true while showing confidence means that you accept it as a neutral or even positive statement, devoid of any negative evaluation: thus the combination of the two movements means "I will really do what they accuse me of", conveying a meaning of defiance, hence giving the impression of an even higher self-confidence.

4. Persuasion in Gaze

In another work, Poggi and Vincze (2008) investigated the persuasive use of gaze in political discourse by analysing another pair of political debates: one is again the discourse of Prodi, and the other is a pre-electoral interview with Ségolène Royal before the elections of April 2007 in France. The two fragments were analysed by two independent expert coders.

Also in this study the hypothesis was that the persuasive import of gaze, just as that of words and gestures, depends on the meanings it conveys. Therefore, to assess how persuasive the gaze exhibited in a discourse might be, you have to assess its meanings. For the analysis of gaze in the fragments of Royal's and Prodi's discourse we used an annotation scheme similar to that used for gestures. Table 2 shows the analysis of two gaze items in Royal's discourse.

In example 1, while talking of the top managers who spoil the enterprises, like Mr. Forgeat (Col.2), Royal *looks at the Interviewer*, Arlette Chabot, with a *fixed gaze* (col.3) which means "I am severe, I do not let you avert gaze" (4); an information about Royal's personality, her being serious and determined (5), aimed at a strategy of Ethos competence (6), and possibly to indirectly conveying that she is one who struggles against injustice (7): again information on her personality (8), bearing on the moral side of ethos, benevolence (9). Then Royal, leaning her head on the left, *looks at the Interviewer obliquely and with half-closed eyelids*, an expression of anger and indignation: information about her emotion, which she possibly wants to induce in the audience, thus pursuing a *pathos* strategy.

In 13, by referring to a proposal made by Sarkozy that the unemployed people should be induced to choose a job out of no more than two, and lest they do so, they should lose their unemployment subsidy, Royal is arguing that this choice can only be acceptable if the conditions of the two jobs are not very punitive. So, while saying *il faut accepter cet emploi* ("you have to accept this job"), she *looks down, first on the right then on the left*, as if looking at two things before deciding, thus referring to the choice between the two jobs. This is an iconic use of gaze, elicit a negative evaluation of the ridiculed people. And by inducing a negative evaluation of the opponents, Prodi providing Information on the World, namely an action of choice, by miming it. After that, she *raises her eyebrows while keeping her eyelids in the défault position*: one more iconic gaze that means "order", miming the expression of someone who orders to the unemployed to make his choice. With these two gaze items Royal is playing the roles of both, the unemployed people and the job proposer, thus dramatising the scene of Sarkozy's proposal. On the basis of the following argumentation, in which Royal is very critic about it, we can interpret her dramatisation as a parody, a way to make fun of Sarkozy's proposal, thus conveying a negative evaluation of her opponent through a pathos strategy.

5. Gesture, gaze and political discourse

By computing the gesture and gaze items in the fragments analysed, we singled out the patterns of persuasive strategies in the observed subjects (Tables 3 and 4). For example, as results from Table 3, Occhetto has a higher percentage of persuasive gestures than Prodi out of the total of communicative gestures (Occhetto 20 out of 24, 83%, Prodi 34 out of 49, 69%), but this is also because Prodi sometimes uses iconic gestures, that convey Information on the World and have no persuasive import except for some in expressivity. Moreover, Occhetto relies much more on *pathos* than on *logos* gestures (30% vs. 5%); Prodi uses the two strategies in a more balanced way, but with a preference for *logos* (23% vs. 12%). In both, the majority of gestures (65%) pursue an *ethos* strategy, and both tend to project an image of competence more than one of benevolence, but this preference for competence holds more for Prodi (50% vs. 15%) than for Occhetto (45% vs. 20%).

These differences can be accounted for both by specific aspects of the fragments analysed, and by the different political origins of the two politicians. On the former side, in the fragment under analysis Occhetto is attacking his opponent Berlusconi from an ethical point of view, and therefore he aims to project an ethically valuable image of himself, while Prodi is describing his program and thus he wants to project the image of one who is able to carry it on in an effective way. On the latter side, the different political origins, Prodi is a centre-leftist coming from a former catholic party (the Christian Democrats), while Occhetto is a communist, while Berlusconi still makes appeal to the old prejudice that the Communists "eat the kids"! Hence, Occhetto has a higher need to show his image of benevolence.

Coming to the persuasive use of gaze, from a comparison between Romano Prodi and Ségolène Royal (Table 4), it comes out that Prodi mainly shows the strategies of showing *competence* (62%), then *logos* (25%) and *pathos* (13%), and no *ethos benevolence* at all; in Royal the *logos* strategy is the most frequent (54%), followed by *competence* (27%), but also some *benevolence* (5%), with *pathos* similar to Prodi's (14%). Also in this case, the differences found in the pattern of persuasive gaze can be due to differences in political strategies. Prodi does not aim to show benevolence because he does not need to

enhance his image of a good honest man, especially as opposed to Berlusconi who is often supposed to deal with either on her goal to contrast the stereotype of female as keener to emotion or irrationality, or to a cultural difference between Italian and French orator: a higher disposition of the French to rational argumentation; a French *esprit de géométrie*.

politics mainly for the sake of his financial interests. And the high level of *logos* strategy in Royal might depend

6. Conclusion

We have presented an annotation scheme for detecting the persuasive import of gesture and gaze. By applying it to the analysis of political discourse in different politicians, we have seen that the pattern of persuasive strategies in their gesture and gaze is quite coherent either with the specific context of their discourse or with their political line and political style. The plausibility of this result may confirm the descriptive adequacy of our annotation scheme.

1. Time	2. Speech	3. Gesture description	4. Literal meaning	5. Meaning type	6. Persuasive import	7. Indir. meaning	8. Meaning type	9. Persuasive import
1 0.00.1	*"Si è detto recentemente con ironia"* *Recently people ironically said*	hands palms up oblique open outward Sp.ext: +1 Fluid: +1 Power: -1 Temp.ext: 0 Rep.: 0	Open, public I show, exhibit, show off	ISM Metadiscursive				
2 0.00.6	*"Ma guarda Prodi fa il discorso con la CGIL e con la confindustria"* *Ok look Prodi is talking to both trade unions and factory owners*	Left arm near body, hand on Hip + Shoulder shaking Sp.ext: 0 Fluid: + 1 Power: - 1 Temp.ext: 0 Rep.: 0	I am miming those who ironically judge by looking down to us	ISM Metadiscursive		I want you to laugh about them	ISM Performative	PERS (Pathos)
3 0.00.8	*Sì, **fa**ccio il disc**o**rso con la CGI**L** e la cofind**u**stria* Ya I talk to trade unions and factory owners	Left arm near body, hand on hip, Bowing rhythmically Sp.ext: + 1 Fluid: - 0.5 Power: + 0.5 Temp.ext: + 1 Rep.: 4	I defy you	ISM Performative		I am self-confident in doing so	ISM Certainty	PERS (Ethos Competence)

Legend: IW: Information on the World; ISM = Information on the Sender's Mind; ISI = Information on the Sender's Identity; PERS = Persuasive

Table 1. The persuasive import of gestures

1. Time	2. Speech	3. Gesture description	4. Literal meaning	5. Meaning type	6. Persuasive import	7. Indir. meaning	8. Meaning type	9. Persuasive import
1 48.10	Et aux hauts dirigeants qui abîment l'entreprise en faillite comme M. Forgeat And as to the top managers who ruin the enterprises, like Mr. Forgeat	Fixed gaze to the Int.Looks at Interviewer 1 leaning head leftward, from down to up, with half-closed eyelids	I'm severe, I feel anger and indignation	ISI Personality ISM Emotion	ETHOS Competence	I struggle against injustice I ask you to feel indignation	ISI Personality ISM Performative	ETHOS Benevolence PATHOS
13 49.10	*Non, là, il faut… il faut accepter cet emploi,* No, you have…. You have to accept this job	She looks down, first right then left as if looking at two things to decide between them Eyebrows raised, Eyelids défault	Choice, choose a job I order you (to choose one)	IW Action ISM Performative		I am ridiculing S.'s proposal His proposal is too punitive	ISM Emotion ISM Negative Evaluation of opponent	PATHOS LOGOS

Table 2. The persuasive import of gaze

	Occhetto		Prodi	
Length	30”		1’32”	
Gestures	14		27	
Communicative units	24		49	
Persuasive units	20		34	
	n.	%	n.	%
Logos	1	5	8	23
Pathos	6	30	4	12
Ethos competence	9	45	17	50
Ethos benevolence	4	20	5	15

Table 4. Prodi’s and Royal’s gaze

	Prodi		Royal	
Length	53”		1’20”	
Gaze items	20		20	
Communicative units	25		25	
Persuasive units	16		22	
	n.	%	n.	%
Logos	4	25	12	54
Pathos	2	13	3	14
Ethos competence	10	62	6	27
Ethos benevolence	0	0	1	5

Table 3. Prodi’s and Occhetto’s gestures

7. References

Aristotle. (1973). Retorica. Bari: Laterza

Atkinson, M., (1984). *Our Master’s Voices. The Language and Body Language of Politics*. London, Methuen.

Attardo, S., Eisterhold, J., Hay, J., Poggi, I. (2003). Multimodal markers of irony and sarcasm. In *Humour. International Journal of Humour Research*. 16, 2, 243--260.

Bucy, E.P., Bradly, S.D. (2004) Presidential Expressions and Viewer Emotion: Counter Empathic Responses to Televised Leader Displays. In *Social Science Information*, 43(1), 59--94.

Bull, P.E. (1986). The Use of Hand Gestures in Political Speeches: Some Case Studies. In*: Journal of Language and Social Psychology*, 5, 102--118.

Calbris, G. (2003). *L'expression Gestuelle de la Pensée d'un Homme Politique*. Paris, Ed. du CNRS.

Castelfranchi, C., Parisi, D. (1980). *Linguaggio, Conoscenze e Scopi*. Bologna: Il Mulino.

Cicero, M.T. (55 B.C.). De Oratore.

Conte, R., Castelfranchi, C. (1995). Cognitive and Social Action. London: University College

Frey, S. (1998). *Die Macht des Bildes. Der Einfluss der nonverbalen Kommunikation auf Kultur und Politik*. Bern, Huber.

Hartmann, B., Mancini, M., Pelachaud C. (2002). Formational Parameters and Adaptive Prototype Instantiation for MPEG-4 Compliant Gesture Synthesis. In *Computer Animation 2002*, 111--119.

Kendon, A. (2004). Contrasts in Gesticulation. A Neapolitan and a British Speaker Compared. In C.Müller and R.Posner. (Eds.), *The Semantics and Pragmatics of Everyday Gestures*. Berlin: Weidler.

Kendon, A. (2004) *Gesture. Visible Action as Utterance*. Cambridge. Cambridge University Press.

Miceli, M., Poggi, I., de Rosis, F. (2006) Emotional and Non-Emotional Persuasion *IJIS, Special Issue on Natural Argumentation*.

Poggi, I. (2005). The Goals of Persuasion. In: *Pragmatics and Cognition* 13, 2005, pp.298--335.

Poggi, I. (2007). *Mind, Hands, Face and Body. A Goal and Belief View of Multimodal Communication*. Berlin: Weidler.

Poggi, I., Castelfranchi, C. : Dare consigli. In Ch. Humphris (a cura di) Atti del 2^{o} Seminario Internazionale per Insegnanti di Lingua. *Bollettino DILIT*, 3, 1990, pp.29-49.

Poggi, I., Pelachaud, C. (2008) Persuasive gestures and the expressivity of ECAs. In I.Wachsmuth, M.Lenzen, G.Knoblich (eds.): *Embodied Communication in Humans and Machines*. Oxford, Oxford University Press.

Poggi, I., Vincze, L. (2008), Persuasive Gaze in political discourse. AISB 2008, Symposium on *Persuasive Agents*, Aberdeen, April, 1-2, 2008.

Quintilianus, M.F. (95) *Institutio Oratoria*. Le Monnier, Firenze, Italy.

Serenari, M. (2003). Examples from the Berlin Dictionary of Everyday Gestures. In M.Rector, I.Poggi & N.Trigo (Eds.) *Gestures. Meaning and Use*. Porto: Edicoes Universidade Fernando Pessoa.

Streeck, J. (forth.) Gesture in Political Communication. A Case Study of the Democratic Presidential Candidates during the 2004 Primary Campaign. To be published in *Research on Language and Social Interaction*, forth-coming.

.

On the Contextual Analysis of Agreement Scores

Dennis Reidsma, Dirk Heylen, Rieks op den Akker
Human Media Interaction
University of Twente
E-mail: {dennisr,infrieks,heylen}@ewi.utwente.nl

Abstract

Annotators of multimodal corpora rely on a combination of audio and video features to assign labels to the events observed. The reliability of annotations may be influenced by the presences or absence of certain key features. For practical applications it can be useful to know what circumstances determined fluctuations in the interannotator agreement. In this paper we consider the case of annotations of addressing on the AMI corpus.

1. Introduction

To a large extent multimodal behaviour is a holistic phenomenon in the sense that the contribution of a specific behaviour to the meaning of an utterance needs to be decided upon in the context of other behaviours that coincide, precede or follow. A nod, for instance, may contribute in different ways when it is performed by someone speaking or listening, when it is accompanied by a smile, when it is a nod in a series of 3 or 5, etcetera. When we judge what is happening in conversational scenes, our judgements become more accurate when we know more about the context in which the actions have taken place. The record of gaze, eye-contact, speech, facial expressions, gestures, and the setting determine our interpretation of events and help to disambiguate otherwise ambiguous activities.

Annotators, who are requested to label certain communicative events, be it topic, focus of attention, addressing information or dialogue act get cues from both the audio and the video stream. Some cues are more important than others, some may be crucial for correct interpretation whereas others may become important only in particular cases. The reliability of annotations may crucially depend on the presence or absence of certain features. Also one annotator may be more sensitive to one cue rather than another. This means that the agreement between annotators may vary with particular variations in the input. Rather than relying simply on a single overall reliability score, it can be informative to know whether there are particular features that account for some of the disagreements. This may influence the choice of features to use for training machine learning algorithms.

The rest of the paper is organised as follows. First we introduce the AMI[1] project and corpus (Carletta, 2007). Then we summarize the role of addressee in interaction and its place in the AMI corpus. For the case of determining who is being addressed in the AMI data, we have looked at the reliability scores of the annotations under different circumstances. The rest of the paper discusses the results and some implications of that analysis.

2. The AMI Corpus

The AMI corpus that was used in this study consists of more than 100 hours of audio and video data of non-scripted, role played meetings (Carletta (2007). In a series of four meetings, a group of designers, marketing experts and project leaders (4 people each time) go through different phases of discussing the design of a new type of remote control. The data has been annotated on many levels. The addressee labels that are the subject of this paper are part of the dialogue act annotations (Jovanovič, 2007). For, more or less each dialogue act, annotators were instructed to indicate whether it was addressed to the group, to a specific individual. Annotators could also use the label unknown.

3. Addressee in Interaction

Addressing occurs in a variety of flavors, more explicitly or less so, verbally or non-verbally. Thus, deciding whether or not the speaker addresses one individual partner in particular can be far from trivial an exercise. In small group discussions, like those in the AMI meetings with 4 participants, most contributions are addressed to the whole group. But sometimes speakers direct themselves to one listener in particular. Group members bring in different expert knowledge and have different tasks in the design process. If someone says to a previous speaker *"can you clarify what you just said about ..."* it is clearly addressed to that previous speaker. This doesn't rule out that a non-addressed participant takes the next turn. But generally this will not happen in an unmarked way.

The basis of our concept of addressing originates from Goffman (1981). The addressee is the participant "oriented to by the speaker in a manner to suggest that his words are particularly for them, and that some answer is therefore anticipated from them, more so than from the other ratified participants". Thus, according to Goffman, the addressee is the listener the speaker has selected because he expects a response from that listener. The addressee coincides with the one the speaker has selected to take the next turn. But addressing an individual does not always imply turn-giving. For example, a speaker can invite one of the listeners to give feedback (either verbally, or non-verbal by eye-gaze) when he thinks that is required, but continue speaking.

Lerner distinguished explicit addressing and tacit addressing. To characterize the latter he writes: "When the requirements for responding to a sequence-initiating action limit eligible responders to a single participant, then that participant has been tacitly selected as next speaker. Tacit addressing is dependent on the situation and content." (Lerner, 2003). An example from our corpus is when a presenter says *"Next slide please"* during his

[1] http://www.amiproject.org.

presentation, a request that is clearly addressed to the one who operates the laptop.
Explicit addressing is performed by the use of vocatives (*"John, what do you think?"*) or, when the addressee's attention need not be called, by a deictic personal pronoun: *"What do you think?"*. There is one form of address that always has the property of indicating addressing, but that does not itself uniquely specify who is being addressed: the recipient reference term "you" (Lerner, 2003). The use of "you" as a form of person reference separates the action of "addressing a recipient" from the designation of just who is being addressed. In interactional terms, then, "you" might be termed a recipient indicator, but not a recipient designator. As such, it might be thought of as an incomplete form of address (Lerner, 2003).

Inherent Ambiguity in Addressing

At a party the host asks Ben - Ben's wife at his side - whether he wants another drink. Ben answers "No, thanks, it was an enjoyable evening, but we should go now," gazing at his wife while uttering the final excuse to his host. What is an excuse for the host is an urgent request addressed to his wife. The example shows that the same words can simultaneously express different speaker intentions directed to different addressees. The AMI annotation scheme was not devised to handle these cases. In the corpus we hardly see cases where addressing is a problem for the participants themselves. Only in a few instances, for example when the speaker uses a wrong address term, or when his utterances containing "you" is not supported by eye gaze to his intended addressee, confusion with respect to the intended addressee occurs for the participants involved in the interaction (see Op den Akker and Theune 2008 for more examples).

4. Addressee Annotation in AMI

Addressing information is part of the Dialogue Act Annotations in the AMI meeting corpus. The AMI dialogue act scheme distinguishes between 16 labels. Some of these labels are not really referring to a speech act as such but mark the special status of the utterance as a *stall*, *fragment*, *backchannel*, or *other*. Excepting these 'non-real' dialog act types, the annotators have indicated for all dialog acts of the remaining 11 types who was being addressed by the speaker: the group or a particular individual. We used meeting IS1003d of the AMI corpus which was annotated by four different annotators. Table 1 shows the confusion matrix for the full set of addressee labels for all 'agreed segments' of two of the annotators (i.e. segments where the annotators agreed on the start and end boundaries).
We ran a series of pairwise agreement analyses for each pair of annotators on the addressee labels assigned to dialogue act segments (i.e. excluding the 'non-real' dialog act types). The agreement is expressed using Krippendorff's alpha (1980). In the following sections we discuss several cases comparing scores for different label sets or class maps and different conditions (contexts).

	A	B	C	D	G	U	Σ
A	46				26	2	74
B	1	25			12	1	39
C			38	1	10	1	50
D				63	16	4	83
G	7	5	9	10	155	5	191
U	16	1	4	4	15	2	42
Σ	70	31	51	78	234	15	479

Table 1: Confusion matrix for two annotators for addressee labels of agreed segments.

Besides annotation of Goffman's notion of addressing, the meetings in the AMI corpus were also annotated with visual Focus Of Attention (FOA), an important cue for addressing behavior (see Section 3). This annotation marks for every participant in the meeting at all times throughout the meeting whom or what he is looking at. The FOA annotation was done with a very high level of agreement at a very precision: changes are marked in the middle of eye movement between old and new target (Jovanovič, 2007).

5. 'Unknown Addressee'

Annotators indicated whether an utterance was addressed to a particular person or to the whole group (note that the AMI meetings are multi-party meetings involving four participants). The annotators also had the choice to use the label *unknown addressee* in case they could not decide who was being addressed.
One can imagine two possibilities for the subset of dialog acts annotated with the *unknown* addressee label. Firstly, annotation of addressee may be a task containing inherently ambiguous instances as discussed by Poesio and Artstein (2005), with the intended addressee of some utterances being ambiguous by design. Secondly, the use of the *unknown* addressee label may reflect more the attentiveness of the annotator or his certainty in his own judgement rather than inherent properties of certain dialog acts.
The difference between the two has clear consequences for machine learning applications of the addressee annotations. It might make sense to try and learn to classify dialog act instances that are *inherently ambiguous with respect to addressing* as such, but less so to train a classifier to emulate the uncertainty of the annotators.
It is not possible to determine solely from the instruction manual which of the two interpretations most accurately reflects the meaning of the *unknown* addressee label as it was applied in the AMI corpus. Inspection of the confusion matrices however suggests that the *unknown* label is about randomly confused with every other possible addressee label. This strongly hints at the second interpretation. This conclusion is also borne out by the alpha agreement score for addressee computed on all dialog act segments vs the alpha agreement on only those dialog act segments not annotated with this *unknown* label. Leaving out the unknown addressee cases shows consistent improvements on the alpha scores, not only for the overall data set reported in Table 2 but also for each and every contextual selection of the data set reported later in this paper.

	Inc. unknown	Excl. unknown
1 vs 2	0.57	0.67
3 vs 4	0.31	0.47
4 vs 2	0.50	0.63
3 vs 2	0.36	0.47
1 vs 4	0.46	0.59
3 vs 1	0.32	0.43

Table 2: Alpha agreement for all segments vs only segments not annotated with the *unknown* addressee label.

For machine learning this suggests that it is better not to try to learn this label. For training and testing the addressee one should probably ignore the unknown addressee segments. The rest of this paper therefore reports only on segments *not* annotated with the unknown label.

6. Group/Single vs Group/A/B/C/D

The second aspect of the annotated data that we investigated in more depth was the difference between dialog act segments annotated as being *group addressed* and segments annotated as being *single addressed*, i.e. addressed to one of the individual meeting participants *A*, *B*, *C*, or *D*. Informal inspection of the confusion matrices suggests that making the global distinction between *group* and *single* addressed utterances is a difficult task: there is a lot of confusion between the label *G* on one hand and *A*, *B*, *C* and *D* on the other hand. However, if annotators see an utterance as *single* addressed they subsequently do not have much trouble determining *who* of the single participants was addressed: there is much less confusion between the 'single' labels *A*, *B*, *C* and *D*.

This is made more concrete by calculating alpha agreement for a class mapping of the addressee annotation in which the 'single' labels *A*, *B*, *C* and *D* are all mapped onto the label *S*. Table 3 shows pairwise alpha agreement for this class mapping, beside the values for the normal label set (excluding all segments annotated with the *unknown* addressee label, as described in Section 5). The consistent differences between the two columns make it is clear that agreement on *who* of the participants was addressed individually is a major factor in the overall agreement.

	Normal label set	Class map (A,B,C,D) => S
1 vs 2	0.67	0.55
3 vs 4	0.47	0.37
4 vs 2	0.63	0.52
3 vs 2	0.47	0.37
1 vs 4	0.59	0.46
3 vs 1	0.43	0.32

Table 3: Pairwise alpha agreement for full label set (left) and for class mapping (A, B, C, D) => S (right), both excluding the segments labelled *unknown*.

Agreement between annotators as to whether an utterance is addressed to the group or to an individual participant is low, but if two annotators agree that a segment is addressed to a single individual instead of the group they also agree on who this individual is.

7. Context: Focus of Attention

The visual focus of attention (FOA) of speakers and listeners is an important cue in multimodal addressing behaviour. To what extent is this cue important for annotators who observe the conversational scene and have to judge who was addressing whom?

We can start answering this question when we compare cases where the gaze is directed towards any person versus those cases where the gaze is directed to objects (laptop, whiteboard, or some other artefact), or nowhere in particular. One might expect that in the second case the annotation is harder and the agreement between annotators lower. When, during an utterance, a speaker looks at only one participant, the agreement may also be higher than when the speaker looks at more (different) persons during the utterance.

To investigate this difference we compare pairwise alpha agreement for four cross sections of the data:

1. all segments irrespective of FOA
2. only those segments during which the speaker does not look at another participant at all (he may look at objects, though)
3. only those segments during which the speaker does look at one other participant, but not more than one (he may also intermittently look at objects)
4. only those segments during which the speaker does look at one or more other participants (he may also intermittently look at objects)

In all four cross sections, only those segments were considered that were *not* annotated with the 'unknown' addressee label. Table 4 presents the pairwise alpha scores for the four conditions. Agreement is consistently lowest for condition 2 whereas conditions 3 and 4 consistently score highest.

	(1)	(2)	(3)	(4)
1 vs 2	0.67	0.60	0.78	0.77
3 vs 4	0.47	0.41	0.57	0.57
4 vs 2	0.63	0.59	0.69	0.66
3 vs 2	0.47	0.42	0.48	0.51
1 vs 4	0.59	0.57	0.63	0.62
3 vs 1	0.43	0.32	0.53	0.56

Table 4: Pairwise alpha agreement for the four contextual FOA conditions, all excluding the segments labelled *unknown*.

This shows that focus of attention is being used as an important cue for the annotators. When a speaker looks at one or more participants, the agreement between annotators on addressing consistently becomes higher. Contrary to our expectations there is no marked difference, however, between the cases where, during a segment, a speaker only looks at one participant or at more of them (cases (3) versus (4)).

8. Context: Elicit Dialog Acts

The last contextual agreement analysis that we present here concerns the different types of dialog acts. Goffman's notion of addressing that was used for the annotation of

the corpus seems to be more applicable to initiatives than to responsive acts, given that it is formulated in terms of "that some answer is therefore anticipated from [the addressee]" (Goffman, 1981). Table 5 presents the pairwise alpha agreement for only the 'elicit' dialog acts opposed to that for all dialog acts. Clearly, the agreement for 'elicit' acts is a lot higher. Apparently the intended addressee of elicits is relatively easy to determine for an outsider (annotator); a closer inspection of the instances concerned may reveal what exactly are the differences in how speakers express 'elicit' acts and other acts (see also op den Akker and Theune, 2008).

	All 'real' dialog acts	Elicits only
1 vs 2	0.67	0.87
3 vs 4	0.47	0.84
4 vs 2	0.63	0.80
3 vs 2	0.47	0.58
1 vs 4	0.59	0.76
3 vs 1	0.43	0.57

Table 5: Pairwise alpha agreement for all 'real' dialog acts (left) and for only the elicit dialog acts (right), both excluding the segments labelled *unknown*.

9. Interaction Between the Different Views

Throughout this paper we presented pairwise alpha agreement scores for different class mappings or cross sections of the addressee annotations in the AMI corpus. The different effects noted about those scores were consistent. That is, although we report only a few combinations of scores, different combinations of mappings and cross sections consistently show the same patterns. For example, all differences for the different FOA conditions hold both for the 'all segments' and the 'excluding unknown labels' condition, and for the (*A*, *B*, *C*, *D*) => *S* class mapping, etcetera. Only the 'elicit' scores were not calculated in combination with each and every other cross section.

10. Discussion

Determining who is being addressed as an outsider is not easy as the alpha scores demonstrate. The above analysis shows some of the factors that influence annotators in the choices they make by comparing alpha values for different conditions.

Reidsma and Carletta (to appear) point out that reliability measures should not be treated as simple one shot indicators of agreement between annotators. A more detailed analysis is required to judge the usability of annotations for further analysis or machine learning.

Vieira (2002) and Steidl (2005) claim that it is 'unfair' to blame machine learning algorithms for bad performance in case human annotators are equally bad or worse in reaching agreement. In general, we agree with this point of view, but we want to argue for a more fine-grained analysis that allows one to understand better the disagreements between annotators. It is very well possible that an algorithm performs badly because of completely different reasons, for which one could "blame" the algorithm. On the other hand, creating algorithms can be improved by knowing the situations in which humans disagree and the reasons that lie behind this.

Acknowledgements

The authors would like to thank the developers of the AMI annotation schemes and the AMI annotators for all their hard work, as well as Nataša Jovanovič for many discussions about addressing. This work is supported by the European IST Programme Project FP6-033812 (AMIDA, publication 99). This article only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

11. References

op den Akker, R. and Theune, M. (2008), How Do I Address You? Modelling addressing behavior based on an analysis of multi-modal corpora of conversational discourse. In: Proceedings of the AISB symposium on Multi-modal Output Generation, MOG'08, Aberdeen.

Carletta, J.C. (2007), Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus, in: Language Resources and Evaluation, 41:2(181-190)

Goffman, E. (1981), Footing. In: Forms of Talk, pages 124-159. Philadelphia: University of Pennsylvania Press.

Gupta, S., John N., Matthew P., and Jurafsky, D. (2007), Resolving "you" in multiparty dialog. In: Proceedings of 8th SigDial Workshop, pages 227-230.

Jovanovič, N. (2007), To Whom It May Concern - addressee identification in face-to-face meetings, PhD Thesis, University of Twente

Krippendorff, K. (1980), Content Analysis: An Introduction to its Methodology, Sage Publications, The Sage CommText Series, volume 5

Lerner, G.H. (2003), Selecting next speaker: The context-sensitive operation of a context-free organization. Language in Society, 32:177-201.

Poesio, M. and Artstein, R. (2005), The Reliability of Anaphoric Annotation Reconsidered: Taking Ambiguity into Account, in: Proceedings of the ACL Workshop on Frontiers in Corpus Annotations II: Pie in the Sky, pages 76-83

Reidsma D. and Carletta. J. (to appear), Reliability measurement: there's no safe limit, to appear in Computational Linguistics

Steidl, S. and Levit, M. and Batliner, A. and Nöth, E. and Niemann, H. (2005), "Of all things the measure is man" Automatic classification of Emotion and Intra Labeler Consistency. ICASSP 2005, International Conference on Acoustics, Speech, and Signal Processing

Vieira, R. (2002), How to evaluate systems against human judgment in the presence of disagreement. Proceedings of the Workshop on Joint Evaluation of Computational Processing of Portuguese

Dutch Multimodal Corpus for Speech Recognition

Alin G. Chiţu, Leon J.M. Rothkrantz
Faculty of Information Technology and Systems
Delft University of Technology
Mekelweg 4, 2628CD Delft,
The Netherlands
E-mail: {A.G.Chitu,L.J.M.Rothkrantz}@tudelft.nl

Abstract

Multimodal speech recognition gets increasingly more attention from the scientific society. Merging together information coming on different channels of communication, while taking into account the context, seems the right thing to do. However, many aspects related to lipreading and to what influences the speech are still unknown or poorly understood. In the current paper we present detailed information on compiling an advanced multimodal data corpus for audio-visual speech recognition, lipreading and related domains. This data corpus contains synchronized dual view acquired using high speed camera. We paid careful attention to the language content of the corpus and to the used speaking style. For recordings we implemented a prompter like software which controlled the recording devices and instructed the speaker to get uniform recordings.

1. Introduction

Multimodal speech recognition is getting more and more importance in the scientific community. There are however, still, many unknowns about what aspects are important when doing speech recognition especially on the lipreading side (i.e. what features hold the most useful information or how accurate must the sampling rate be and of course how do people lip-read). There is an increasing belief, common sense but also based on scientific research [see McGurk and MacDonald 1976], that people use context information acquired through different communication channels to improve the accuracy of their speech recognition. This is the case for almost everything people do throughout their existence. Hence, merging aural and visual data seems more than natural.

Although some level of agreement was achieved on what is important when trying to recognize speech, there is still large space for improvement. To answer as many questions as possible about speech recognition we need real data recordings that cover as many aspects as possible of the speech process. Hence it is not needed to say that data corpora are an important part of any sound scientific study. The data corpus should provide the means for understanding most of the important aspects of a given process, direct the development of the techniques toward an optimum solution by allowing for the necessary calibration and tuning of the methods and also give good means for evaluation and comparison. Having a good data corpus (i.e. well designed, capturing both general and also particular aspects of a certain process) is of great help for the researchers in this field as it greatly influences the research results.

There are a number of data corpora available in the scientific community; however these are usually very small and are compiled ad-hoc tailored to a specific project. Moreover, they are usually meant for person identification rather than for speech recognition. However, the time and effort needed to build a good dataset are both very large. We strongly believe that there should be some general guidelines that researchers should follow when building a data corpus. Having a standard guarantees that the resulted datasets have common properties which will give the opportunity to compare the results of different approaches of different research groups even without sharing the same data corpus. Some of the questions that need an answer and that should be taken into account are given in the following paragraphs.

The paper continues then with the main section, that presents in detail the recordings setup. We introduce here the prompter software, the video device, audio device, dual view recording, demographic data recorded and the language content. Our preliminary take home findings are given at the end, just before the references.

2. Requirements for the data corpus

A good data corpus should have a good coverage of the language such that every speech and visual item is well represented in the database, including co-articulation effects. The audio and video quality is also an important issue to be covered. An open question is however, what is the optimum sampling rate in the visual domain? Current standard for video recording frame rate ranges from 24 up to 30 frames per second, but is that enough? There are a number of issues related to the sampling rate in the visual domain. A first problem and the most intuitive is the difficulty in handling the increased amount of data, since the bandwidth needed is many times larger. A second problem is technical and is related to the techniques used for merging the audio and video channels. Namely, since it is common practice to sample the audio stream at a rate of 100 feature vectors per second, in the case when the

information is merged in an early stage, we encounter the need to use interpolation to match the two data sampling rates. A third issue, that actually convinced us to use a high speed camera, is related to the coverage of the visemes during recording, namely the number of frames per visemes. In the paper [Chițu and Rothkrantz 2007b] it was showed that the visemes coverage becomes a big issue when the speech rate increases. Figure 1 shows the poor coverage of the visemes in the case of fast speech rate as found in the DUTAVSC [Wojdeł et. al 2002]. Hence, in the case of fast speech rate the data becomes very scarce; we have a mean of 3 frames per viseme which can not be sufficient. Therefore, during the recordings we asked the speakers to alternate their speech rate, in order to capture this aspect as well.

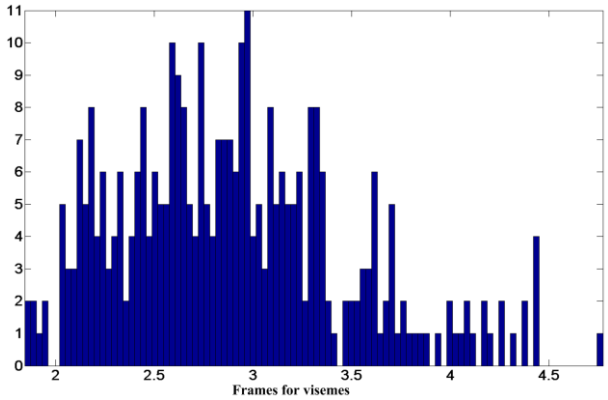

Figure 1: Viseme coverage by data in the case of fast speech rate in DUTAVSC data corpus.

A good coverage of the speaker variability is also extremely necessary. Hence there should be a balanced distribution of gender, age, education levels, and so on. An interesting aspect related to the language use was shown in the paper [Pascal and Rothkrantz 2007]. The authors show in this paper that there is an important language use difference between men and women, and between different age groups.

As we said in the beginning we aim at discovering where the most useful information for lipreading lies. We also want to give the possibility for developing new applications for lipreading. Therefore we decided to include side view recordings of the speaker's face in our corpus. A useful application could be lipreading through the mobile phone's camera. The idea of side view lipreading is not entirely new [Yoshinaga et. al 2003 and 2004]. However, it is in our opinion poorly investigated, less than a hand full of papers is dealing with this problem. Besides that, a data corpus with side view recordings is nowhere to find at this moment.

One more question would be about whether we need to have a controlled environment and alter the recording later towards the application needs or is better to record directly for the targeted application.

A thorough study of the existing data corpora can be found in [Chițu and Rothkrantz 2007]. In that scientific paper we tried to identify some of the requirements of a good data corpus and comment on the existing corpora.

Since we already had some experience with a data corpus that was built in our department, which was rather small and unfortunately insufficient for a proper training of a good lipreader or speech recognizer we decided to build a new corpus from scratch. We present in this paper, in sufficient detail, the settings of the experiment and the problems we encounter during the recordings. We believe that sharing our experiences is an important step forward towards standardized data corpora.

3. Recordings' settings

This section presents the settings used when compiling the data corpus. Figure 4 shows the complete image of the setup. We used a high speed camera, a professional microphone and a mirror for dual view synchronization. The camera was controlled by the speaker, through a prompter like software. The software was presenting the speaker the next item to be uttered together with directions on the speaking style required. This provided us with a better control of the recordings.

3.1 Prompter

Using a high speed camera increases the storage needs for the recordings. It is almost impossible to record everything and than in the annotation post process cut the clips at the required lengths. One main reason is that when recording in high speed high resolution the bandwidth limitation requires that the video be captured in the memory (e.g. on a RAM Drive). This makes the clips to have a maximum length of approximately 1 minute, depending on the resolution and color subsampling ratio used. However, we anyway needed to present the speakers with the pool of items required to be uttered. We build therefore a prompter like tool that provided the user the next item to be uttered together with some instructions about the speaking style and also controlled the video and audio devices. The result was synchronized audio and video clips already cropped to the exact length of the utterance. The tool provided the speaker the possibility to change the visual themes to maximize the visibility, and offer a better recording experience. Figure 2 shows a screenshot with the tool.

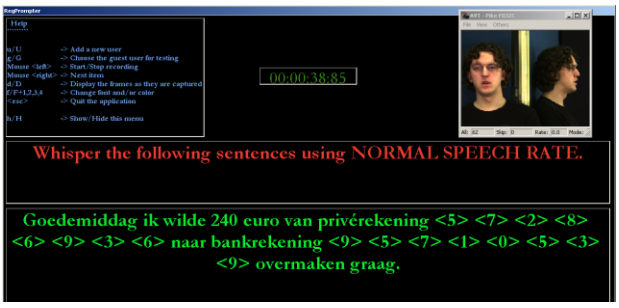

Figure 2: Prompter view during recordings.

The control of the software was done by the speaker

through the mouse buttons of a wireless mouse that was taped on the arm of the chair. After a series of trials we conclude that this level of control is sufficient and not very disruptive for the speaker. The tool was also used to keep track of the user's data, recording takes and recording sessions.

3.2 Video device

When one goes outside the range of consumer devices, things become extremely more complicated and definitely more expensive. The quality of the sensors and the huge bandwidth necessary to stream high speed video to the PC makes high speed video recording very restrictive. Fortunately, lately, by the advance made in image sensors (i.e. CCD and CMOS technology), it is possible to develop medium speed computer vision cameras at acceptable prices. We used for recording a Pike F032C camera built by AVT. The camera is capable of recording at 200Hz in black and white, 139Hz when using the chroma subsampling ratio 4:1:1 and 105Hz when using the chroma subsampling ratio 4:2:2 while capturing at maximum resolution 640X480. By setting a lower ROI the frame rate can be increased. In order to increase the Field Of View (FOV), as we will mention later, we recorded in full VGA resolution at 100Hz. To be able to guarantee a fix and uniform sampling rate and to permit an accurate synchronization with the audio signal we used a pulse generator as an external trigger. A sample frame is shown in Figure 3.

In the case of video data recording there are a larger number of important factors that control the success of the resulted data corpus. Hence, not only the environment, but also the equipment used for recording and other settings is actively influencing the final result. The environment where the recordings are made is very important since it can determine the illumination of the scene, and the background of the speakers. We use mono-chrome background so that by using a "chroma keying" technique the speaker can be placed in different locations inducing in this way some degree of visual noise.

Figure 3: Sample frame with dual view.

3.3 Audio device

For recording the audio signal we used NT2A Studio Condensators. We recorded a stereo signal using a sample rate of 48kHz and a sample size of 16bits. The data was stored in PCM audio format. Special laboratory conditions were maintained, such that the signal to noise ratio (SNR) was kept at controlled level. We considered that it is more advantageous to have very good quality recordings and degrade them in a post process as needed. The specific noise can be simulated or recorded in the required conditions and later superimposed on the clear audio data. An example of such database is NOISEX-92 [Varga and Steeneken 1993]. This dataset contains white noise, pink noise, speech babble, factory noise, car interior noise, etc. As said before special attention was paid to the synchronization of the two modalities.

3.4 Mirror

The mirror was placed at 45 degrees on the side of the speaker so that a parallel side view of the speaker could be captured synchronized with the frontal view. The mirror covered the speaker face entirely. Since the available mirror was 50cm by 70cm the holder gave the possibility to adjust the height of the mirror, thus tailoring it for all participants.

Figure 4: The setup of the experiment.

4. Demographic data recorded

As we specified in the introduction a proper coverage of the variability of the speakers is needed to assure the success of a data corpus. We also have seen that there is a language use difference between speakers. This can be used for instance to develop adaptive speech recognizers. Therefore we recorded for each speaker the following data: gender, age, education level, native language (as well as whether he/she is bi-lingual) and region where he/she had grown up. The last aspect is used to identify possible particular groups of the language's speakers, namely language dialects.

5. Language content

The language coverage is very important for the success of a speech data corpus. The language pool of our new data corpus was based on the DUTAVSC data corpus, however, enriched to obtain a better distribution of the phonemes. Hence, the new pool contains 1966 unique words, 427 phonetically rich unique sentences, 91 context aware sentences, 72 conversation starters and endings and 41 simple open questions (i.e. for these questions the user was asked to utter the first answer that they think of. In this way we expect to collect more spontaneous aspects of the speech). For each session the speaker was asked to utter 64 different items (sentences, connected digits combination, random words, and free answer questions) divided in 16 categories with respect to the language

content and speech style: normal rate, fast rate and whisper. Table 1 gives the complete set of the indications used for the speaker. The total recording time was estimated to lie in the range 45-60 minutes. The complete dataset should contain some 5000 utterances; hence a few hours of recordings, thus we target 30-40 respondents that should record 2-3 sessions. However, the data corpus is at this moment still under development.

Indication present to the speaker	Num. takes
Utter the following random combinations of digits using NORMAL SPEECH RATE.	**3**
Utter the following random combinations of digits using FAST SPEECH RATE.	**3**
Whisper the following random combinations of digits using NORMAL SPEECH RATE.	**3**
Spell the following words using NORMAL SPEECH RATE.	**3**
Spell while whispering the following words using NORMAL SPEECH RATE.	**3**
Utter the following random combinations of words using NORMAL SPEECH RATE.	**3**
Utter the following random combinations of words using FAST SPEECH RATE.	**3**
Whisper the following random combinations of words using NORMAL SPEECH RATE.	**3**
Utter the following sentences using NORMAL SPEECH RATE.	**10**
Utter the following sentences using FAST SPEECH RATE.	**10**
Whisper the following sentences using NORMAL SPEECH RATE.	**10**
Utter the following “common expressions” using NORMAL SPEECH RATE.	**5**
Answer the following questions as natural as possible.	**5**

Table 1: Indications presented to the speaker. The second column shows the number of recordings per category.

6. Conclusions

We presented in this paper our thoughts and investigation on building a good data corpus. We presented the settings used during the recordings, the language content and the recordings progression. The new data corpus should consist of high speed recordings of synchronized dual view of speaker’s face while uttering phonetically rich speech. It should provide a sound tool for training, testing, comparison and tuning a highly accurate speech recognizer.

There are still many questions to be answered with respect to building a data corpus. For instance which modalities are important for a given process, and moreover what is the relationship between these modalities. Is there any important influence between different modalities?

An important following step is to develop an annotation schema for the multimodal corpus. This is well another research topic. For an example of such a schema see [Cerrato 2004]

7. Acknowledgments

The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024.

We would like to thank Karin Driel (K.F.Driel@student.TUDelft.NL), Pegah Takapoui (pegahtak@gmail.com) and Martijs van Vulpen (mathijs@ch.tudelft.nl) for their valuable help with building the language corpus and setting up and conducting the recording sessions.

8. References

Chițu, A.G. and Rothkrantz, L.J.M. (2007). Building a Data Corpus for Audio-Visual Speech Recognition. In *Proceedings of Euromedia2007*, Delft, The Netherlands, ISBN 9789077381328, pp. 88-92.

Chițu, A.G. and Rothkrantz, L.J.M. (2007). The Influence of Video Sampling Rate on Lipreading Performance. In *Proceedings of the 12-th International Conference on Speech and Computer*, Moscow State Linguistic University, Moscow, ISBN 6-7452-0110-x, pp. 678-684.

Cerrato, L. (2004). A coding scheme for the annotation of feedback phenomena in conversational speech. In *Proceedings of the LREC Workshop on Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, Lisboa 25 May 2004 (p. 25-28)

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices *Nature*, 1976, *264*, pp. 746 - 748

Varga A. and Steeneken, H. (1993). Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. In *Speech Communication*, vol. 12, no. 3, pp. 247-251.

Wiggers, P. and Rothkrantz, L.J.M. (2007). Exploring the Influence of Speaker Characteristics on Word Use in a Corpus of Spoken Language Using a Data Mining Approach. In *Proceedings of the 12-th International Conference on Speech and Computer (SPECOM'2007)*, Moscow State Linguistic University, Moscow, ISBN 6-7452-0110-x, pp. 633-638.

Wojdeł, J.C., Wiggers, P. and Rothkrantz, L.J.M. (2002). An audio-visual corpus for multimodal speech recognition in Dutch language. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP2002)*, Denver CO, USA, September, pp. 1917-1920.

Yoshinaga, T., Tamura, S., Iwano, K. and Furui, S. (2003). Audio-Visual Speech Recognition Using Lip Movement Extracted from Side-Face Images. In AVSP2003, pp. 117–120.

Yoshinaga, T., Tamura, S., Iwano, K. and Furui, S. (2004). Audio-Visual Speech Recognition Using New Lip Features Extracted from Side-Face Images. In Robust 2004, August 2004.

Multimodal Data Collection in the AMASS++project

Scott Martens[1], Jan Hendrik Becker[2], Tinne Tuytelaars[2], Marie-Francine Moens[3]
[1] Centrum voor Computerlinguïstiek
[2] IBBT-PSI - Center for Processing Speech and Images
[3] Department of Computer Science
K.U.Leuven
Leuven, Belgium
E-mail: Scott.Martens@ccl.kuleuven.be, JanHendrik.Becker@esat.kuleuven.be,
Tinne.Tuytelaars@esat.kuleuven.be, Marie-Francine.Moens@cs.kuleuven.be

Abstract

The AMASS++ project is a project sponsored by Flemish public interests aimed at increasing the usefulness and usability of multimedia archives - notably combined text, audio and visual data - through the application of natural language and image processing technologies. To this end, we are collecting digital print media news reports, and television news programming. This data will be thematically organized and includes annotated television news programming and text media in both English and Dutch. The project's purpose is to develop and implement technologies to perform cross-language and cross-media search, summarization and user-friendly, productive presentation of results.

1. Introduction

Digital multimedia archives are now the first place many people turn to for information about current and relatively recent events, not just for ordinary media consumers but also professionals in journalism, business, academia and government. Improving the accessibility and usefulness of this class of resource is the objective of a number of public and private initiatives.

The AMASS++ project (*A*dvanced *M*ultimedia *A*lignment and *S*tructural *S*ummarization) is a project sponsored by Flemish public interests[1] aimed at increasing the usefulness and usability of multimedia archives – notably combined text, audio and visual data – through the application of natural language and image processing technologies. AMASS++ touches not only on the problem of finding materials relevant to queries, but also on the importance of presenting them in the most productive manner, rather than simply as a Google-style list of ranked pointers.[2]

The goals of this project are the development of:

- Technologies to align comparable content across media, e.g. text and video news reporting of the same events.
- Technologies for providing a structured, cross-media and cross-language summary of information about topics encompassing results from different sources, e.g. both text summaries and images.

Media firms of various kinds have shown their interest in the outcome of the project through their participation in the AMASS++ user committee, and some of them are providing us with text and video data. Additional resources are acquired through Internet crawling and recording broadcasted material.

Proof-of-concept and evaluation are performed using news materials – text news articles and televised news reports – because they are the kinds of materials whose producers can most immediately profit from the results of this project, because they are readily classifiable topically, and because they are widely available.

The scope of this project has been limited to natural language texts and image processing in order to simplify the problem and given the broad availability of subtitling and reliable transcripts. We are not considering the classification or alignment of audio or processing of the output of a speech recognition system.

2. Data contents and collection

The test data consists of text news reports (including accompanying still images in many cases), video capture data, subtitling acquired along with the video capture data, and video transcripts (where available). Because this project involves the grouping of comparable materials, data capture methods are in part oriented towards the collection of news media concerning specific events, such as the American presidential elections or the Olympic games, although for some sources, more longitudinal collection processes are also at work. Sources include news reports - in text format and televised video - in both English and Dutch originating from British, Dutch and Flemish sources.

In total, we are aiming for approximately 200 hours of

[1] AMASS++ is funded by IWT (Institute for Innovation in Science and Technology) project No. 060051, and funding for some of the video research is provided by a fellowship from the FWO (Fund for Scientific Research Flanders).
[2] See http://www.cs.kuleuven.be/~liir/projects/amass/ for further details about AMASS++.

video with subtitles, and an as yet indeterminate amount of text and web derived multimedia data.

2.1 Text data

The text data collected in this project comes from a number of sources: newspaper reports from the Internet, transcripts of TV programming, captured subtitles and autocues.

We have crawled the Google News website for URLs to articles in Dutch that Google has classified in political categories including foreign politics. Those articles are then downloaded as raw HTML and accompanying images. The HTML is filtered to separate essential content from advertisements, navigation menus, and other peripheral materials. This is challenging because articles are retrieved from a variety of news providers, each of whom structures their website somewhat differently. We use tools developed in-house for retrieving and filtering web pages. The crawler follows a breadth-first rule. The HTML filter integrates heuristic rules that balance the generality and accuracy of the filtering procedure. To date we have collected roughly 1GB of processed text data from the web.

Preprocessing. The language of the text is first identified, using procedures that select not only which language the material is in, but also assess if it is a language other than those we are prepared to process. It is then tokenized and tagged using the TnT statistical tagger (Brants, 2001), trained for Dutch using the *Corpus Gesproken Nederlands* (Oostdijk et al., 2002) and for English using the British National Corpus (Aston & Burnard, 1998); and chunked using the ShaRPa2.1 chunker (Vandeghinste, 2008). Dutch texts are also processed using an in-house decompounder (Vandeghinste 2008). The Dutch POS tagger has been trained to use a subset of the CGN/D-COI tagset (Van Eynde, 2005) and the English tagger uses the C5 tagset deployed in the British National Corpus (Aston & Burnard, 1998).

2.2 Video data

Video data is captured using a Hauppauge WinTV PVR-350 card from analog broadcasts and stored as MPEG2 at the standard 768x576 PAL resolution, although letterboxing reduces the actually used area to 750x430. Subtitles are extracted by capturing text from the subtitle teletext page. This produces HTML output for the subtitles, which includes text color information that often designates changes of speaker. Timing information for subtitles is also preserved so that realignment with the video is possible (although some manual adjustment is currently still needed). Although we have chosen not to focus on audio processing within this project, the audio is stored in 48kHz stereo format, generally encoded at roughly 200kbit/sec in MPEG2-Layer3 (a.k.a. MP3) format.

At present, we are capturing a daily news broadcast from Flemish public broadcaster VRT, and from Flemish commercial broadcaster VTM, as well as one daily news program from the BBC. We also receive higher quality video and autocue data directly from VTM.

To date, we have collected more than 50 hours of English language news broadcasts and 50 hours of Dutch news broadcasts (mostly from VTM), covering February and March 2008. Our intention is to expand our video capture procedures to cover more broadcasts oriented towards specific events in order to obtain more parallel coverage.

Preprocessing. This data is subjected to shot-cut detection and automatic keyframe extraction (Osian & Van Gool, 2004). This is necessary for the later stages of processing, in which computationally more intensive processes can be restricted to keyframes only. Additionally, we detect frontal faces in the keyframes using the method proposed by Viola and Jones (Viola & Jones, 2004) and fit a 3D morphable face model to the data (De Smet et al., 2006). We also extract local features (Bay, Tuytelaars & Van Gool, 2006; Matas, Koubaroulis & Kittler, 2002), and plan to track these over time in the near future. These local features serve as basic image representation for further high-level recognition tasks.

Ongoing work includes cleaning up the teletext output (removing repetitions), as well as an automatic tool for aligning the VTM autocues (without precise timing information) with the subtitles extracted from the teletext (aligned with the actual video data).

Figure 1: Example ground truth segmentation from a BBC news broadcast.

For a subset of the video material (30-50 news broadcasts), we are generating detailed ground truth information about story segmentation, story classification, and scene classification. See Figure 1 for an example. For this same subset, we also plan to generate ground truth connecting people's names with their faces.

3. Applications

This data is collected and processed as a test bed for the development of new algorithms for multimodal search and retrieval using advanced video and natural language processing technologies. It is, therefore, oriented towards the collection of data that provides useful tests of such technology.

For example, we intend to collect extensive coverage of the US Presidential elections in November 2008. The vast number of stories this event will no doubt generate, in all media and in many languages, makes it an excellent example of the type of event we expect to be able to use. Events will be explicitly located in time and space, in both print and visual media, and routinely described in relation to the present – the time of the news broadcast or article's publication. The elections also offer good conditions for testing algorithms to identify proper nouns and connect them to recognizable faces. Various media of different types will, inevitably, report on the same events, yielding many different reports on the same things and creating a large corpus of comparable texts and video reports. Furthermore, events will be reported on before they happen, and referred to afterwards, offering a straightforward test of text understanding algorithms designed to fix events in time.

We also intend to use coverage of the 2008 Summer Olympics to construct a corpus with narrower focus, as the coverage of the Olympics refers to fewer events outside of the Olympics themselves, and touches less on larger news stories. Furthermore, because the coverage of the Olympics is driven by specific names and times of events, named sportsmen, and explicit data about results (e.g. heights of high jumps, the times of athletes in races), it offers a chance to test algorithms designed to extract structured data from real world information sources.

4. Issues in multimodal data collection

There are a number of issues in constructing multimodal corpora that this project will have to address.

First, our focus on video and text means that there is a rather large disparity in the quantities of data we collect of each type. An hour of video with transcripts is a great deal of image data and a quite small quantity of text. A text corpus of moderate size may easily include thousands of times as much text data as the transcripts of even the largest video corpora. In order to apply the most effective techniques of corpus linguistics, we will need far more text data than the video sources alone will produce. This will involve extending the text portion of the corpus with additional materials.

Second, video data requires considerably more storage space than most other kinds of media. One high quality news broadcast of 30 minutes requires about 10GB of storage, while a comparable length of CD quality stereo audio requires only some 300MB and the corresponding teletext takes on the order of 10KB of space with lossless compression. Video data can be compressed, but generally not without some reduction in image quality. This project, therefore, has had to make a trade-off between required video quality and available storage.

Third, it is not clear what the basic unit of analysis in multimodal corpora should be. In broadcast video corpora, this is typically the *shot*. In textual media, there are a variety of possible units of analysis - documents, paragraphs, sentences, words or other units – none of which corresponds in an obvious way to shots. This poses a significant challenge both to corpus indexing and to processing.

Fourth, a sizable portion of daily news broadcasts is devoted to coverage of local events and events with little long term interest that are ill-suited to an archive of comparable materials. Using materials from different nations means that many stories that generate significant attention from one broadcaster will be completely ignored by the others. Automatically identifying these cases is another challenge to this project.

5. Related work

There have been a few other efforts to collect large multimodal datasets focused on news. TrecVid (Smeaton et al., 2006) organizes a competition focused on the analysis of video material on a yearly basis and makes a large dataset of American, Chinese and Arab news broadcasts (including some talk shows and commercials) available to its participants. However, since 2005, these have no longer included transcripts or subtitles.

The European IST project *Reveal-This*, has also made a large effort to collect multimodal data (Pastra, 2006). They have collected news programming as well as recordings of European parliament sessions and travel programs. The dataset contains materials in English and Greek.

Lastly, the Informedia project (Hauptman, 2005) has developed a fully automated process for daily content capture, information extraction and online storage. Their library contains more than 1,500 hours of daily news and documentaries produced for government agencies and public television over several years. Only a small part of their dataset has been made available through the OpenVideo Project.[3]

None of these efforts has focused on combining visual and textual material. Although this list is not intended to be exhaustive, as far as we know, there is no substantial corpus containing both video and a corresponding clean, accurate text, such as might be obtained from subtitles. Moreover, no effort to date includes Dutch language data.

6. Conclusion

We believe this to be a fairly novel class of multimodal resource and one of immediate value in the production of useful natural language and image processing systems

[3] http://www.open-video.org

with immediate real world applications.

By constructing a multi-modal corpus of news reports indexed topically, we intend to concurrently construct tools for performing alignment across media and using materials in one medium to assist in the segmentation and discovery of searchable features in another. This has potentially broad application in information retrieval, as it lends itself to the production of multimedia summaries and the enhancement of search applications.

7. References

Aston, G. and Burnard, L. (1998) *The BNC Handbook*. Edinburgh Univ. Press, Edinburgh, UK.

Bay, H., Tuytelaars, T., & Van Gool, L. (2006) Surf: Speeded up robust features. In: *Proceedings European Conference on Computer Vision 2006.*

Brants, T. (2001) *TnT – A Statistical part-of-Speech Tagger*. Published online at http://www.coli.uni-sb.de/thorsten/tnt.

De Smet M., R. Fransens, L. Van Gool. (2006) A generalized EM approach for 3D model based face recognition under occlusions, In: *Proceedings IEEE computer society conference on computer vision and pattern recognition - CVPR2006*. vol.2, p.1423-1430.

Hauptmann, A. (2005) Lessons for the Future from a Decade of Informedia Video Analysis Research. In: *International Conference on Image and Video Retrieval - CIVR'05*. LNCS 3568, pp. 1-10.

Matas, J., Koubaroulis, D. & Kittler, J. (2002) Robust wide baseline stereo from maximally stable extremal regions. In: *Proceedings of the British Machine Vision Conference*. Vol. 1. pp. 384-389.

Osian, M., Van Gool, L. (2004) Video shot characterization. *Machine Vision and Applications*, 15 (3), pp. 172-177.

Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.P., Moortgat, M., Baayen, H. (2002) Experiences from the Spoken Dutch Corpus Project. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC).* Vol.1, pp. 340-347

Pastra K. (2006) Beyond multimedia integration: corpora and annotations for cross-media decision mechanisms, In: *Proceedings of the 5th Language Resources and Evaluation Conference (LREC)*

Smeaton, A., Over, P., & Kraaij, W. (2006) Evaluation campaigns and TRECVid. In: *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval MIR '06*. pp. 321-330.

Van Eynde, F. (2005) Part-of-Speech Tagging en Lemmatisering van het D-Coi corpus. Annotation Protocol. Centrum voor Computerlinguïstiek, KU Leuven.

Vandeghinste, V. (2008) *A Hybrid Modular Machine Translation System – LoRe-MT: Low Resources Machine Translation*. LOT, Utrecht.

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

The Nottingham Multi-Modal Corpus: A Demonstration

Knight, D., Adolphs, S., Tennent, P. and Carter, R.

The University of Nottingham

The School of English Studies, The University of Nottingham, University Park, Nottingham, NG7 2RD, UK
E-mail: aexdk3@nottingham.ac.uk, Svenja.Adolphs@nottingham.ac.uk, pxt@Cs.Nott.AC.UK, Ronald.Carter@nottingham.ac.uk

Abstract

This software demonstration overviews the developments made during the 3-year NCeSS funded *Understanding New Forms of the Digital Record for e-Social Science* project (DReSS) that was based at the University of Nottingham. The demo highlights the outcomes of a specific 'driver project' hosted by DReSS, which sought to combine the knowledge of linguists and the expertise of computer scientists in the construction of the multi-modal (MM hereafter) corpus software: the Digital Replay System (DRS). DRS presents 'data' in three different modes, as spoken (audio), video and textual records of real-life interactions, accurately aligning within a functional, searchable corpus setting (known as the Nottingham Multi-Modal Corpus: NMMC herein). The DRS environment therefore allows for the exploration of the lexical, prosodic and gestural features of conversation and how they interact in everyday speech. Further to this, the demonstration introduces a computer vision based gesture recognition system which has been constructed to allow for the detection and preliminary codification of gesture sequences. This gesture tracking system can be imported into DRS to enable an automated approach to the analysis of MM datasets.

1. Introduction

This paper, and accompanying software demo, reports on some of the developments made to date on the 3-year ESRC (Economic and Social Research Council) funded DReSS (Understanding Digital Records for eSocial Science) interdisciplinary research project, based at the University of Nottingham. The linguistic concern of the project was to explore how we can utilise new textualities (MM datasets) in order to further develop the scope of Corpus Linguistic (CL hereafter) analysis. This paper discusses selected linguistic and technological procedures and requirements for developing such a MM corpus. We focus on the NMMC (Nottingham Multi-Modal Corpus, a 250,000 word corpus of single and dyadic conversational data taken from an academic discourse context), and we outline key practical issues that need to be explored in relation to mark-up and subsequent codification of linguistic and gesture phenomena.

2. Outlining the DRS

The Digital Replay System (DRS), the software used to interrogate the NMMC, aims to provide the linguist with the facility to display synchronised video, audio and textual data. In addition, perhaps most relevantly, it is integrated with a novel concordance tool which is capable of interrogating data constructed both from textual transcriptions anchored to video or audio and from coded annotations. Figure 1, below, shows an example of the concordance tool in use within the DRS environment. In this window, concordance lines for the search term *yeah* are displayed in the top right-hand panel and, as each concordance line is selected, the corresponding source video file is played to the left-hand side of the user-interface (UI hereafter).

For text based corpora (including current spoken corpora), concordance tools are nothing new. Wordsmith for example (http://www.lexically.net, see Scott, 1999) is a well known tool allowing an analyst to carry out concordance searches across large corpora of spoken or written discourse. It would be possible to export transcriptions from DRS to such a tool. However, by making such an export, we sacrifice many of the benefits of having a MM analysis tool such as DRS. DRS contains its own concordance tool. At its most basic level this allows the analyst to search across a transcription or collection of transcriptions (constituting a text only corpus) creating a concordance which displays the textual context of words or regular expressions.

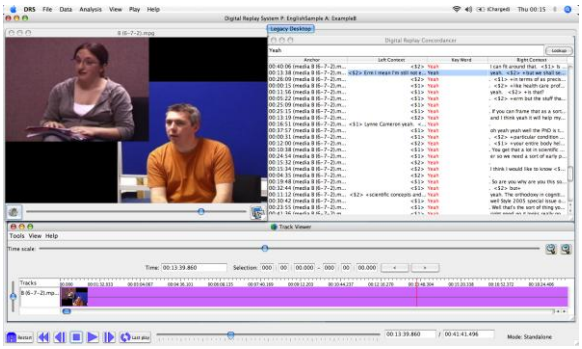

Figure 1: The concordance tool in use within the DRS environment.

Perhaps the most immediate difference between a standard text-based corpus and a MM corpus is the need to use a timeline as a means of aligning all the different data streams. This may have originally been included as a logistical necessity, but in practice it allows a degree of flexibility that standard corpus software tools do not have. Further to this, it is important to note that the trend in corpus linguistics has been towards having all the data and metadata together in one file. DRS is much more flexible in this regard. Because it uses a timeline as an anchor, the user can attach as many transcripts or annotations to that timeline as is desired. This means that data and metadata can be stored in separate files, the text can be read easily by the user without being buried by hoards of metadata records, and vice versa.

Since such reference media can be organized, indexed and stored within the DRS, we can provide more than just a textual context. Simply clicking on an instance of the utterance, we can immediately display the video of that utterance occurring, providing a far greater degree of context than is available with more traditional text-only tools. In addition, we also have coded gestures as part of the NMMC, so the DRS concordancer allows the analyst to search across the codes as well, treating them in the same way as spoken utterances. The user can therefore use DRS as an analysis tool rather than just a read-only tool already provided by existing software, thus making DRS a useful interface for a wide variety of users.

It is important to note that the concordancer is still being enhanced in order to provide frequency counts of the data. The integration of this utility will eventually allow the linguist to research statistical or probabilistic characteristics of corpora, as well as to explore specific tokens, phrases and patterns of language usage (both verbal and non-verbal) in more detail. The current version of the DRS concordancer allows the analyst to search across texts as well as within texts, and provides a reference to the text from which specific concordances were derived. Once the tracker is integrated within DRS (see below), this feature can be used to allow the linguist to search for key terms and related 'tracked' gestures in order to start to map relationships between language and gesticulation.

This novel MM concordancer has led to the need for developing new approaches for coding and tagging language data, in order to align textual, video and audio data streams (see Adolphs and Carter, 2007 and Knight, 2006). Subsequently, this demo also reports on findings of explorations (using the concordance search facility) of relationships between the linguistic characteristics and context of specific gestures, and the physically descriptive representations of those gestures extracted from video data.

The study of this relationship leads to a greater understanding of the characteristics of verbal and non-verbal behaviour in natural conversation and the specific context of learning. This will allow us to explore in more detail the relationships between linguistic form and function in discourse, and how different, complex facets of meaning in discourse are constructed through the interplay of text, gesture and prosody (building on the seminal work of McNeill, 1992 and Kendon, 1990, 1994).

3. Coding using DRS

Codes in DRS are stored in a series of 'coding tracks'. Each of these tracks is based on a timeline and associated with a particular media file. Similarly, transcripts are stored as 'annotation tracks' which behave in the same way as coding tracks, though with free rather than structured annotations. Because each utterance or code has a time associated with it, as well as a reference media, it is possible to search across these different types looking for patterns with the original media instantly accessible in the correct place. This allows the analyst to examine the context of each artifact. In order to search the data effectively, a suitable tool is required.

DRS is equipped to support the annotation and coding of raw and semi-structured data through a multistage iterative process which includes "quick and dirty" qualitative exploration of the data. This is particularly useful where rapid accessing of data and rough annotation/coding is required in order to identify passages of interest and possible variables to be included in a coding scheme.

4. Annotating MM Corpora

Traditionally linguists have relied on text as a 'point of

entry' for corpus research. However, one of the fundamental aims of this project is that all modes should be equally accessible to corpus searches, allowing not only text-based linguists but also researchers investigating the use of gesture to access data.

This principle has led to the need for new approaches for annotating and coding textual language data, in order to align them with video and audio data streams, thus enabling subsequent analysis (see Adolphs & Carter, 2007; Knight, 2006). For a MM corpus to be of use to the broader research community all streams should be accessible in order to facilitate research.

Current annotation schemes that are equipped for both gesture and speech (including, but not limited to those used within the field of linguistics) tend to only look at each mode in turn, as Baldry and Thibault (2006: 148) emphasise:

> 'In spite of the important advances made in the past 30 or so years in the development of linguistic corpora and related techniques of analysis, a central and unexamined theoretical problem remains, namely that the methods adapted for collecting and coding texts isolate the linguistic semiotic from the other semiotic modalities with which language interacts…. [In] other words, linguistic corpora as so far conceived remains intra-semiotic in orientation…. [In] contrast MM corpora are, by definition, inter-semiotic in their analytical procedures and theoretical orientations.'

Many schemes do exist, however, which depict the basic semiotic relationship between verbalisations and gesture (early coding schemes of this nature are provided by Efron 1941 and Ekman and Friesen 1968, 1969). These mark-up the occasions where gestures co-occur (or not) with the speech, and state whether the basic discoursal function of the gestures and speech 'overlap', are 'disjunct' and so on, or if the concurrent verbalisation or gesture is more 'specific' than the other sign at a given moment (for more details see Evans et al., 2001: 316). These schemes may be a useful starting point for labeling information in each mode, which can be further supplemented to cater for the semantic properties of individual features.

An example of a coding scheme that deals with defining a range of gestures based upon sequences of kinesic movements (that occur during speech) was been drawn up by Frey et al. (1983). Other more detailed kinesic coding schemes exist which attempt to define more explicitly the specific action, size, shape and relative position of movements throughout gesticulation (see Holler and Beattie, 2002, 2003, 2004; McNeill, 1985, 1992; Ekman & Friesen 1968, 1969). However, these schemes are limited in their utility for marking up the linguistic function of such sequences, and their explicit relationship to spoken discourse Other available coding schemes are not designed to provide the tools for more pragmatic analyses of language, nor to facilitate the integration of analyses of non-verbal and verbal behaviour as interrelated channels for expressing and receiving messages in discourse.

Current schemes that do classify the verbal and the visual only tend to deal with the typological features of MM talk. An example of this is given by Cerrato (2004: 26, also see Holler & Beattie's 'binary coding scheme for iconic gestures', 2002) who marks up a range HH and HCI conversations according to, primarily, whether it is a word (marked as W), phrase (marked as P), sentences (marked as S) and gestures (marked as G). Indeed steps to facilitate the exploration of both modes in conjunction have been made by various researchers and research teams (for example Cerrato 2004, discussed in more detail in chapter 4, and Dybkjær & Ole Bernsen, 2004).

Other key limitations with current coding and annotation schemes and tools are that they are not always available for general use. Instead they are often designed to meet the address a particular research question and so are difficult to expand beyond the remit of their associated research projects. For example, more extensive coding schemes that are equipped for dealing with both gesture and speech (a variety of schemes are discussed at length by Church & Goldin-Meadow, 1986 and Bavelas, 1994) are generally designed primarily to model sign language and facial expressions specifically (which can also be used for determining mouth movements in speech, as is common in HCI studies). Examples of such coding schemes are the HamNoSys (Hamburg Notation System, see Prillwitz et al., 1989), the MPI GesturePhone (from the Max Planck Institute; which transcribes signs as speech) as well the MPI Movement Phase Coding Scheme which is designed to code gestures and signs which co-occur with talk (Kita et al., 1997).

The coding scheme that is perhaps closest to the requirements of MM corpora is the MPI Movement Phase Coding Scheme. This is described as 'a syntagmatic rule system for movement phases that applies to both co-speech gestures and signs' (Knudsen

et al., 2002). However, this system does not provide detailed codes for the functional significance of the different characteristics of talk. It is a scheme that was developed at the MPI in order to allow for the referencing of video files. Annotations made with this scheme can be conducted using another MPI tool, MediaTagger, and are input into the software EUDICO for further analysis and the representation of data.

The development of a coding system that is more transferable across the different data streams in the MM corpus would be useful for the purpose of linguistic analysis. It would allow us to connect the pragmatic and semantic properties of the two gesture and speech and enable cross referencing between the two. This would make it easier to search for patterns in concordances of the data in order to explore the interplay between language and gesture in the generation of meaning.

Despite this, it has been widely recognised that it would be beneficial to create 'International Standards for Language Engineering' (known as the ISLE project, see Dybkjær & Ole Bernsen, 2004: 1). These standards are labeled as NIMMs in the ISLE project; Natural Interaction and MM Annotation Schemes. ISLE is based on the notion that there is a need for a 'coding scheme of a general purpose' to be constructed to deal with the 'cross-level and cross modality coding' of naturally occurring language data (Dybkjær & Ole Bernsen, 2004: 2-3, also refer to Wittenburg et al., 2000). Such set standards may, therefore, be of use to the development of MM corpora.

However, since gesticulations are so complex and variable in nature, it would appear difficult to create a comprehensive scheme for annotating *every* feature of gesture-in-use. Depending on the perspective of research, gestures may also be seen to have different semantic or discursive functions in the discourse. Thus it would be difficult to mark-up each respective function.

5. Coding the NMMC

Despite practical constraints, we have aimed to encode the NMMC data in a way that will 'allow for the maximum usability and reusability of encoded texts' (Ide, 1998: 1). The basic coding rubric adopted can be seen in figure 2. In order to explore the 'interaction of language and gesture-in-use for the generation of meaning in discourse', which has been the central aim for linguistic analyses using the NMMC, we initially focused upon classifying movement features and linguistic features independently.

Gestures are classified in a top-down fashion, with the analyst working to firstly define the specific form of gesture, before proceeding to establish the linguistic function (pragmatic category) of such. Knowledge from the tracking output and manual analyses determine the shape and direction of hands and whether one or both of the hands are moving at any one point. . These are then classified using the typological descriptions of gestures that are available in gesture research. The aim is to establish whether the movement features of the gesture best attribute it to being *Iconic*, *Metaphoric*, *Beat-like*, *Cohesive* or *Deictic* in nature (see McNeill, 1995, 1985, similar paradigms are seen in McNeill et al., 1994: 224; Richmond et al., 1991: 57; Kendon, 1994).

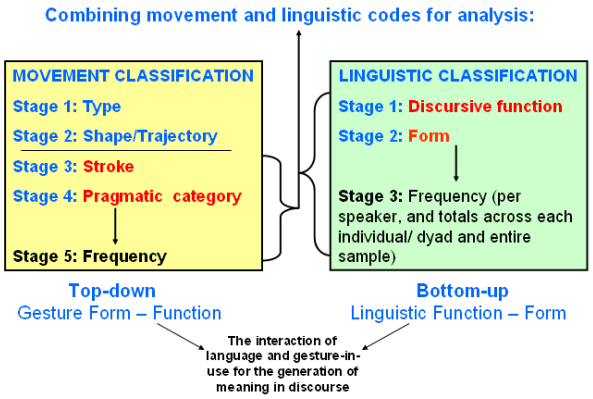

Figure 2: Coding verbal and non-verbal features of talk

In instances whether gestures co-occur specifically with speaker verbalisations (rather than with recipient gesticulations), we are working with a separate classification system in a bottom-up manner, exploring first the discursive function of co-occurring text before looking in more detail at specific tokens and phrases (which are separately encoded, see Knight and Adolphs, 2008 for details). As a final measure this information is combined in order to explore more closely specific words or phrases that are likely to co-occur with specific gestures throughout the gesture phase (and at the *stroke;* the most emphatic point, in particular).

6. Analysing gesture in the NMMC

In addition to the NMMC DRS interface, a further aim for DReSS was to develop tools which model gesture-in-talk, with the ability to monitor the *function*, *timing* and *response* (if any) of all participants, to gain an increased understanding of their role in discourse.

Although it may be feasible to manually extract and observe specific sequences of gesticulation as they occur in 10 minutes or even 5 hours of video data, it should be acknowledged that an increase in length of video data makes this method less practical and cost-effective to use. In the same way that the manual examination of pre-electronic corpora was time-consuming and error prone, the manual strategies presented here are not yet automated and rely on manual analysis The linguist has to trawl through each second of data to find features of interest, before manually marking up and encoding those features, and manipulating them before patterns and general observations can be explored.

It may therefore be appropriate to exploit a more automatic digital approach for such analysis in future. This should detect and ultimately define and encode gesture-in-talk (based on parameters pre-determined by the analyst) at high speed, and thus reduce the amount of time required to undertake such operations. In addition to this, automated methods should help to provide scientifically verifiable parameters of gesture categories and codes. One such 'automatic' approach has been developed by Computer Vision experts at the University of Nottingham (and has been tested as part of the DReSS project) in the form of a 2D gesture algorithm, which can be seen in figures 3 and 4 (for information on the technological specifications of the algorithm see Knight et al., 2006 and Evans & Naeem, 2007).

The tracker is applied to a video (represented in the form of circular nodes) of a speaker and reports in each frame the position of, for example, the speaker's hands in relation to his torso. These targets, which can be adjusted in terms of size in relation to the image, are manually positioned at the start of the video and subsequently, as the tracking is initiated, we are presented with three vertically positioned lines marking four zones on the image, R1 to R4 (R2 and R3 mark the area within shoulder width of the participant, acting as a perceived natural resting point for the arms, hence R1 and R4 mark regions beyond shoulder width).

The algorithm tracks the video denoting in which region the left hand (labeled as **R** by the tracker, since it is located to the right of the video image) and right hand (labeled as **L** by the tracker, since it is located to the left of the video image) are located in each frame. So as the video is played movement of each hand is denoted by changes in the x-axis position of R and L across the boundaries of these vertical lines. Figure 4 (overleaf) shows an alternative location matrix that can used with the tracker, dividing the video image into 16 separate zones (based on McNeill's diagram for gesture space encoding, 1992: 378) for a more detailed account of specific the horizontal and vertical movements of each hand.

The movement of each hand can therefore be denoted as a change in x-axis based region location of the hand. So when using the tracker seen in figure 3 (overleaf), we see a sequence of outputted zone 3 for frames 1 to 7, which changes to a sequence of zone 4 for frames 8 to 16 for **R**, this notifies the analyst that the left hand has moved across one zone boundary to the right during these frames. In theory, in order to track larger hand movements, the analyst can pre-determine a specific sequence of movements which can be searched and coded in the output data. So if, for example, the analyst had an interest in exploring a specific pattern of movement, considered to be of an *iconic* nature, i.e. a specific combination of the spontaneous hand movements which complement or somehow enhance the semantic information conveyed within a conversation, it would be possible to use the hand tracker to facilitate the definition of such gestures across the corpus (for in-depth discussions on iconics and other forms of gesticulation, see studies by Ekman and Friesen, 1969; Kendon, 1972, 1980, 1982, 1983; Argyle, 1975; McNeill, 1985, 1992; Chalwa and Krauss, 1994 and Beattie and Shovelton, 2002).

In both cases the tracker outputs 'raw' data into the Excel spreadsheet consisting of a frame-by-frame account of the region location of each hand (in terms of it's position within the numbered matrix; comprising of a sequence of numbers for each frame for **R** and **L**). The movement of each hand is therefore denoted as a change in region location of the hand, so for example for **R** hand (the left hand), we see a sequence of outputted zone 3 for frames 1 to 7, which changes to a sequence of zone 4 for frames 8 to 16. Ergo this notifies the analyst that the **R** hand has moved across one zone boundary to the right during these frames. Using this output the analyst would be required to 'teach' the tracking system be means of pre-defining the combination of movements to be coded as 'iconic gesture 1', for example (so perhaps a sequence of **R** or **L** hand movements into from R1 to R4 and back to R1 across x amounts of frames, for the tracker seen in figure 3), in order to convert the raw output into data which is useable.

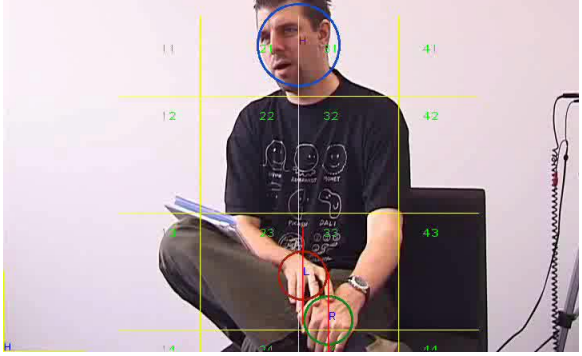

Figure 4: A 16 region version of the Hand Tracker

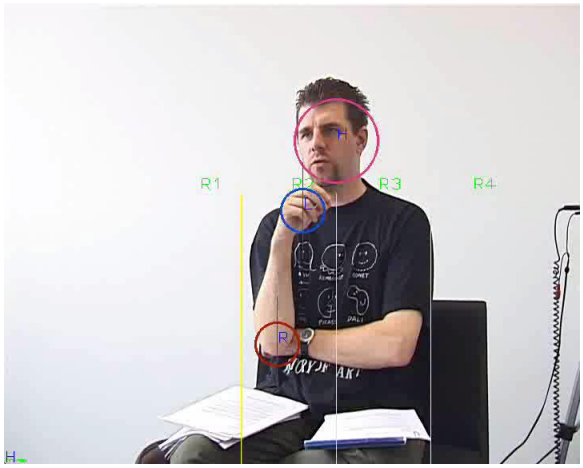

Figure 3: The initial 4 regions of the Hand Tracker

The raw data can, however, be plotted on to a basic graph, as seen in figure 5, which as a basic measure, informs the analyst whether movement does or does not occur at points throughout the video (this plot can be integrated into the DRS software). The graph maps the movement of the **L** and **R** hand across each region on the movement matrix, thus denoting movements which occur in a left or right location. This notion of movement Vs no-movement acts a useful preliminary step to classifying and encoding specific movement sequences, one which can be enacted automatically, again decreasing the amount of time required to manually extract such information. However, further to this the analyst is obviously required to determine whether these movements are in fact examples of gesticulation rather than fidgeting, for example, as regardless of how sensitive the system is, the complex nature of bodily movement makes it near impossible to determine such a difference fully automatically.

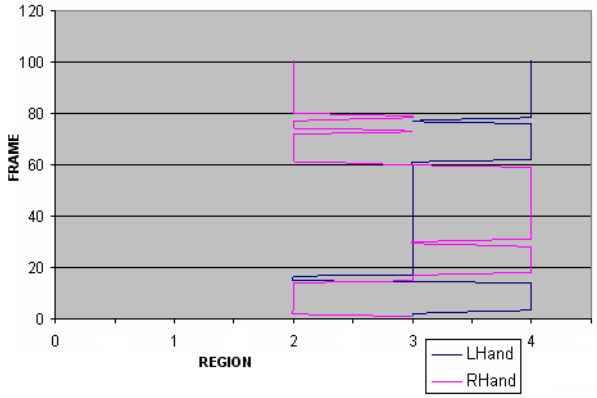

Figure 5: Plotting the tracking information (using the initial 4 region information, seen in figure 3)

It is important to note that the tracker is designed to allow the analyst to track more than one image in the same frame at the same time. In other words it has the ability for the user to apply the tracker on pre-recorded, digitised images which in theory can include up to two participants in each recorded image frame, so both participants as recorded in the NMMC corpus data comprising of dyadic academic supervisions. However,

after extensive testing, it was discovered that the tracker appears to be at its most effective when the video is of high quality (.avi) with a high resolution, with the image of each participant shown as close-up and large scale as possible. This is because smaller, lower quality images were more likely to lose the tracking target locations instantly. This requirement proved to be slightly problematic to adhere to when dealing with the streamed two-party videos from the supervision sessions because the reduction in the physical size and associated quality of the image seen in such aligned videos causes the tracker to readily lose the target locations, making it difficult for the CV algorithm to adequately track these locations. In such situations it was found that even when frequent *debugging* (when the tracker loses it's desired targets and is thus manually stopped by the analyst and the target features are redefined and relocated before the tracking is resumed) was undertaken, target locations were often instantly lost when tracking recommenced.

We further attempted to run the tracker off the original, individual .avi videos from each recording and found that, in general, the tracking algorithm was able to track the desired bodily locations with increased levels of consistency (i.e. with decreased amounts of debugging required) and accuracy with such data. Using the individual source videos rather than those which have been aligned makes the process of tracking even more lengthy as each individual participant needs to be tracked in turn rather than simultaneously. However, using the split screen version of the videos, it can be even more difficult to watch both images simultaneously and accurately stop and debug the tracker as required. Consequently, it was deemed more beneficial, and in the long term more accurate, to deal with each image individually, before attempting to align results at a later date.

Kapoor & Picard point out, in their development of a 'real-time detection' and classification tool, even with salient gestures, for example head nods and shakes, it is difficult to fully automate this stage (2001). This is due to the fact that gestures-in-talk are spontaneous, idiosyncratic (Kendon, 1992) and transient (Bavelas, 1994: 209), and are generally seen to contain no standard forms in conversation (it is unlikely that two hand motions will be exactly the same, for example). Instead, they differ according to user, intensity, meaning and in terms of how the head or hand is rotated, i.e. whether it is simply a rigid up and down, left or right movement, or whether there is more of an up and slight rotation. This complexity in form means it is not easy to accurately encode and quantify particular movement features, especially if relying on purely automated methods. The intervention of the expert human analyst is still paramount in this context.

The gesture tracker, in its current, generates a fairly simplistic set of codes. It is generally possible to tell if a large gesture has occurred, but difficult to differentiate between different types of gesture. To some extent it serves to create a 'code-template' from which a skilled analyst can apply a more detailed coding scheme to generate a more complete description of the gestures captured in a given video session. Even from the tracker's simplistic codes it is possible to search for co-occurrences of gesture and utterance, but with a more detailed coding track – generated either by simply hand-coding the video using DRS's comprehensive coding tools, or by taking the code-template generated by the tracker and filling out the detail.

7. Summary

This demonstration paper has started to outline some of the technical and practical problems and considerations faced in the development and exploration of MM corpora. It presents a novel MM corpus UI (user-interface), the DRS. DRS provides the analyst with an easy-to-use corpus tool-bench for the exploration of relationships between the linguistic characteristics and context of specific gestures, and the physically descriptive representations of those gestures extracted from video data (using the novel MM concordancer).

8. Acknowledgements

The research discussed on which this article is based is funded by the UK Economic and Social Research Council (ESRC), e-Social Science Research Node *DReSS* (Grant N$^{o.}$ RES-149-25-0035, http://www.ncess.ac.uk/research/digital_records), and the ESRC e-Social Science small grants project *HeadTalk* (Grant N$^{o.}$ RES-149-25-1016).

9. References

Adolphs, S. & Carter, R. (2007). Beyond the word: New challenges in analysing corpora of spoken English. *European Journal of English Studies 11*(2).

Argyle, M. (1975). *Bodily Communication*. London: Methuen.

Baldry, A. & Thibault, P.J. (2006). *Multimodal Transcription and Text Analysis: A multimedia toolkit and coursebook.* London: Equinox.

Bavelas, J.B. (1994). Gestures as part of speech: methodological implications. *Research on Language and Social Interaction 27,* 3: 201-221.

Beattie, G. & Shovelton, H. (2002). What properties of

talk are associated with the generation of spontaneous iconic hand gestures? *British Journal of Social Psychology 41*, 3: 403-417.

Cerrato, L. (2004). A coding scheme for the annotation of feedback phenomenon in everyday speech. *LREC Workshop on Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, Lisboa. pp. 25-28.

Chawla, P. & Krauss, R. M. (1994). Gesture and speech in spontaneous and rehearsed narratives. *Journal of Experimental Social Psychology 30*: 580-601.

Church, R.B. & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition 23*, 1: 43-71.

Dybkjær, L. & Ole Bernsen, N. (2004) Recommendations for natural interactivity and multimodal annotation schemes. *Proceedings of the LREC'2004 Workshop on Multimodal Corpora*, Lisbon, Portugal. pp. 5-8.

Efron, D. 1972. (1941). *Gesture, Race and Culture.* The Hague: Mouton & Co.

Ekman, P. & Friesen, W. (1968). Nonverbal behavior in psychotherapy research. In J. Shlien (ed), *Research in Psychotherapy*. Vol. III. American Psychological Association. pp.179-216.

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

Evans, J.L., Alibali, M.W. & McNeill, N.M. (2001). Divergence of verbal expression and embodied knowledge: Evidence from speech and gesture in children with specific language impairment. *Language and Cognitive Processes 16*, 2-3: 309-331.

Evans, D. and Naeem, A. (2007). "Using visual tracking to link text and gesture in studies of natural discourse", *Online Proceedings of the Cross Disciplinary Research Group Conference 'Exploring Avenues to Cross-Disciplinary Research'*, November 7, University of Nottingham.

Frey, S., Hirsbrunner, H.P., Florin, A., Daw, W. & Crawford, R. (1983). A unified approach to the investigation of nonverbal and verbal behaviour in communication research. In Doise, W. & Moscovici, S. (Eds.), *Current issues in European Social Psychology*. Cambridge: Cambridge University Press.

Holler, J. & Beattie, G. (2002). A micro-analytic investigation of how iconic gestures and speech represent core semantic features in talk. *Semiotica 142*, 1-4: 31-69.

Holler, J. & Beattie, G. (2003). How iconic gestures and speech interact in the representation of meaning: Are both aspects really integral to the process? *Semiotica 146*, 1-4: 81-116.

Holler, J. & Beattie, G.W. (2004). The interaction of iconic gesture and speech. *5th International Gesture Workshop*, Genova, Italy. Selected Revised Papers. Heidelberg: Springer Verlag.

Ide, N. (1998). Corpus encoding standard: SGML guidelines for encoding linguistic corpora. *First International Language Resources and Evaluation Conference*, Granada, Spain.

Kapoor, A. & Picard, R.W. (2001). A Real-Time head nod and shake detector. *ACM International Conference Proceedings Series*. pp.1-5.

Kendon, A. (1972). Some relationships between body motion and speech. In Seigman, A. & Pope, B. (Eds.), *Studies in Dyadic Communication*. Elmsford, New York: Pergamon Press. pp.177-216.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In Key, M.R. (Ed), *The Relation between Verbal and Non-Verbal Communication*. pp. 207-227.

Kendon, A. (1982). The organisation of behaviour in face-to-face interaction: observations on the development of a methodology. In Scherer, K.R. & Ekman, P. (eds) *Handbook of Methods in Nonverbal Behaviour Research*. Cambridge: Cambridge University Press.

Kendon, A. (1983). Gesture and Speech: How they interact. In Wiemann, J. & Harrison, R. (Eds.), *Nonverbal Interaction*. California: Sage Publications. pp.13-46.

Kendon, A. (1990) *Conducting Interaction.* Cambridge: Cambridge University Press.

Kendon, A. (1992). Some recent work from Italy on quotable gestures ('emblems'). *Journal of Linguistic Anthropology 2*, 1: 77-93.

Kendon, A. (1994). Do gestures communicate? A review. *Research on Language and Social Interaction 27*, 3: 175-200.

Kita, S., van Gijn, I., & van der Hulst, H. (1997). Movement Phase in Signs and Co-Speech Gestures, and Their Transcriptions by Human Coders. *Gesture Workshop 1997*: 23-35.

Knight, D. (2006). 'Corpora: The Next Generation', Part of the AHRC funded online *Introduction to*

Corpus Investigative Techniques, The University of Birmingham. http://www.humcorp.bham.ac.uk/

Knight, D. and Adolphs, S. (In Press, 2008) Multi-modal corpus pragmatics: the case of active listenership. In Romeo, J. (ed.) *Corpus and Pragmatics*. Berlin and New York: Mouton de Gruyter.

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

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

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review 92*, 3: 350-371.

McNeill, D. (1992). *Hand and Mind.* Chicago: The University of Chicago Press.

McNeill, D. (1995). *Hand and mind: What gestures reveal about thought.* Chicago: The University of Chicago Press.

McNeill, D., Cassell, J., McCullough, K-E. (1994). Communicative effects of speech-mismatches gestures. *Research on Language and Social Interaction 27*, 3: 223-237.

Prillwitz, S., Leven, R., Zienert, H., Hanke, T. & Henning, J. (1989). *HamNoSys. Version 2.0. Hamburg Notation System for Sign Language. An Introductory Guide.* Hamburg: Signum.

Richmond, V.P., McCroskey, J.C. & Payne, S.K. (1991). *Nonverbal Behaviour in Interpersonal Relations*. Prentice Hall: New Jersey.

Scott, M. (1999). *Wordsmith Tools*. Oxford: Oxford University Press.

Wittenburg, P., Broeder, D. & Sloman, B. (2000). Meta-description for language resources. *EAGLES/ ISLE White paper.* Available online from http://www.mpi.nl/world/ISLE/documents/papers/white_paper_11.pdf. [Accessed 2006-02-10]

Analysing Interaction: A comparison of 2D and 3D techniques

Stuart A. Battersby[1], Mary Lavelle[1], Patrick G.T. Healey[1] & Rosemarie McCabe[2]
[1] Interaction, Media & Communication, Department Of Computer Science, Queen Mary, University of London
[2] Unit for Social and Community Psychiatry, Queen Mary, University of London

E-mail: stuart@dcs.qmul.ac.uk, maryl@dcs.qmul.ac.uk, ph@dcs.qmul.ac.uk, r.mccabe@qmul.ac.uk

Abstract

Human interaction is inherently three dimensional and multi-party. The human body is 3D, performing actions within a 3D space. 3D aspects of communication such as the orientation and spatial arrangement of participants, collaborative gestures and synchrony are *central* to how communicative signals are formed and interpreted within an interaction. Traditional interaction research has relied on 2D audio-visual techniques to record interaction, however a comprehensive analysis of the 3D aspects which guide the interpretation and formation of interaction cannot be conducted by viewing interaction on a 2D plane. An emerging alternative to the traditional 2D techniques is the use of 3D motion capture technology. Although this technology has had extensive use within the film industry, few studies have used motion capture as a tool for examining interaction. This paper provides a comparative account of 3D motion capture techniques and traditional 2D audio-visual techniques for analysing interaction using data from a pilot study. The merits of 3D techniques are demonstrated. The potential of 3D motion capture technology to enhance and build upon previous studies within the field of human interaction is discussed and we present one example of an application area.

1. Introduction

Natural human interaction presents two key analytic and technical challenges. The first of these is that human communication is multi-modal. Modalities such as speech and gesture are not interpreted as separate channels but as composite, integrated communicative signals (Engle,1998). For example, if people hear the sound “ba” while watching the lip movements for “ga” the perceived phoneme is a blend “da” - not “ba” + “ga” (McGurk & MacDonald,1976). The second challenge, and the one of primary concern here, is that human communication is also multi-party. The way we produce and interpret communicative signals is also sensitive to the way they are placed, in space and time, with other people's utterances, gestures, body position and orientation (Kendon, 1990,1992; Ozyurek, 2000, 2002; Scheflen, 1975). In face-to-face interactions people use the three-dimensional arrangement of their bodies and gestures in space as a resource for organising and interpreting their interaction. Where a gesture is placed matters just as much as its specific shape.

Previous studies have used audio or audio-visual techniques to observe interaction, ranging from studies in the late sixties and early seventies (Condon & Ogston,1966; Kendon, 1970) through to those of the current day (McCabe, Leudar & Antaki, 2004). Video recordings provide considerably richer data than audio recordings in analysing interaction. Nevertheless, these techniques are restricted in the granularity of information they provide. These limitations become more pronounced when analysing multi-party interaction. As audio-visual techniques record information on a 2D plane, important features of interaction such as the interpretation of gestures, use of interactional space, directionality of gaze, and differing participant perspectives are diminished.

An emerging alternative to these techniques is the use of 3D motion capture technologies. Although motion capture has, to date, had extensive use within the film industry and for the capture of an individual's movements, few studies have made use of motion capture as a tool for analysing interaction. With the increased granularity from 3D motion capture information, we can examine more subtle aspects of interaction which are not accessible when restricted to 2D visual data.

This paper will discuss the possibilities of 3D motion capture systems to enhance and build on previous interaction studies focusing upon the use of space, collaborative gesture and interactional synchrony. We will begin by discussing the 3D nature of interaction, and how this can be analysed using motion capture compared to traditional audio-visual techniques. We will also present a pilot study comparing these two approaches and discuss how new advances in 3D techniques could be employed to advance research in the field of interaction generally and present one example of an application area, i.e. mental health.

2. The 3D Nature of Interaction

Interaction is inherently 3D; the human body is 3D, and in interaction our bodies are oriented in 3D space performing 3D actions. A number of research studies have identified the relevance of the 3D aspects of interaction, especially the spatial arrangement of interactional participants (Kendon, 1992, 1990, 1973; Scheflen, 1975; Vine, 1975), collaborative gesture (Ozyurek, 2000, 2002) and synchrony (Condon & Ogston,1966; Kendon, 1970). It is important to note that these aspects are not peripheral add-ons, but *central* to how interaction is constructed and comprehended.

Kendon proposes that each individual has a space directly in front of them within which all their activities occur, this is termed the *transactional segment*. When an interaction occurs, individuals position themselves to allow their transactional segments to overlap, creating a jointly managed interactional *o-space*. The system of spatial and postural arrangements of participants in maintaining and sustaining this interactional space is termed the *f-formation* (Kendon, 1990, 1992). It is clear that these formations are highly 3D; the spaces created fall on the horizontal, sagittal and vertical planes.

Within the interactional o-space we find participants' gestures. These are crucial 3D aspects of the interaction, particularly when formed collaboratively. Previous research suggests that gesture formation is strongly influenced by the location of the addressee (Ozyurek, 2000, 2002), thus the spatial arrangement of the individuals will not only impact what formation the gestures take but also how they are interpreted by the addressee. Therefore a comprehensive analysis of gesture cannot be achieved without taking into account where they occur in space and in relation to the addressee.

Early work by Condon & Ogston and Kendon, suggested that an interactional synchrony of body movement exists around the o-space and this can be related to the management of turn-taking and topic change. Condon & Ogston observed a distinct lack of interactional synchrony in patients with a diagnosis of schizophrenia, a population noted for their impaired social functioning (Ihnen, 1998). In this study, Condon & Ogston transcribed a behavioural stream coding both body movement and speech simultaneously. For the purpose of this paper we will firstly show how the audio-visual techniques used within this study only allow for a single view point on the 3D aspects, and secondly, demonstrate how motion capture can offer improvements to these traditional techniques.

3. 2D & 3D Techniques

Traditionally, two-dimensional audio-visual techniques such as video recording have been used to analyse interaction. This allows for visually rich capture *from one perspective*. As we have discussed, human interaction is 3D; given a single 2D perspective, much of the interaction is missed. At the most basic level, participants' bodies and limbs may obstruct the view of key areas of interaction from the video camera; this will be especially true in a multi-party situation. There may also be more than one key area of interest at one time, which cannot be captured by a video camera.

From an analytical point of view, with audio-visual recordings the analyst is unable to view the impact of the interaction from the perspective of each of the interactional participants; for example, a gesture may appear vastly different from one viewpoint to another, and if this gesture was created and designed to be delivered to a specific participant, adopting the incorrect viewpoint will lose vital information.

When using 3D systems, we eliminate the issue of a single perspective; this may simply be achieved by combining many 2D video cameras to create a 3D image. However, by using more precise motion capture systems to perform an analysis, the analyst can capture the exact 3D co-ordinates of each area of interest on the body, with any number of participants. This information can be used, amongst other things, to build 3D wire frame representations of the subjects. These representations can then be viewed from any angle or position desired. The playback of this captured data has the added benefits of being able to exclude unwanted subjects and can be viewed at any speed. Alternatively, the capture data can be fed to statistical analysis applications, where it can be directly analysed using cross-correlations to identify co-ordinated spatial and temporal patterns of body movement; a technique which would have unrealistic time requirements without this automation.

4. Example from Pilot study

To demonstrate the use of motion capture as a tool for interaction research, we performed a pilot study in the Augmented Human Interaction Laboratory at Queen Mary, University Of London. This is equipped with a Vicon motion capture system consisting of twelve infra-red cameras which track reflective markers attached to the clothing of subjects. The study saw three participants improvise a scene for a soap opera. Two of the participants (one male, participant A and one female, participant B) were asked to play the parts of a couple, participant B (the female partner) was told that she was having an affair with the third participant (participant C). The scene begins with participant C joining the couple at a pub table moments after participant B has confessed the affair to her partner (participant A). The scene which ensued was both video recorded (See Figure 1) inline with traditional audio-visual techniques, and motion captured (see Figure 2).

Figure 1: 2D Video – Fixed Viewpoint

These figures demonstrate some of the problems which can arise using 2D methods which could be overcome with the use of 3D motion capture. If we focus on the movements of the male participant (participant A) at the far left of the 2D image (see Figure 1), we can see his left arm is obscured from view by his right arm; equally both of his arms are blocking participant C's body from full

view. Further, we are unable to determine how this gesture could be interpreted from the perspectives of the other participants; we may only adopt the perspective of an outsider looking in on the interaction.

We can now view the same scene using reconstructed data from the 3D capture (Figure 2). Here we have adopted the perspective of the female, participant B, looking towards the open armed gesture of participant A. From this we have full view of both of his arms and their positions in 3D space, and we can adopt an equally impressive view of participant C (or indeed the perspective of participant C).

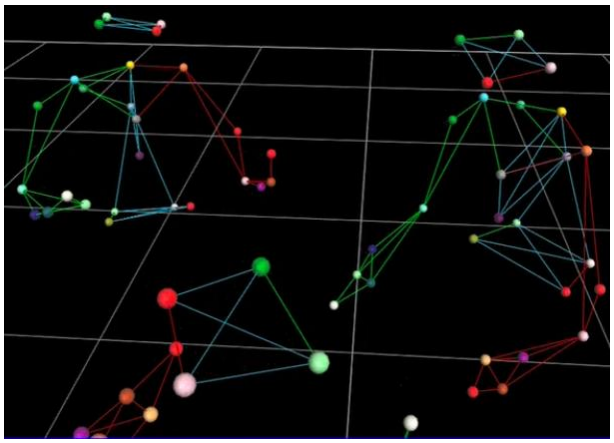

Figure 2: 3D Capture

This 3D reconstruction allows us to examine the scene at a much finer level of granularity than was available from the video recording. This is due to our knowledge of the exact 3D co-ordinates of the body markers, captured at 120 frames per second.

Given these features of the motion capture system, we are able to detect elements of interaction and body movements at levels not visible using the traditional audio-visual techniques employed by Condon & Ogston and Kendon. This process is further aided in the 3D representation by removing unwanted visual 'noise' and only presenting the discrete body segments of interest for the coding. The computerised nature of the motion capture system allows for the potential of an automated coding system to be developed; we currently have in place a system to detect overlapping head orientations in dyadic interaction which can be extended to detect other interactional phenomena.

5. Conclusion

We have presented a comparative account of 3D motion capture systems and traditional 2D techniques for recording interaction. Using data from a pilot study we have demonstrated the merits of 3D systems over 2D techniques in observing the central features of interaction such as the spatial arrangement of participants, collaborative gestures and synchrony.

It is clear from this preliminary study that motion capture technology has great potential for advancing interaction research with implications for other areas of research such as mental health, in particular schizophrenia. As mentioned previously, impaired social functioning is a hallmark of schizophrenia (Ihnen, 1998), indeed it may be that interactional impairments are more important than symptoms (or cognitive functioning) in predicting long term outcome in schizophrenia (Couture, 2006). Although a growing body of work has attempted to explain these interactional deficits there is little understanding of what underlies these interactional problems. 3D technology would allow for a more comprehensive analysis of social interaction within this patient group, potentially allowing us to identify the precise nature of the impairment, which in-turn would enable these deficits to be targeted therapeutically.

6. References

Condon, W., Ogston, W. (1966). Sound film analysis of normal and pathological behaviour patterns. *The Journal of Nervous and Mental Disease*. 143, pp. 338--347.

Couture, S., Penn, D., Roberts, D. (2006). The functional significance of social cognition in schizophrenia: a Review. *Schizophrenia Bulletin*, 32, (S1) pp. 44–63.

Engle, R. (1998). Not Channels but composite signals: Speech, gesture, diagrams and object demonstrations are integrated in multimodal explanations. In M Gernsbacher, & S. Derry (Eds.), *Proceedings of the 20th Annual Conference of the Cognitive Science Society*. pp. 321—326.

Ihnen, G., Penn, D., Corrigan, P., Martin, J. (1998), Social perception and social skills in schizophrenia. Psychiatry Research, 80, pp. 275—286.

Kendon, A. (1970). Movement coordination in social interaction: Some examples described. *Acta Psychologica*. 32, pp. 100—125.

Kendon, A. (1973). The role of visible behaviour in the organization of social interaction. In M. Von Cranach & I. Vine (Eds.) *Social Communication and Movement*. London Academic Press.

Kendon, A. (1990). *Conducting Interaction: Patterns of behaviour in focused encounters*. Cambridge University Press.

Kendon, A. (1992). The negotiation of context in face-to-face interaction. In A. Durati & C. Goodwin (Eds.) *Rethinking context: Language as in interactive phenomenon.* Cambridge University Press.

McCabe, R., Leudar, I., Antaki, C. (2004). Do people with a diagnosis of schizophrenia display theory of mind deficits in naturalistic interaction? *Psychological Medicine*. 34, pp. 401--412.

McGurk, H., MacDonald, J. (1976). Hearing lips and seeing voices. *Nature.* 264, pp. 746—748

Ozyurek, A. (2000). The influence of addressee location on spatial language and representational gestures of direction. In D. McNiell (Ed.) *Language and Gesture*. Cambridge University Press.

Ozyurek, A. (2002). Do speakers design their cospeech gestures for their addressees? The effects of addressee location on representational gestures. *Journal of Memory and Language*. 46. pp. 688—704.

Scheflen, A. E., (1976). *Human Territories: how we behave in space-time*. Prentice-Hall.

Vine, I. (1975). Territoriality and spatial regulation of interaction. In A. Kendon, R.M. Harris, &R.M, Key (Eds.), *Organisation of Behaviour in face-to-face interaction*. Mouton Publishers The hague: Paris. pp. 357—388