The 10th Workshop on Multimodal Corpora: Combining Applied and Basic Research Targets

Workshop Programme

09:00 – 09:15 Registration

09:15 – 09:30 Welcome

09:30 – 10:30 Keynote
Hannes Högni Vilhjálmsson: *TBA*

10:30 – 11:00 Coffee break

11:00 – 13:00 Session 1 (Oral)
Masashi Inoue, Toshio Irino, Ryoko Hanada, Nobuhiro Furuyama and Hiroyasu Massaki: *Continuous Annotations for Dialogue Status and Their Change Points*
Eli Pincus and David Traum: *Towards a Multimodal Taxonomy of Dialogue Moves for Word-Guessing Games*
Bayu Rahayudi, Ronald Poppe and Dirk Heylen: *Gaze Patterns in the Twente Debate Corpus*
Kirsten Bergmann, Ronald Böck and Petra Jaecks: *EmoGest: Investigating the Impact of Emotions on Spontaneous Co-speech Gestures*

13:00 – 14:00 Lunch break

14:00 – 15:00 Session 2 (Oral)
Ivan Gris, David Novick, Mario Gutierrez and Diego Rivera: *The "Vampire King" (Version 2) Corpus*
Kristina Nilsson Björkenstam and Mats Wirén: *Multimodal Annotation of Synchrony in Longitudinal Parent-Child Interaction*

15:00 – 16:00 Session 3 (Poster)
Federica Cavicchio, Amanda Brown, Reyhan Furman, Shanley Allen, Asli Özyürek, Tomoko Ishizuka and Sotaro Kita: *Annotation of space and manner/path configuration in bilinguals' speech and manual gestures*
Oliver Schreer and Stefano Masneri: *Automatic Video Analysis for Annotation of Human Body Motion in Humanities Research*
Trine Eilersen and Costanza Navarretta: *A Multimodal Corpus of Communicative Behaviors of Disabled Individuals during HRI*
Jens Edlund, Mattias Heldner and Marcin Wlodarczak: *Catching wind of multiparty conversation*

16:00 – 16:30 Coffee break

16:30 – 10:30 Session 4 (Oral)
Gabriel Murray: *Resources for Analyzing Productivity in Group Interactions*
Nesrine Fourati, Jing Huang and Catherine Pelachaud: *Dynamic stimuli visualization for experimental studies of body language*

17:30 – 18:00 Business meeting, close

Editors

Jens Edlund	KTH Royal Institute of Technology, Sweden
Dirk Heylen	University of Twente, The Netherlands
Patrizia Paggio	University of Copenhagen, Denmark/University of Malta, Malta

Workshop Organizers

Jens Edlund	KTH Royal Institute of Technology, Sweden
Dirk Heylen	University of Twente, The Netherlands
Patrizia Paggio	University of Copenhagen, Denmark/University of Malta, Malta

Workshop Programme Committee

Jens Allwood	University of Gothenburg, Sweden
Susanne Burger	Carnegie Mellon University, USA
Jens Edlund	KTH Royal Institute of Technology, Sweden
Dirk Heylen	University of Twente, The Netherlands
Costanza Navarretta	University of Copenhagen, Denmark/University of Malta, Malta
David Novick	University of Texas at El Paso, USA
Patrizia Paggio	University of Copenhagen, Denmark
Ronald Poppe	University of Twente, The Netherlands
Albert Ali Salah	Boğaziçi University, Turkey
David Schlangen	Bielefeld University, Germany
David Traum	Institute for Creative Technologies, USA

Table of contents

Author Index

Introduction

We are pleased that in 2014, the 10th Workshop on Multimodal Corpora is once again returning home and is collocated with LREC, this time in Reykjavik, Iceland. The workshop follows in a series previously held at LREC 2000, 2002, 2004, 2006, 2008, 2010; at ICMI 2011; at LREC 2012; and at IVA 2013 (all workshops of the series are documented under www.multimodal-corpora.org).

As always, we present a wide cross-section of the field, with contributions ranging from collection efforts, coding, validation and analysis methods, to tools and applications of multimodal corpora. Given that LREC this year emphasizes the use of corpora to solve language technology problems and develop useful applications and services, we aim for this workshop also to highlight the usefulness of multimodal corpora to applied research as well as basic research. Many of the unimodal speech corpora collected over the past decades have served a double purpose: on the one hand, they have enlightened our view on the basic research question of how speech works and how it is used; on the other hand, they have forwarded the applied research goal of developing better speech technology applications. This reflects the dual nature of speech technology, where funding demands often require researchers to follow research agendas that target applied and basic research goals in parallel.

Multimodal corpora are potentially more complex than unimodal corpora, and their design poses an even greater challenge. Yet the benefits to be gained from designing with a view to both applied and basic research remain equally desirable. Against this background, the theme for this instalment of Multimodal Corpora is how multimodal corpora can be designed to serve this double purpose.

Continuous Annotations for Dialogue Status and Their Change Points

Masashi Inoue[1], Toshio Irino[2], Ryoko Hanada[3], Nobuhiro Furuyama[4]
Hiroyasu Massaki[5]

[1]Yamagata University
[2]Wakayama University, [3]Kyoto University of Education
[4]National Institute of Informatics/Tokyo Institute of Technology, [5]Graduate University for Advanced Studies
3-16, 4 Jyonan, Yonezawa-shi, Yamagata, Japan
mi@yz.yamagata-u.ac.jp

Abstract

This paper presents an attempt to continuously annotate the emotion and status of multimodal corpora for understanding psychotherapeutic interviews. The collected continuous annotations are then used as the signal data to find change points in the dialogues. Our target dialogues are carried between clients with some psychological problems and their therapists. We measured two values, namely the degree of the dialogue progress and the degree of clients being listened to. The first value reflects the goal-oriented nature of the target dialogues. The second value corresponds to the idea of active listening that is considered as an important aspect in psychotherapy. We have modified an existing continuous emotion annotation toolkit that has been created for tracking generic emotion of dialogues. By applying a change point detection algorithm on the obtained annotations, we evaluated the validity and utility of the collected annotation based on our method.

Keywords: Continuous Annotation, Emotion, Change Point

1. Introduction

We created a multimodal video corpus of about 20 psychotherapeutic dialogues (Inoue et al., 2012) to better understand the nature of psychotherapy. In this paper, we describe a new modality annotation assigned to the corpus using a continuous annotation toolkit and the result of the analysis on the annotation. Although the annotation is generic and can be used in various analyses, we first focus on its utility in identifying the change points of the dialogues. In the following sections, we shall describe the toolkit used, the dimensions of collected annotations, and the result of the initial analysis of data.

2. Emotion Tracking Tool

We are currently investigating the emotional scores assigned to the video sequences. For this purpose, we developed a scoring interface called Emotional Movement Observation (EMO). EMO is designed for continuous measurement of emotion in a conversation. The video of the target dialogue is shown in a window. The users of the interface, called an annotator hereafter, listens to the dialogue between two people facing each other. The evaluation window contains a square-shaped grid as shown in Figure 1. The current values are highlighted by a circle pointer. Various interfaces have been proposed for the tracking of emotions. A circular-shaped evaluation interface has been used for emotion annotation in speech (Cowie et al., 2000). An alternative interface design is an array of sliding bars. Such interfaces have been used for long-term emotion tracking (McDuff et al., 2012) and for the evaluation of dialogue systems (Inventado et al., 2011). A button selection interface with representative emotive images has been applied for evaluating the web interface (Huisman et al., 2013). EMO's interface is most similar to EMuJoy that was developed for measuring music emotion (Nagel et al., 2007). The two-dimensional evaluation window of our EMO system contains axes for pairs such as pleasant/unpleasant, roused/sleepy, dominant/submissive, credible/doubtful, interested/indifferent, and positive/negative. Grading terms are shown along both the axes and in both directions: 'very', 'fairly', and 'somewhat'. These scaling terms were used to minimise individual differences regarding the use of an evaluation window; without any verbal assistance, some annotators use only the centre area whereas others use only areas around the edges. An annotator moves the pointer on the evaluation window using a laptop touch pad to assign a score to the emotional state of the conversation segment they are observing. The values are recorded between -1 to 1 for each axis. The colour of the pointer changes when it deviates from the grid to inform the annotators about their irregular movements. When the pointer is outside the grid, the value is recorded either as -1 or 1 along the axis direction. In addition, there is a pause and resume functionality. Annotators can pause the video and its evaluation by clicking a button. This functionality has been introduced because dialogue videos are often longer than 20 minutes and annotators sometimes wish to rest during the process.

3. Emotional and Situational Axes

3.1. Progress/Recess

The psychotherapeutic dialogues help in finding solutions to the psychological problems faced by clients. Therefore, they are goal-oriented dialogues and it is important to know if an ongoing dialogue leads to the

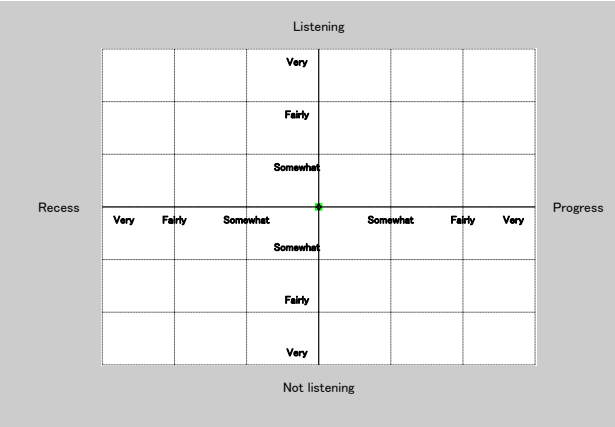

Figure 1: The input window for the EMO system.

solution. In psychotherapy, the evaluation of the outcome of dialogues is often used in practice. After a session, clients are asked to grade their mental situations on a certain scale. Then, these scores are compared with the pre-session scores. Although these session-wise evaluations are important for the psychotherapeutic outcome research, they are insufficient to examine the process of dialogues. Occasionally, participants hesitate to openly discuss their problems. In such situations, the therapists fail to understand the client's problems leading to a failure in counselling or creating an insufficient change in the participants. These irregularities are of particular interest to us, and we wish to assign the annotations of the degree of progression or recession during the psychotherapeutic interviews. Therefore, we have set the first axis as the degree of progression or recession.

3.2. Listening/Not listening

It is said that the relationship between therapists and clients is more important than the techniques used by the therapists during the interview sessions (Lambert and Barley, 2001). To establish a beneficial relationship, the feeling of being listened to is considered important. Therefore, we set the second axis as the degree of being listened to or not.

4. Dialogue Segmentation

4.1. Three Levels of Segmentation Clues

In segmenting dialogues, we can take the distinction between the semantic and signal clues into consideration. For example, spoken words and gesture occurrences were used for segmenting a multimodal dialogue (Takahashi and Inoue, 2014). Words are extracted from speech sounds and gestures are extracted from hand movements. During the discretization of signals, some information is lost. Moreover, extraction and labeling of semantic tokens is cumbersome for human annotators. The emotion and status annotations of dialogues, which are the focus of our annotation, often result in large cognitive load. For example, recall and verbalization are involved in the Interpersonal Process Recall method (Elliott, 1979). We intended to collect annotations using our continuous annotation tool without manual discretization effort but still usable as dialogue segmentation clues.

4.2. Change Point Detection

Segmenting dialogues into parts implies finding boundaries of the dialogues. Such boundaries are often topical or emotional change points of dialogues. Various techniques are available for change point detection. We are interested in the recession of problem-solving and its recovery from stagnation. Therefore, we calculate the extrema of timeseries, local minima in particular, by taking derivatives. That is, we consider that there are changes in the dialogue when the annotation values first decrease and then increase. Further, at that time, we can expect some actions to have occurred that changed the dialogue mood.

5. Annotation and Analysis

5.1. Data

We expanded our psychotherapeutic dialogue corpus with a new situation where graduate students majoring in psychotherapy were clients for a counselling session conducted by experienced therapists. In this situation, the counselling session was to be completed within one day, even if a reasonable solution was not found. We did not oversee the topic and the student clients talked about their actual problems. The interviews were not role-plays and we were able to witness conversations of emotionally depressed or confused participants. Participants agreed to the use of their dialogue data for research purposes through written consent forms.
The annotations were assigned by the therapists and clients for the same dialogue video. Although we show only annotations by the clients in this paper, this duplicated annotation was considered necessary for psychotherapeutic understanding. The feeling of being listened to by the clients is correlated with the outcome of therapy (Barrett-Lennard, 1962); whereas the feeling of listening to by the therapists does not always lead to better outcomes (Barrett-Lennard, 1981).

5.2. Annotation result

As a pilot study to check the annotation tool, we acquired two-dimensional annotations for 10 dialogues by respective clients focusing on both the therapists and clients. Out of these data, we used one dialogue and its annotation focussing on the therapist as an example. In the example dialogue, the client talked about his bad habit and the therapist worked with him to find methods to overcome the habit. The annotations for the dialogue are shown as a one-dimensional (progress/recess) timeseries as illustrated in the upper graphs of Figure 2 and another dimensional (listening/not-listening) timeseries as shown in Figure 3. The dots in these graphs indicate the measurement points. The position of the mouse pointer was recorded once every 0.1 second.

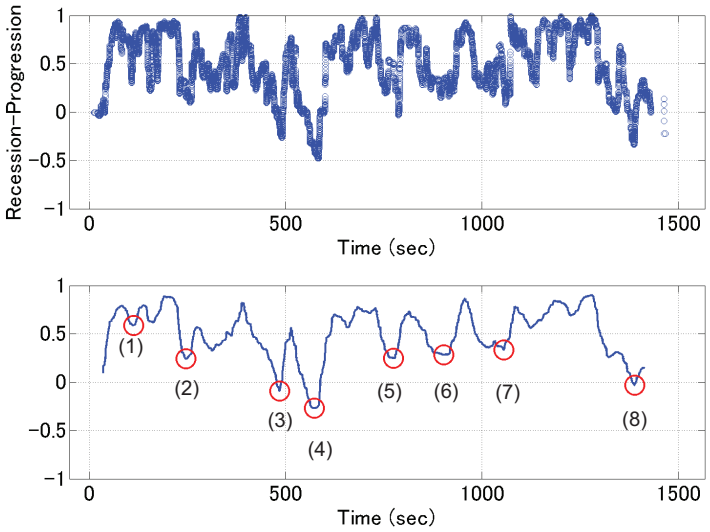

Figure 2: An example of acquired timeseries focusing on the therapist along the first axis. The numbers correspond to change points whose dialogue contents are shown in Table 1.

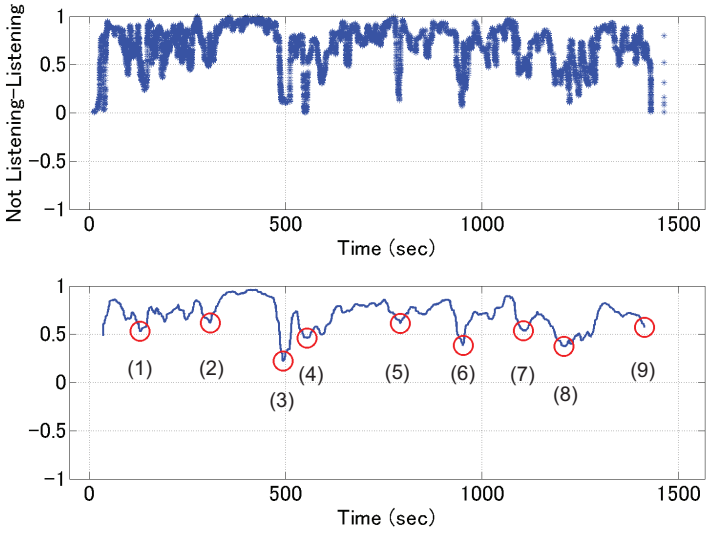

Figure 3: An example of acquired timeseries focusing on the therapist along the second axis. The numbers correspond to change points whose dialogue contents are shown in Table 2.

5.3. Change points

After applying a smoothing by taking an average of 25 points (2.5 seconds), we obtained the curves as shown in the lower graphs of Figures 2 and 3. Smoothing was employed to reduce the effects of fluctuations during the cursor movements. It reveals movement trends intended by the annotators. Then, the derivatives of the curves were calculated and the valleys deeper than a threshold value were extracted. The change points were plotted as circles in the lower graphs of Figure 2 and Figure 3. There were eight change points along the first axis (Figure 2) and nine change points along the second axis (Figure 3). The dialogue situations around the points were summarised in Table 1 and Table 2. Points common to both the axes were marked with asterisks subsequent to its indices. By comparing these points, we found the following patterns: the client felt that the therapist was not attending to him when the therapist was thinking and not talking to him or when the therapist followed an idea that the client was not happy with.

6. Conclusion

In this paper, we explained our procedure for adding dialogue emotion and status annotations to a multimodal dialogue video corpus to better understand con-

Table 1: Change points of progress/recess status when the client looked at the therapist.

Index	Time (min)	Description
1(*)	1.9	The client explained his problem to the therapist.
2	4.1	The therapist asked a question about the assumed state of the client.
3(*)	8.1	There were moments of silence and fillers because the therapist was unsure.
4(*)	9.6	The therapist was thinking intently and the client swayed his body while waiting.
5	12.9	The therapist summarised the client's behaviour but the client replied ambiguously.
6	15.0	The client explained the details of his bad behaviour and provided his own interpretation.
7	17.6	The client talked about a similar incident of his brother's bad behaviour.
8(*)	23.2	The therapist continued talking about the solution that the client did not accept.

Table 2: Change points of listening/not listening status when the client looked at the therapist.

Index	Time (min)	Description
1(*)	2.2	The client explained his problem to the therapist.
2	5.2	The client talked about a solution that he had not tried.
3(*)	8.2	There were moments of silence and fillers because the therapist was unsure.
4(*)	9.3	The therapist was thinking intently and the client swayed his body while waiting.
5	13.2	After the client explained his worst experience, the therapist proposed a new idea.
6	15.9	The client responded negatively to the therapist's idea.
7	18.4	The client suggested a possible solution but the therapist did not accept it.
8	20.2	The therapist let the client decide the action plan.
9(*)	23.6	The therapist continued to talk about the solution that the client did not accept.

versations that occur during psychotherapy. By using our continuous emotion and status-tracking tool, the annotators annotated the videos within the video time. The annotations were then utilised to find change points in the dialogues. The extrema of the annotation timeseries corresponded to the change points of the psychotherapeutic dialogues. In the future, we will analyse various dialogues annotated by different annotators to investigate individual differences. Moreover, the annotation will be combined with automatically recorded signal data such as head movements in our corpus.

Acknowledgements

This research was partially supported by Grants-in-Aid for Scientific Research 24500321, 23650111, and the Telecommunications Advancement Foundation.

7. References

G. T. Barrett-Lennard. 1962. Dimensions of therapist response as causal factors in therapeutic change. Psychological Monographs: General and Applied, 76(43):1–36.

G. T. Barrett-Lennard. 1981. The empathy cycle: refinement of a nuclear concept. Journal of Counseling Psychology, 28(2):91–100, March.

Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey, and Marc Schroder. 2000. FEELTRACE: An instrument for recording perceived emotion in real ttime. In ITRW on Speech and Emotion, Newcastle, UK.

Robert Elliott. 1979. How clients perceive helper behaviors. Journal of Counseling Psychology, 26(4):285.

Gijs Huisman, Marco van Hout, Elisabeth van Dijk, Thea van der Geest, and Dirk Heylen. 2013. Lemtool: Measuring emotions in visual interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 351–360.

Masashi Inoue, Ryoko Hanada, Nobuhiro Furuyama, Toshio Irino, Takako Ichinomiya, and Hiroyasu Massaki. 2012. Multimodal corpus for psychotherapeutic situation. In LREC Workshop on Multimodal Corpora for Machine Learning, pages 18–21, Istanbul, Turkey, May.

Paul Salvador Inventado, Roberto Legaspi, Masayuki Numao, and Merlin Suarez. 2011. Observatory: A tool for recording, annotating and reviewing emotion-related data. In Third International Conference on Knowledge and Systems Engineering, pages 261–265, Hanoi, Vietnam.

M. J. Lambert and D. E. Barley. 2001. Research summary on the therapeutic relationship and psychotherapy outcome. Psychotherapy: Theory, Research, Practice, Training, 38(4):357–361.

Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. 2012. Affectaura: An intelligent system for emotional memory. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 849–858.

Frederik Nagel, Reinhard Kopiez, Oliver Grewe, and Eckart Altenmuller. 2007. EMuJoy: Software for continuous measurement of perceived emotions in music. Behavior Research Methods, 39(2):283–290.

Kodai Takahashi and Masashi Inoue. 2014. Multimodal dialogue segmentation with gesture post-processing. In LREC 2014, Reykjavik, Iceland, May.

Towards a Multimodal Taxonomy of Dialogue Moves for Word-Guessing Games

Eli Pincus, David Traum

Institute for Creative Technologies, University of Southern California
12015 Waterfront Drive, Playa Vista, CA 90094
pincus@ict.usc.edu, traum@ict.usc.edu

Abstract

We develop a taxonomy for guesser and clue-giver dialogue moves in word guessing games. The taxonomy is designed to aid in the construction of a computational agent capable of participating in these games. We annotate the word guessing game of the multimodal Rapid Dialogue Game (RDG) corpus, RDG-Phrase, with this scheme. The scheme classifies clues, guesses, and other verbal actions as well as non-verbal actions such as gestures into different types. Cohen kappa inter-annotator agreement statistics for clue/non-clue and guess/non-guess are both approximately 76%, and the kappas for clue type and guess type are 59% and 75%, respectively. We discuss phenomena and challenges we encounter during annotation of the videos such as co-speech gestures, gesture disambiguation, and gesture discretization.

Keywords: rapid dialog, RDG-Phrase, clues, guesses, gesture disambiguation, gesture discretization

1. Introduction

In this work we develop a taxonomy of dialogue moves for team word guessing games in which one or more team members (called clue receivers) try to guess a target word or phrase known to the other partner (clue giver). The clue giver can use verbal or non-verbal means to elicit the target from the receiver. Generally, there are also restrictions on what the giver can say or do, which includes not saying (parts of) the target, but also might include other forbidden words or expressions. Variations of this game are popular as parlor games, card games, electronic games, and television game shows.

The taxonomy, presented in Section 2., seeks to capture strategies and typical behavior of both givers and receivers. This is done as a first step towards construction of computational agents capable of simulating human players of word-guessing games. To this end, we define categories for different types of clues, different delivery methods of clues, different types of guesses, as well as more generic actions such as hesitations. We also define several attributes that these actions can possess.

This taxonomy was used to annotate parts of the multimodal Rapid Dialogue Game (RDG) corpus (Paetzel et al., 2014). One of the games in this corpus, called RDG-Phrase, is a word-guessing game. This game has a single clue receiver, who is face to face contact with the clue giver. The clue giver has an opportunity to view, order, and prioritze the set of target words before each round. There is also a strict time limit, encouraging rapid interaction. Each pair in the corpus alternates rounds as clue giver and clue receiver. An interested reader should refer to Table 2 for a sample dialogue (with annotation) or for a longer sample dialogue (Paetzel et al., 2014).

2. Annotation Scheme

We divide actions that occur during word guessing game play into two categories according to role: clue giver or clue receiver. Both giver and receiver actions come in verbal and non-verbal form. Giver verbal actions are classified as either clues or non-clues. Receiver verbal actions are classified as either guesses or non-guesses. In order to address the multi-functionality nature inherent in utterances as discussed in (Bunt, 2010) , we use the "code high" approach (Condon and Cech, 1995) and specifiy a hierachy of tag types, so that lower priority tags are used only if no higher priority tags are used. Clues and guesses are higher priority than non-clue and non-guess. We further subdivide all of these categories by type. Clues are also associated with a delivery method attribute according to the structure of the sentence(s) utilized by the giver to deliver the clue to the receiver. Besides delivery method; we have defined several other attributes for verbal actions that we will define below. Non-verbal actions are broken into 7 categories: turn-taking, metaphoric, iconic, deictic, positive symbolic, negative symbolic, and other.

2.1. Giver Verbal Clues

There are 16 clue types defined below. Example instances can be found in Table 1. Each clue type is given a priority [A,B, or C], shown to in parentheses to the right of the type name, below. The "code high" principle is used to code clue types only from the highest category, in the case that more than one applies.

Analogy ***(A)*** clues set up a relationship between two entities and then attempt to elicit the receiver to recognize the same relationship between the target and another entity.
AssocAction ***(B)*** clues are utterances that describe what the target word does, what it is used for, or what uses it.
CitePast ***(B)*** clues reference previous turns or segments.
Contrast ***(A)*** clues supply a contrasting word or concept.
DescriptionDef ***(C)*** clues either describe or define the target word.
Disabuse ***(B)*** clues are meant to convey to the receiver that his guesses are off track.
Hypo ***(A)*** and **Hyper** ***(A)*** clues occur when the giver provides hyponyms or hypernyms of the target, respectively.
GeneralContext ***(A)*** clues cite knowledge that depends on aspects of the conversation situation (time, visible objects, etc.) concepts such as the current time, the current specific location, or the objects that are present in the room.
PartialPhrase ***(A)*** clues refer to instances where the giver states words that are commonly used with the target word or describe words that are commonly used with the target.
SemanticClass ***(A)*** clues are giver utterances that contain words with the same hypernym as the target word or utterances that request the receiver to say words with the same hypernym as the receiver's previous guess.
Synonyms ***(A)*** provide a synonym of the target word or a

Clue Type	Delivery Method	Target	Instance	Next Guess
Analogy	Complete	Night	"light versus dark but daytime"	Incorrect
Descr/Def	Fill-in-Blank	Alley	"The pathway behind a building is called a"	Correct
Contrast	Fragment	Video	"Not audio"	Correct
CitePast	Complete	Today	"you mentioned it before"	Incorrect
AssocAction	Complete	Doll House	"A place little girls play in"	Correct
Hyper	Fragment	Bus	"Public Transportation"	Correct
Hypo	Fragment	Gas Guzzler	"Cadillac"	Incorrect
SemanticClass	Fragment	Hour	"Um minute"	Incorrect
Partial-Phrase	Fill-in-Blank	Cabin	"Abraham Lincoln lives in a log"	Correct
Synonym	LeadingQue.	Main Street	"Whats another word for major"	Incorrect
Disabuse	Fragment	Electric	"nope" (after prius)	Incorrect
GeneralContext	Fragment	Today	"friday" (said on a Friday)	Incorrect
RequestSynonym	Complete	Hair Care	"another word for that" (after guess of "nurture")	Incorrect
Widen	Fragment	Hair Care	"more in general" (after guess of "washing hair")	Incorrect

Table 1: Example Clue Types

close approximation to a synonym of the target word.
RequestSynonyms ***(A)*** and **RequestAntonyms** ***(A)*** are clues where the giver directs the receiver to provide synonyms or antonyms, respectively, of words recently said.
Rhyming ***(A)*** clues have words that rhyme with the target.
Widen ***(A)*** clues ask the receiver to generalize what he is saying while **Narrow** ***(A)*** clues ask the receiver to state something more specifically.
Each clue has a **Delivery Method** that specifies the manner in which it is said. **Fill-in-Blank** clues are given as a sentence containing a missing word that is intended to be the target. Clues given in the form of a **LeadingQuestion** are expressed in the form of a question whose answer is supposed to be the target. A clue stated as a full sentence that does not fall into the other categories is considered **Complete** while a clue that is not a fully formed sentence and is not a Fill-in-Blank is a **Fragment**. If the delivery method of the clue is not clear, the clue's delivery method is said to be **None**. Refer to Table 1 for some example clues and their associated delivery methods.

2.2. Receiver Verbal Guesses

Receiver guess types are assigned to one of 6 categories. **Correct** guesses state the target word. **PartialCorrect** guesses contain the target within a larger word or phrase while **AbbreviatedCorrect** guesses state an abbreviated version of the target. **Partial** guesses are ones that state a part of the target but not the whole target. A guess is considered **Incorrect** if it contains no part of the target. Finally, guesses that are incomplete and therefore can not be unambiguously classified into one of the other categories are labeled as **None**. If a receiver utterance contained multiple guesses annotators marked the guess in the following order of priority: Correct, Partial Correct, Abbreviated Correct, Partial, Incorrect, None.

2.3. Non-Clues & Non-Guesses

Giver and receiver non-clue and non-guess actions have several categories in common. Both players can state an **Acknowledgement** indicating understanding of what the other player has said or a **Clarification** indicating that the player requires additional information about what was just said. Alternatively, either player can state a **Delay**, a filler utterance said while a player is thinking about his next action. The former three non-clue/non-guess types are instances of core dialogue dimensions discussed in (Bunt, 2010) as none of the types qualify as a RDG-Phrase dependent dialogue act. Acknowledgement and Clarification lie in the *Auto-Feedback* dimension and Delay has the communicative function *Stalling* in the *Time-Management* dimension. In addition, either player can utter an **Encouragement** in an attempt to boost the other player's morale or request to **Skip** to the next target. Either player can also **Evaluate** their performance by expressing thoughts on current game-play or emit **Laughter**. Evaluate, Skip, and Encouragement lie in the *Task* core dimension defined in (Bunt, 2010).
The giver can state a **Confirmation** in order to convey to the receiver that he has made a correct guess or partially correct guess. On the other side, the receiver may **Reject** by communicating his lack of knowledge of the target based on current information or **RequestRepeat** by asking the giver to repeat his last clue. Confirmation and RequestRepeat can be viewed as lying in Bunt's *Auto-Feedback* dimension while Confirmation can be viewed as lying in Bunt's *Task* dimension. Note that we only consider Laughter and Delay tags if none of the other tags seem appropriate.

2.4. Additional Verbal Attributes

We have also defined a number of attributes for clues and guesses. **Repeat** clues or guesses have already been used for the current target, **Incomplete** ones have been cut short, while clues or guesses assigned **ProsodyCompletion** are identified by their extended prosody. A **Multiple** guess is a receiver utterance composed of multiple guesses. Any clue labeled as **MultiWord** is a clue intended to elicit only part of the whole target from the receiver. **Recast** clues are clues that have adopted content words used by the receiver to guess the current target. Clues labeled with the **Clarification** attribute are ones that could not be understood without knowledge of previous clues. If the annotator feels that one clue spans either sequential giver utterances or giver utterances that are separated by Delay utterances or Laughter utterances only; then the blocks that span the clue are labeled **Partial** to indicate the multiple-block span nature of the clue. The delivery method attribute is then assigned to each of these blocks by considering all of the blocks as a single entity rather than assigning a delivery method attribute to each individual block. Table 2 shows a partial

Speaker	Utterance	Type	Attributes
Giver	"Not a large car but a"	Contrast	DM:Fill-in-Blank
Receiver	"Small car sedan"	Incorrect	Multiple
Giver	"Small"	Synonym	DM:Fragment;Recast
Receiver	"Small car"	Incorrect	Repeat
Giver	"Small car"	Hyper	DM:Fragment;Recast
Receiver	"Suburban [laughter] oh suburban"	Incorrect	-
Giver	"Sub"	PartialPhrase	DM:Fragment
Receiver	"Oh subcompact"	PartialCorrect	-
Receiver	"Right got you"	Acknowledgment	-

Table 2: Sample RDG-Phrase Dialogue with Target: Compact

transcription of a RDG-Phrase game, with annotations.

2.5. Non-Verbal

Initially, we divided non-verbal actions into 7 categories, loosely based on the categories of (McNeill, 1995), with a few specialized to timed guessing games: turn-taking, metaphoric, iconic, deictic, positive symbolic, negative symbolic, and other.

3. Annotation Method & Evaluation

3.1. Method

We utilize the multi-modal annotation tool Anvil (Kipp, 2012) to perform our annotation. Speech was segmented in the transcriptions of the RDG-Phrase videos if it was separated by 300 milliseconds of silence or more. We automatically convert these segmented utterances to instantiate utterance block elements in Anvil. Each speaker's utterance blocks are assigned their own "track" in Anvil. Each utterance block is labeled with its type in corresponding blocks in either the giver track or the receiver track and appropriate attributes selected.

3.2. Challenges

Several conversational phenomena arose during the course of our annotation. Co-speech gestures occurred frequently during game-play. We came across many verbal utterances whose semantic content was only clear when one considered the gesture the speech co-occurred with. For instance, in an attempt to elicit the target *playing cards* one giver pantomimed dealing cards while saying "I'm just gonna do this."

As pointed out by Susan Duncan[1], gestures are often multi-functional and segmentation can be particularly challenging as gestures repeat and blend into each other. For example, we frequently came across instances where the giver would utter an uninterrupted stream of clues of the same type synchronously with a rhythmic forward-backward hand extension. These gestures were unequivocally beat gestures but also appeared to serve a turn-taking cue function each time the giver's hand extended forward toward the receiver; seemingly to provide a chance for the receiver to interject with a guess. After initial attempts, we deferred non-verbal coding until we can suitably refine the annotation scheme to focus on those elements that are most crucial for game play.

3.3. Scheme Evaluation

We perform a small inter-annotator agreement study on four sequential seventy-second RDG-phrase rounds played by one pair (team), this includes 90 giver and 57 receiver utterances. Table 3 contains Cohen's Kappa statistics and absolute agreement statistics for each of the major verbal categories in our annotation scheme.

Category	Cohen's Kappa	Absolute
Clue/Non-Clue	76.18%	88.89%
Guess/Non-Guess	75.63%	89.47%
Giver Type	59.00%	64.44%
Receiver Type	74.96%	80.70%
Clue Delivery Method	53.00%	64.71%

Table 3: Inter-Annotator Agreement Statistics

The tags causing the most disagreement for utterances both annotators label as clue are DescriptionDef and AssocAction. This type of disagreement accounts for 3 out of the 10 or 30% of clue type disagreements. One example of this disagreement occurs with the giver utterance "yeah and then this one is on the ocean" where the target had been beach house and the receiver had just correctly guessed country house. This clue seems to fit in both categories as it describes the target like a DescriptionDef but in some sense it also answers the question: what is it used for? like an AssocAction. Instances such as this might lead us to further refine the definitions of these two categories for future annotation efforts.

The most common disagreement for the clue delivery method attribute occurs when one annotator feels the delivery method is not clear and therefore chooses the None value. This scenario accounts for 7 of the 18 tags that did not match; close to 40%. None of the other delivery method disagreements account for more than 3 of the delivery method tags that do not match.

4. Preliminary Annotation Results

The first author annotates all of the speech in 18 70-second RDG-phrase rounds played by three different pairs of people. The speech was segmented into 762 utterances according to our 300 milliseconds of silence criterion. 439 (58%) of the total utterances were said by the giver while 323 (42%) utterances were said by the receiver. See Table 4 for a further breakdown of these utterances.

4.1. Clues & Guesses

Figure 1 shows the relative frequency of Clue types. We find no instances of RequestAntonym or Rhyming clues in the annotated rounds and therefore these two types do not appear in Figure 1. The two most common clue types are AssocAction clues (28%) and DescriptionDef clues

[1] http://mcneilllab.uchicago.edu/pdfs/susan_duncan/Annotative_practice_REV-08.pdf

Giver Utt. Categ.	**# of Utt.** (% Giver Utt.)
Clues	247 (60%)
Non-Clues	162 (40%)
Rec. Utt. Categ.	**# of Utt.** (% Rec. Utt.)
Guesses	224 (69%)
Non-Guesses	99 (31%)

Table 4: Giver & Receiver Utterance Breakdown

(16%). One possibility is that this indicates that the definition of AssocAction captures important properties of the most common conceptual model for a noun or noun-phrase (all targets fall into one of these two syntactic categories). These statistics also imply that givers find word-relations (a category most of the other clue-types fall under) either more difficult to construct or consider them a less effective way of eliciting the target. We calculate a little less than

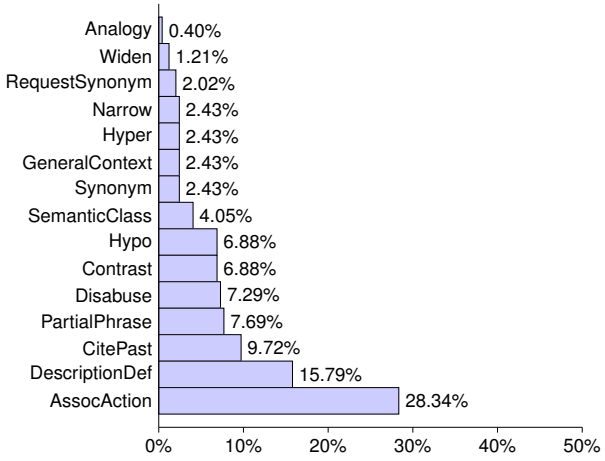

Figure 1: Clue Type Relative-Frequency

a quarter of the total guesses are correct (24%) and 55% contain at least part of the target or an abbreviated version of part of the target. More specifically, the breakdown of guesses are as follows: AbbreviatedCorrect (2.23%), PartialCorrect (2.68%), Correct (23.66%), Partial (26.79%), Incorrect (44.64%).

Clue Delivery Method Table 5 shows clue delivery statistics. The Fragment (39%) and Complete (28%) delivery methods were the most common for clues. This indicates that human givers find non-complete sentences the most efficient manner to deliver a clue and frequently consider structuring a grammatically correct sentence a task that does not contribute a significant amount of value. This also implies human givers use Fill-In-Blank and Leading Question delivery methods less often; possibly due to the time needed to construct clues in these forms.

Delivery Method	**# of Clues (%)**
LeadingQuestion	16 (7%)
None	21 (9%)
Fill-In-Blank	38 (17%)
Complete	64 (28%)
Fragment	89 (39%)

Table 5: Clue Delivery Method Statistics

4.2. Non-Clues and Non-Guesses

We tag 133 (17% of all utterances, 51% of Other Verbal utterances) utterances of either the giver or the receiver as Delay. 74 of these delays were said by the giver and 59 by the receiver. One third of all non-clues said by the giver were Confirmations. 18% of receiver's non-guesses were Acknowledgements. The other non-clue categories and the other non-guess categories each comprised a small relative percentage of all non-clue and non-guess utterances; 21% and 22% respectively. Further annotation and deeper investigation into these statistics should provide us data relevant to constructing a computational agent player that is able to perform behaviors such as backchannels, filled pauses, and turn-taking in a natural manner.

5. Conclusions

We present a taxonomy of dialogue moves for word-guessing games as a first step towards implementing a computational agent that can simulate a human player. Evaluation of our scheme yields reasonable inter-annotator reliability.

In future work, we intend to further refine our annotation scheme including providing guidelines for non-verbal annotation that minimize issues such as gesture disambiguation and gesture discretization. We will also continue our study of word-guessing game strategy by examining the relationship between prior clues and a current guess if the current guess is viewed as the current target. This investigation should also help determine how receivers interpret clues. We also have plans to implement a computational giver that is able to generate clue types such as Synonym, Contrast, Hyper, Hypo and DescriptionDef. We will accomplish this task by linking the giver to a database of word relations such as WordNet (Miller, 1995).

6. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1219253. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

7. References

Sherri L. Condon and Claude G. Cech. 1995. Problems for reliable discourse coding systems. In *AAAI Technical Report SS-95-06 Working Notes AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation.*, pages 27–33, March.

Harry Bunt et. al. 2010. Towards an iso standard for dialogue act annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), May.

Michael Kipp. 2012. Anvil: A universal video research tool. In J. Durand, U. Gut, and G. Kristofferson, editors, *Handbook of Corpus Phonology*. Oxford University Press.

David McNeill. 1995. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

Maike Paetzel, David Nicolas Racca, and David DeVault. 2014. A multimodal corpus of rapid dialogue games. In *Language Resources and Evaluation Conference (LREC)*, May.

Gaze Patterns in the Twente Debate Corpus

Bayu Rahayudi, Ronald Poppe, Dirk Heylen

University of Twente
PO Box 217, 7500AE, Enschede, Netherlands
e-mail: b.rahayudi@utwente.nl, r.w.poppe@utwente.nl, d.k.j.heylen@utwente.nl

Abstract

Different patterns of verbal and nonverbal behaviours have been associated with turn-taking in face-to-face conversations. Gaze is one that has been studied extensively. An important factor that determines the exact patterns in a particular conversation is the nature of the conversation; whether it is dyadic or multi-party, whether it is a chat or a heated debate, etcetera. In this paper we present a first analysis of the gaze patterns in the Twente Debate Corpus to investigate how the particular setting that was chosen influences the patterns in gaze behaviour. This analysis is meant to provide us with better insight in the features that are needed to improve automatic prediction algorithms such as those that predict the end of a turn.

Keywords: Gaze, End-of-turn, Multiparty conversation, Visual focus of attention

1. Introduction

Head movements and gaze patterns in face-to-face interaction have received a considerable amount of attention in the analysis of conversations. There are the classical analyses of face-to-face conversations, mainly targeting dyadic settings (Kendon, 1967; Argyle and Ingham, 1972) which show the various patterns and their conversational functions (see Heylen (2006) for a review and further analysis). One of the key roles that has been observed is that certain behaviours frequently occur - by nature or design - at particular places in a conversation such as turn switches. Kendon (1967) observed a clear pattern in gaze and head movements in this process. The speaker ended his utterance by looking at the listener, and the listener started his utterance by looking away from the speaker. It seems natural for a speaker to end the turn by looking at the listener to see whether the listener has understood what was said and is ready to respond. If this is so, then a speaker can refrain from looking at the listener by design to indicate the turn will not yet to be completed or to display ignorance of the signs the listeners wish to take the turn.

How the timings of gaze differ between speaker and listener is a continuous theme in the literature. In face-to-face conversations, it seems that listeners spend more time looking at the speaker than vice versa. Typically, as Bavelas et al. (2002) point out, listeners look at the speaker in uninterrupted intervals, whereas speakers briefly glance at the listener. There are often brief periods of mutual gaze, in which the listener produces a backchannel (nod or short vocalization) to signal continued attention, comprehension and interest. Oertel et al. (2012) analysed gaze patterns in dyadic conversations, and found distinctive gaze patterns associated with smooth speaker changes. In multiparty gaze analysis, Vertegaal et al. (2001) observed similar patterns and relative amounts of gaze towards speaker and listener. Ishii et al. (2013) analysed gaze transition patterns to predict the next speaker.

Recent technological advances have made possible the robust, automatic measurement of head position and orientation in 3D. This brings within reach the automatic analysis of turn-taking behaviours in face-to-face dialogs. In particular, we foresee applications in human-agent or humanrobot interaction, in which the agent or robot can predict when the human will end the turn, in order to facilitate smooth turn-taking behaviour. To this end, we need to understand how a persons head movement and gaze are correlated with stages of the turn-taking process such as starting and ending a turn[1]. The patterns that one can observe in conversations differ according to the setting, in this case a multiparty setting with differences in the make-up of the teams and the fact that it takes the form of a debate. We will show how these aspects are reflected in the gaze patterns, in particular during turn-taking.

Our data consists of three-person debates in which two participants act as a team against a single third participant. Both teams have to persuade the other of their opinion. Apart from the increased complexity, this allows us to look at differences between the behaviour of the participants in terms of whether they are a team consisting of a single person (referred to as single in the rest of the paper) and the team consisting of two persons (referred to as team). We will analyse a couple of features of the visual focus of attention (VFOA), the target of a persons gaze. In particular, we look at the amount of VFOA from and towards a speaker and the listeners, both within the same team and to the other. Ultimately we would like to use the information on such patterns in this type of behaviour to predict the next speaker, the end of the turn and see, for instance, how certain patterns might be indicative of the degree of persuasiveness of a participant.

2. Corpus

The corpus used in this research is the multimodal Twente Debate Corpus (Rahayudi et al., to appear). In addition to the audio-visual recordings, manually annotated VFOA and

[1] See also the series of workshops on Gaze and Human Computer Interaction, such as those recently held at `http://www.ci.seikei.ac.jp/nakano/GAZE_ICMI2012/` and `http://cs.joensuu.fi/~rbednari/GazeIn2013/`).

speaking/not speaking data are provided. The corpus consists of over 2 hours of debates, in 6 groups with 18 participants in total. Every group consists of three persons sitting around a table, arranged to face each other in the same angle, see Figure 1. Participants were recorded with a Kinect sensor, placed at the center of the table. The Microsoft Kinect SDK was used to determine the head position and orientation of each participant.

Figure 1: Picture from the Twente Debate Corpus

For each group, we recorded three sessions with a different participant as the single debater against a team of the other two. The participants had to defend their opinion on topics of which they had previously indicated whether they agreed or not. In addition to the analysis of VFOA in a multiparty setting, the corpus allows us to look into differences between the single and team debaters.

3. Data Analysis

In this section, we explore in more depth the gaze behaviours in this particular setting of three-party conversations, with a focus on gaze patterns at turn changes. We will also look in more depth at the differences between the single and the team participants which is a peculiar part of our setting. In Rahayudi et al. (to appear), we presented some basic numbers about the gaze of the participants (single and team) when they spoke and listened. When the single debater spoke, he would be looked at by the other participants (team) on average 74.15% time of the turn. When one of the team members spoke, he would be looked at by the other team member on average 61.01% of the time of a turn, and would be looked at by single on average 70.48% of a turn.

The total number of turns in the corpus is 627. Some of these turns are very short, for example short responses and short feedback or confirmation utterances. Therefore, we divided the turns into two categories, i.e. long turns for a turn longer than 3 seconds, and short turn otherwise. There are 458 long turns and 169 short turns in the corpus. Figure 2 shows the distribution of the length of the turns.

We analysed gaze in relation to the start and the end of turn. Table 1 shows some aspects of gaze in relation to the start-of-turn and end-of-turn. From Table 1, we observe that the speaker receives around 76.7% gaze from listeners at the end of a turn. We also can see that the speaker is looked at

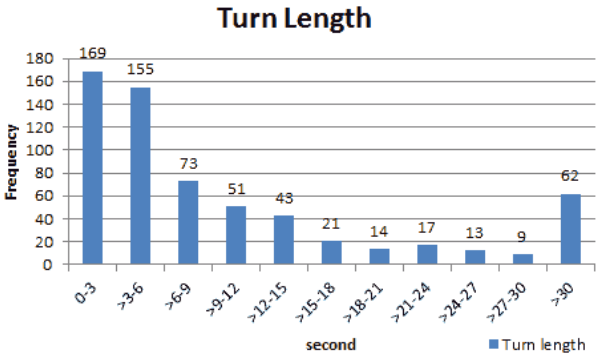

Figure 2: Distribution of turn lengths

by the next speaker at the end of the turn 52.8% of the time, and that this probability is somewhat lower for short turns. This table support the intuition that the next speaker looks at the current speaker when he wants to take the turn. It is also shown by the mutual gaze numbers at the end or at the start of the turn, respectively 56.8% and 45.1%.

	Short	**Long**	**All**
Speaker looked by listener at end-of-turn	79.3%	75.7%	76.7%
Speaker looked by next speaker at end-of-turn	46.7%	55%	52.8%
Mutual gaze at the end-of-turn	59.7%	55.7%	56.8%
Mutual gaze at the start-of-turn	43.8%	45.6%	45.1%

Table 1: Gaze in relation to start-of-turn and end-of-turn

To see how the amount of gaze to the speaker varied within the turn, we divided every turn into 20 equal periods (each corresponding to 5% turn progress) and calculated the average amount of gaze received by the speaker from the listener in each period. The graphs for short, long and all turns are shown in Figure 3. We can see from this figure that, at the beginning and end of the turn, the speaker receives less gaze from the listeners. This effect is stronger for longer turns, in which the speaker receives approximately 85% gaze from the listeners. In comparison, the speaker receives around 70% in turns shorter than 3 seconds. It shows that listeners tend to look at the speaker especially in the middle of turn.

Let us next compare the differences of the gaze behaviour of speakers and listeners when the speaker is either part of the team, or the single debater. Figure 4 shows the percentages of gaze when the speaker is a single speaker. The percentages from each participant do not add up to 100% as each participant can also be looking at other targets such as the table. These numbers also differ slightly from those reported by Rahayudi et al. (to appear) as regions with speech overlap are here also taken into account. A speaker is thus a person who holds the turn, and the listeners are in this case the other two.

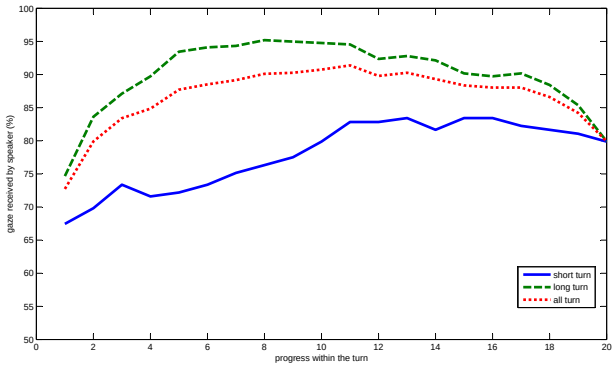

Figure 3: Amount of gaze received by the speaker during the course of the turn

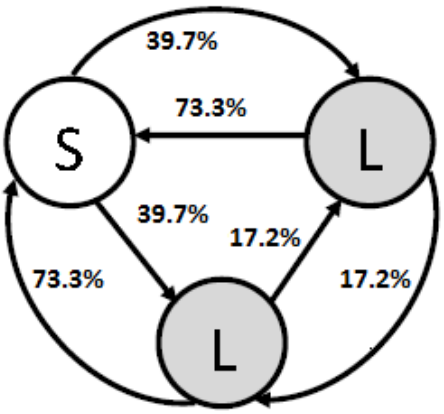

Figure 4: Percentages of gaze for speaker (S) and listener (L), when the speaker is the single debater. Shaded circles correspond to team members

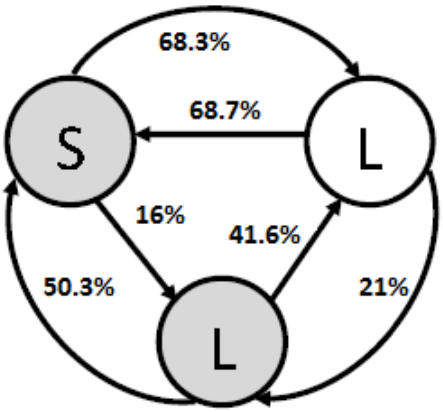

Figure 5: Percentages of gaze for speaker (S) and listener (L), when the speaker is a team debater. Shaded circles correspond to team members.

	Gaze(s)
Single S → Team L	5.5
Team L → Single S	10.0
Team L → Team L	4.7
Team S → Single L	8.0
Team S → Team L	3.5
Single L → Team S	9.6
Single L → Team L	5.6
Team L → Team S	6.1
Team L → Single L	5.4

Table 2: Average length of gaze period for each participant to the other, in seconds

What can be seen here is that the single speaker divides his attention over both listeners and the listeners that form a team mainly look at the speaker and only in a limited percentage of the time at each other. Now what happens if the speaker is in a team? Figure 5 shows the various percentages of gaze. Whereas the single speaker in Figure 4 divides his attention between the two team listeners, the team speaker is looking mainly at the single listener and only for a limited amount to his team mate.

The team mate listener, on the other hand, divides his attention between his team mate speaker and the single listener. Why is this? One can explain this by the fact that the team listener is not only interested in what his team mate has to say, but also very much in what the opposing single listener thinks of this.

In addition to the amount of gaze to each participant in the debate, we also analysed the length of the gaze intervals to the other participants. In line with Bavelas et al. (2002), we expected fairly long periods of gaze from the listener to the speaker, and relatively brief moments of gaze from the speaker to the listener. Our data indeed confirms this hypothesis. Table 2 summarizes the average gaze length for different pairs of participants.

From Table 2, it becomes clear that, when the single debater is the speaker, the average length of the periods of gaze towards him is twice that of the gaze from this speaker. When the single debater is not the speaker, however, the periods of gaze from the speaker to the single listener are much longer (8.0s). Another interesting finding is that the listeners periods of gaze towards to the speaker in the same team are much shorter (6.1s, compared to 10.0 when the speaker is in the other team). These numbers again show that participants in the debate closely monitor the participants in the other team, and that these numbers bias the common speaker-listener statistics to a large degree.

Finally, we show four typical frequency histograms to emphasise the previous statements. We present the gaze periods from the single speaker towards a team listener (Figure 6), from the team listener towards the single speaker (Figure 7), from the team speaker towards the single listener (Figure 8), and from the team speaker towards the team listener (Figure 9). In Figure 6, there is a clear peak in the very brief (shorter than 1 second) gazes from single speaker to team listener. It validates the statement that a single speaker tends to look at team listener frequently but briefly. Figure 7 shows that a team listener looks at a single speaker for longer periods of time. On the other hand, when one of the team became speaker, then the team speaker tended to

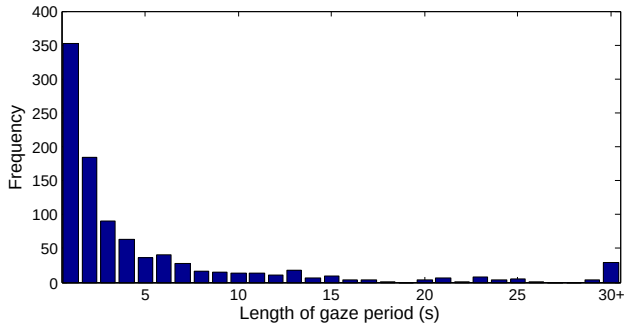

Figure 6: Frequency histogram for gaze period length from single speaker to team listener.

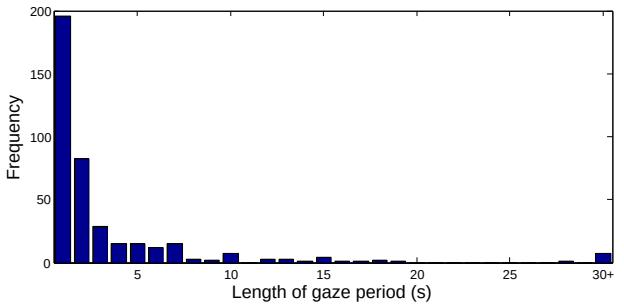

Figure 9: Frequency histogram for gaze period length from team speaker to team listener.

look at single listener in longer time compared to his team listener (Figure 8 and 9).

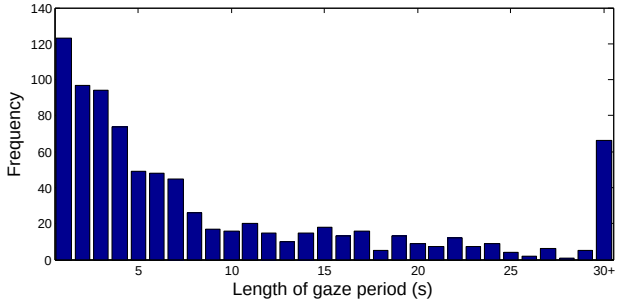

Figure 7: Frequency histogram for gaze period length from team listener to single speaker.

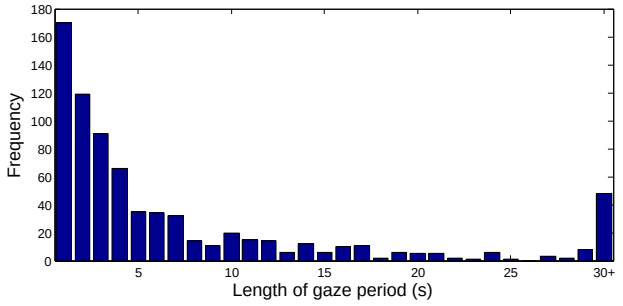

Figure 8: Frequency histogram for gaze period length from team speaker to single listener.

4. Conclusion and Future Work

The above analysis of the gaze patterns in our corpus shows that the nature of the conversation determines to a huge extent who will be looking at whom for how long. It is obvious that these aspects need to be taken into account when building classifiers of conversational structure and predictors for who will take the next turn if gaze will be taken in as one of the features. In our future work, we will address this issue from a computational point of view.

5. Acknowledgements

This research has been supported by the EU's FP7 Program SSPNet. The first author also would like to thank the Directorate General for Higher Education, Indonesian Ministry of Education and Culture (DIKTI) for the scholarship that allows him to take the PhD program at the University of Twente.

6. References

M. Argyle and R. Ingham. 1972. Gaze, mutual gaze, and proximity. *Semiotica*, 6(1):32–49.

J.B. Bavelas, L. Coates, and T. Johnson. 2002. Listener Responses as a Collaborative Process: The Role of Gaze. *Journal of Communication*, 52(3):566–580, September.

D. Heylen. 2006. Head gestures, gaze and the principles of conversational structure. *International Journal of Humanoid Robotics*, 3(3):241–267, September.

R. Ishii, K. Otsuka, S. Kumano, M. Matsuda, and J. Yamato. 2013. Predicting next speaker and timing from gaze transition patterns in multi-party meetings. *Proceedings of the 15th ACM on International conference on multimodal interaction - ICMI '13*, pages 79–86.

A. Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica*, pages 22–63.

C. Oertel, M. Wlodarczak, J. Edlund, P. Wagner, and J. Gustafson. 2012. Gaze patterns in turn-taking. Proceedings of Interspeech 2012.

B. Rahayudi, R. Poppe, and D. Heylen. to appear. Twente Debate Corpus A Multimodal Corpus for Head Movement Analysis. to appear.

R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt. 2001. Eye Gaze Patterns in Conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 301–308.

EmoGest: Investigating the Impact of Emotions on Spontaneous Co-speech Gestures

Kirsten Bergmann[1], Ronald Böck[2], Petra Jaecks[3]
[1]Faculty of Technology, Bielefeld University, P.O. Box 100 131, 33501 Bielefeld, Germany
[2]Cognitive Systems Group, Otto von Guericke University Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany
[3]Faculty of Linguistics and Literary Studies, Bielefeld University, P.O. Box 100 131, 33501 Bielefeld, Germany
kirsten.bergmann@uni-bielefeld.de, ronald.boeck@ovgu.de, petra.jaecks@uni-bielefeld.de

Abstract

Spontaneous co-speech gestures are an integral part of human communicative behavior. Little is known, however, about how they reflect a speaker's emotional state. In this paper, we describe the setup of a novel body movement database. 32 participants were primed with emotions (happy, sad, neutral) by listening to selected music pieces and, subsequently, fulfilled a gesture-eliciting task. We present our methodology of evaluating the effects of emotion priming with standardized questionnaires, and via automatic emotion recognition of the speech signal. First results suggest that emotional priming was successful, thus, paving the way for further analyses comparing the gestural behavior across the three experimental conditions.

Keywords: Emotions, priming, multimodal, co-speech gestures, corpus collection, audio feature analysis

1. Introduction

There is a large body of empirical evidence demonstrating that emotional states manifest themselves in different aspects of communicative behavior. For speech, research has demonstrated various effects in terms of acoustic features such as loudness, speaking rate, intonation, voice quality etc., as well as lexical choice, use of syntactic construction etc. (see, e.g., Bänziger et al. (2014)). Likewise, facial expressions have been studied extensively as a major medium of expressing emotions (see, e.g., Keltner et al. (2003), Russell et al. (2003)). In addition, there is a substantial amount of evidence demonstrating that particular body postures are associated with a specific mood or attitude (e.g., Crane and Gross (2013), Dael et al. (2012)).

Despite all this, what we know about the impact of particular emotional states on *co-speech gestures* is still sparse. Existing corpora like the *Belfast database* (Douglas-Cowie et al., 2000), the *EmoTV* corpus (Abrilian et al., 2005), or the *GEMEP* corpus (Bänziger et al., 2010) do not focus on co-speech gestures in detail. There are, however, a few studies which have begun to address the impact of emotional states on speech-accompanying gestures. Castellano et al. (2007) conducted a study in which participants performed one and the same gesture while expressing different emotional conditions. An approach of automated video analysis has been employed to investigate whether expressive motion cues, such as movement amplitude or speed/fluidity of movement, allow to discriminate between emotions. Results showed that expressive motion cues allow to discriminate between high and low arousal emotions as well as positive and negative emotions. Kipp and Martin (2009) investigated how basic gesture form features (handedness, hand shape, palm orientation, motion direction) are related to components of emotion. The analysis was based on a corpus of segments from two versions of a movie in which the protagonist displays a wide range of emotions. The analysis revealed that handedness in gestures is closely correlated with emotion categories. A positive correlation was demonstrated for high pleasure and left-handed gestures, while right-handed gestures were more likely to occur when low pleasure was expressed. With a similar approach, Fourati and Pelachaud (2013) recently set up a larger database of acted emotional body behavior. 3D motion capture data synchronized with full HD video was recorded from 11 actors who expressed different emotional states while describing several actions. In advance, the actors had gone through a training to express emotions in daily actions while avoiding exaggerated and expressiveless behavior.

The present corpus collection aims to advance this previous work by providing detailed data on the interrelation of emotions and co-speech gestures in spontaneous face-to-face interaction. While the aforementioned studies took important steps in providing first data and evidence that different aspects of gesture use are affected by the speakers' emotional state, they are limited to *acted* emotional states. The question, therefore, remains whether and how *spontaneous* speech-accompanying gestures reflect the speaker's emotional state. Likewise, in the community of speech-based emotion recognition, there is a recent trend towards naturalistic data sets which represent spontaneous emotional reactions (see, e.g., Schuller et al. (2011)).

In this paper, we describe the setup of a novel database of spontaneous co-speech body movement behavior, the *EmoGest corpus*. Participants were primed with emotions by listening to selected music pieces – rather than instructed to express particular emotions – and subsequently fulfilled a gesture-eliciting task. In the following we will first sketch the study setup. Then, we put a focus on our methodology and first results of evaluating the effects of emotional priming in terms of (a) participants' self-ratings with standardized questionnaires as well as (b) automatic emotion recognition of the speech signal. We conclude with a prospect of gesture coding techniques intended to complement the corpus data.

2. Experimental Setup and Data Collection

The corpus was set up based on a linguistic experiment. 32 participants interacted naturally in a tangram task, where they had to describe 12 tangram figures to a confederate interaction partner. Prior to the tangram task, all participants listened to one of three audio files of about three minutes length each presenting classical musical pieces that induce different emotions (happiness, sadness, neutral). The happy and sad stimuli were collected and published by Eerola and Vuoskoski (2011). The items of their "Soundtracks datasets for music and emotion" were evaluated for their power to induce emotions (see Eerola and Vuoskoski (2011) for statistics). The neutral stimuli were generated according to the description and statistics by Hunter et al. (2008). After participants were provided with the music stimuli, they completed self-rating questionnaires to evaluate the priming effect of the musical emotion induction. Subsequently, they listened to the same music stimulus once again before they fulfilled the tangram description task in interaction with a confederate.

The primary data of the corpus consists of audio and HD video recordings of the interactions as well as Kinect data. For the videotape three synchronized camera views were recorded (see Fig. 1). In total, the corpus consists of ∼12 hours of dialogical interaction and contains ∼4.000 representative gestures (projected from first gesture segmentations of ∼25% of the material). The three experimental groups were comparable in handedness according to the Edinburgh handedness inventory ((Oldfield, 1971); 27 right, 4 left, 1 ambidextrous; χ^2=2.651, p=0.618) and gender distribution (χ^2=3.269, p=0.195). They did not differ in age (20-41 years, χ^2=2.327, p=0.312) or years of education (13-25 years, χ^2=1.420, p=0.492).

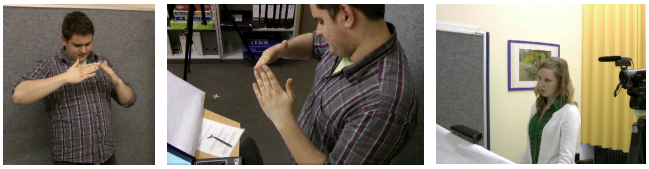

Figure 1: Experimental dialogue situation from three camera views, capturing a participant who describes a stimulus tangram figure displayed on a laptop (left and middle), and the confederate (right).

Several personality questionnaires were conducted (prior to the main experiment). There were no significant differences in personality traits across the three groups (BFI-K, Rammstedt and John (2005); e.g. extraversion: χ^2=4.409, p=0.110), actual mood (UWIST, Matthews et al. (1990); χ^2=0.384, p=0.825) or empathy (SPF/IRI, Paulus (2009); χ^2=0.670, p=0.715).

3. Evaluation of Emotional Priming

3.1. Self-ratings of Emotional State

To evaluate the priming effect of the musical emotion induction, two different scales were applied. After listening to the music, the groups differed in their feelings of 'joyful activation', 'wonder', 'power', 'tension', 'sadness' (GEM Scales, (Zentner et al., 2008)) and valence and activity (dimensional model, Eerola and Vuoskoski (2011)). For example, 'joyful activation' is rated significantly higher in the 'happy' condition (χ^2=16.474, p<.001) providing evidence for a relevant emotional priming effect. Therefore, we argue that it is scientifically sound to compare the three condition groups in further analyses.

3.2. Analysis of Audio Features

To complement the results from participants' self estimation of their emotional state, we employ an automated analysis of acoustic features. In the field of speech data-based emotion recognition two categories of features are widely used, namely spectral and prosodic features. Most speech recognition systems rely on spectral features sets which are based on Mel Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coefficients (LPCs), and Perceptual Linear Predictive coefficients (PLPs). Various researchers showed that these features are also suitable to recognize emotions from speech (cf. Böck et al. (2010), Schuller et al. (2010), Ververidis and Kotropoulos (2006), Vogt and André (2005)). On the other hand, prosodic features like pitch, intensity, voice quality and vocal tract features provide additional information for the classification of emotional speech. Vocal tract features like formants, their bandwidths etc. reflect characteristics of the speaker whereas voice quality features (jitter, shimmer, etc.) characterize the current articulation. Reviews on prosodic features are given in Schuller et al. (2011), Ververidis and Kotropoulos (2006). The most important issue in feature selection is the identification of meaningful features that represent the characteristics of the speaker and the current situation. Especially, in the context of naturalistic interactions existing and well-known feature sets have to be re-evaluated.

The EmoGest corpus provides a naturalistic Human-Human Interaction (HHI) of two partners, the participant and a confederate. The participant was primed to be in a certain emotional state, namely happy or sad (or neutral as a control). To evaluate the priming from a speech perspective we concentrate on the two emotions which can be also classified as positive and negative. From these considerations and based on previous work (Böck et al., 2012), we selected features which will be in the focus of future research: the first to third formant and their corresponding bandwidth, pitch, jitter, and intensity (Scherer, 2001; Vlasenko et al., 2011) are potentially meaningful since these are related to negative as well as high aroused emotions (cf. De Looze et al. (2011), Schuller et al. (2010)).

The feature extraction is conducted on a level of utterances. To extract the features we applied PRAAT (cf. Boersma (2001)) for prosodic features and the Hidden Markov Toolkit (HTK) (cf. Young et al. (2009)) for MFCCs and combined them afterwards. In preliminary tests such a procedure is advisable since the combination of features can be handled more easily.

3.2.1. Classifiers

In the community of emotion recognition from speech several types of classifiers are used whereas Support Vector Machines and Hidden Markov Models (HMMs) are most prominent. HMMs are utilized in the classification of emotional speech (cf. e.g. El Ayadi et al. (2011), Schuller et al. (2011)). In general, each HMM is a finite state automata which passes from state s_i to state s_j in each time slot. While traversing the model a sequence of observations is produced given a certain probability density. Given a set of trained HMMs the most likely sequence of observations is calculated by the Viterbi algorithm. Afterwards, the model providing the highest log-likelihood is selected as the classification result. Further technical details are given in El Ayadi et al. (2011), Young et al. (2009).

Since we are dealing with a multi-modal corpus we have the opportunity to investigate single modalities in the context of naturalistic HHI and further, to combine various modalities. This leads to the issue of fusion. According to (Krell et al., 2013) we suggest a two step classification process. For each modality features are extracted separately and afterwards, are used to achieve a first classification results. This will be finally combined with those results gained by applying the other modalities. To handle gaps in the input sequence of the final classifier., that means, information is partially not available, a suitable combination method has to be identified. As discussed by Krell et al. (2013), Markov Fusion Networks can be a potential solution.

3.2.2. Preliminary results

An automatic emotion recognition from speech was conducted applying HMMs and the feature set described above in a 10-fold-cross-validation. Based on a subset of the data we achieved an unweighted average accuracy of 90.8% in a two class investigation given by the experimental design ('happy' vs. 'sad'). In line with our results from participants' self-rating of their emotional state, these results indicate that the emotional priming was successful and that the speakers' emotional state can be automatically distinguished in speech. As up to now, not all participants of the experiment were processed to enable automatic classification, the presented results do not have high significance, yet. The preliminary study was implemented to verify if the priming could be seen also in emotionally colored speech.

4. Conclusion

Our goal is to provide a corpus which allows to address whether and how *spontaneous* co-speech gesture use in terms of gesture rate, gesture types, physical gesture form, and gesture expressivity (cf. Hartmann et al. (2006)) is affected by emotional states of the speaker. In this paper, we described the experimental setup of the corpus collection and focused on evaluations of the applied emotional priming. First results are promising so that we now continue to set up the full corpus. The audio signal-based evaluation will be continued and further complemented with an observer-based rating of speakers' emotional state. In addition, we will continue to generate secondary data, particularly focusing on speakers' gestural behavior. To this end, we will apply a feature-based coding of physical gesture form as already applied in the SaGA corpus (Lücking et al., 2013) complemented with annotations according to the NEUROGES coding system (Lausberg, 2013). We will further apply automated coding techniques based on Kinect data, e.g., the MINT.tools (Kousidis et al., 2013), or the NovA for social signal analyses (Baur et al., 2013). These codings will enable us to conduct detailed analyses of how spontaneous co-speech gesture use is affected by emotional states, as well as detailed inter-modal analyses of linguistic content, speech, and gestures in emotionally primed speakers.

5. Acknowledgement

We acknowledge support by the Collaborative Research Centre SFB 673 "Alignment in Communication" and the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems", both funded by the German Research Foundation (DFG).

6. References

Abrilian, S., Devillers, S., Buisine, S., and Martin, J.-C. (2005). EmoTV1: Annotations of real-life emotions for the specification of multimodal affective interfaces. In *Human Computer Interaction International*.

Bänziger, T., Scherer, and R., K. (2010). Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) corpus. In Scherer, K. R., Bänziger, T., and Roesch, E. B., editors, *Blueprint for affective computing: A sourcebook*, pages 271–294. Oxford University Press, Oxford, UK.

Bänziger, T., Patel, S., and Scherer, K. (2014). The role of perceived voice and speech characteristics in vocal emotion communication. *Journal of Nonverbal Behavior*, 38(31–52).

Baur, T., Damian, I., Lingenfelser, F., Wagner, J., and André, E. (2013). Nova: Automated analysis of nonverbal signals in social interactions. In Salah, A. A., Hung, H., Aran, O., and Gunes, H., editors, *Human Behavior Understanding*. Springer, Berlin/Heidelberg.

Böck, R., Hübner, D., and Wendemuth, A. (2010). Determining optimal signal features and parameters for hmm-based emotion classification. In *Proceedings of the 15th IEEE Mediterranean Electrotechnical Conference*, pages 1586–1590, Valletta, Malta. IEEE.

Böck, R., Limbrecht, K., Siegert, I., Glüge, S., Walter, S., and Wendemuth, A. (2012). Combining mimic and prosodic analyses for user disposition classification. In Wolff, M., editor, *Proceedings of the 23. Konferenz Elektronische Sprachsignalverarbeitung*, pages 220–228, Cottbus, Germany.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Castellano, G., Villalba, S., and Camurri, A. (2007). Recognizing human emotions from body movement and gesture dynamics. In Paiva, A., Prada, R., and Picard, R., editors, *Affective Computing and Intelligent Interaction*, LNAI 4738, pages 71–82. Springer, Berlin/Heidelberg.

Crane, E. and Gross, M. (2013). Effort-shape characteristics of emotion-related body movement. *Journal of Nonverbal Behavior*, 37(2):91–105.

Dael, N., Mortillaro, M., and Scherer, K. R. (2012). Emotion expression in body action and posture. *Emotion*, 12(5):1085–1101.

De Looze, C., Oertel, C., Rauzy, S., and Campbell, N. (2011). Measuring dynamics of mimicry by means of prosodic cues in conversational speech. In *17th International Congress of Phonetic Sciences*, Hong Kong, China.

Douglas-Cowie, E., Cowie, R., and Schröder, M. (2000). A new emotion database: Considerations, sources and scope. In *Proceedings of the ISCA Workshop on Speech and Emotion*.

Eerola, T. and Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39:18–49.

El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572 – 587.

Fourati, N. and Pelachaud, C. (2013). A new emotional body behavior database. In Edlund, J., Heylen, D., and Paggio, P., editors, *Proceedings of the Workshop on Multimodal Corpora 2013: Multimodal Corpora: Beyond Audio and Video*.

Hartmann, B., Mancini, M., and Pelachaud, C. (2006). Implementing expressive gesture synthesis for embodied conversational agents. In Gibet, S., Courty, N., and Kamp, J.-F., editors, *Gesture in Human-Computer Interaction and Simulation*, pages 45–55. Springer, Berlin/Heidelberg.

Hunter, P. G., Schellenberg, E. G., and Schimmack, U. (2008). Mixed affective responses to music with conflicting cues. *Cognition & Emotion*, 22(2):327–352.

Keltner, D., Ekman, P., Gonzaga, G., and Beer, J. (2003). Facial expression of emotion. In *Handbook of affective sciences*, pages 415–532. Oxford University Press, New York, NY, US.

Kipp, M. and Martin, J.-C. (2009). Gesture and emotion: Can basic gestural form features discrminate emotions? In Cohn, J., Nijholt, A., and Pantic, M., editors, *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII-09)*. IEEE Press.

Kousidis, S., Pfeiffer, T., and Schlangen, D. (2013). MINT.tools: Tools and adaptors supporting acquisition, annotation and analysis of multimodal corpora. In *Proceedings of Interspeech 2013*.

Krell, G., Glodek, M., Panning, A., Siegert, I., Michaelis, B., Wendemuth, A., and Schwenker, F. (2013). Fusion of fragmentary classifier decisions for affective state recognition. In *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*, volume 7742 of *LNAI*, pages 116–130. Springer Berlin Heidelberg.

Lausberg, H., editor. (2013). *Understanding body movement: A guide to empirical research on nonverbal behavior: with an introduction to the NEUROGES coding system*. PL Academic Research, Frankfurt a.M.

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., and Rieser, H. (2013). The bielefeld speech and gesture alignment corpus (SaGA). *Journal on Multimodal User Interfaces*, 7(1-2):5–18.

Matthews, G., Jones, D., and Chamberlain, A. (1990). Refining the measurement of mood: The uwist mood adjective checklist. *British Journal of Psychology*, 81:17—42.

Oldfield, R. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 9:97–113.

Paulus, C. (2009). Der Saarbrücker Persönlichkeitsfragebogen (IRI) zur Messung von Empathie. Psychometrische Evaluation der deutschen Version des Interpersonal Reactivity Index. [The Saarbrücken personality questionnaire (IRI) for measuring empathy: A psychometric evaluation of the German version of the interpersonal reactivity index].

Rammstedt, B. and John, O. P. (2005). Kurzversion des big five inventory (bfi-k): Entwicklung und validierung eines ökonomischen inventars zur erfassung der fünf faktoren der persönlichkeit. *Diagnostika*, 51:195–206.

Russell, J. A., Bachorowski, J., and Fernández-Dols, J. (2003). Facial and vocal expressions of emotion. *Annual Review of Psychology*, 54:329–349.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research,*, pages 92–120.

Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., and Rigoll, G. (2010). Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131.

Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10):1062–1087.

Ververidis, D. and Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181.

Vlasenko, B., Philippou-Hübner, D., Prylipko, D., Böck, R., Siegert, I., and Wendemuth, A. (2011). Vowels formants analysis allows straightforward detection of high arousal emotions. In *2011 IEEE International Conference on Multimedia and Expo (ICME)*, Barcelona, Spain.

Vogt, T. and André, E. (2005). Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *IEEE International Conference on Multimedia and Expo 2005*, pages 474–477, Amsterdam, The Netherlands. IEEE.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2009). *The HTK Book, version 3.4*. Cambridge University Engineering Department.

Zentner, M., Grandjean, D., and Scherer, K. (2008). Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion*, 8(4):494—521.

The "Vampire King" (Version 2) Corpus

Ivan Gris, David Novick, Mario Gutierrez, Diego A. Rivera
The University of Texas at El Paso
500 W. University Ave., El Paso, TX 79912 USA
E-mail: ivangris4@gmail.com, novick@utep.edu, mgutierrez19@miners.utep.edu, darivera2@miners.utep.edu

Abstract

As part of a study examining nonverbal and paralinguistic behaviors in conversations between humans and embodied conversational agents (ECAs), we collected a corpus of human subjects interacting with an ECA in an adventure game. In the interaction, the ECA served as a narrator for a game entitled "Escape from the Castle of the Vampire King," which was inspired by text-based computer games such as Zork. The corpus described here is based on Version 2 of the game, in which a map of the castle was displayed on the wall behind the ECA. The system was not a Wizard-of-Oz simulation; the system responded using speech recognition and utterance generation. The corpus includes 20 subjects, each of whom interacted with the game for 30-minute sessions on two consecutive days, for a total of approximately 1200 minutes of interaction. All 40 sessions were both audiovisually recorded and automatically annotated for speech and basic posture using a Kinect sensor. The corpus includes (a) the automated annotations for speech and posture and (b) manual annotations for gaze, nods and interrupts.

Keywords: multimodal, real conversation, rapport

1. Introduction

This paper reports the collection of a corpus of interactions between humans and an embodied conversational agent (ECA). We developed the corpus to support a study of human-ECA rapport.

One of the main goals of researchers on real-time interaction with ECAs is to strive for increased realism in agents' behavior. One issues is maintaining and adapting to long-term interaction, particularly with respect to rapport. In our view (Novick & Gris, in press), paralinguistic rapport comprises three dimensions: a sense of emotional connection, a sense of mutual understanding, and a sense of physical connection. Because our research focuses on the physical dimension, the corpus was aimed at understanding the results of using an agent with different nonverbal behaviors (familiar and non-familiar). Studies of human-human dialog have suggested that people signal increased familiarity by, among other things, increasing the amplitude of nonverbal communicative behaviors such as hand gestures and head nods (Neff et al., 2010; Cafaro et al., 2012; Clausen-Bruun, Ek, & Haake, 2013). Thus in our system the agent communicated increased familiarity by increasing the amplitude of its gestures.

Because our underlying research on the development of human-ECA rapport depended on having subjects engage in multiple sessions over time, we needed to provide an interaction experience that was highly engaging; participants should want to return for later sessions. Toward this end, we developed an adventure game based on text games such as Zork (Anderson & Galley, 1985) or Colossal Cave (Crowther, Woods & Black, 1976) that follows the same gameplay format. In our game, the user tries to escape from the castle of an evil vampire king. In the place of the traditional console text interface, though, our agent narrated the game situation to the user. Players verbally indicated how they want to react to the presented situation, either by moving to different rooms, picking up items, using items, examining and uncovering secret passages or fighting the vampires.

In this paper we describe our automated and semi-automated annotation systems and our corpus-collection techniques. We provide examples of both verbal and nonverbal annotations generated by our system as well as how we use this functionality to recreate past interactions. We briefly describe the differences between familiar and non-familiar agent behaviors. Finally, we conclude with a discussion of limitations of our study and avenues for future research and applications of our corpus and corpus annotation systems.

2. The Computer Game

The human-ECA interaction took place in the form of a spoken-language version of a text-based adventure game entitled "Escape from the Castle of the Vampire King." The player controlled the game through speech commands rather than through text input, and the game had a graphical interface with a full-sized embodied conversational agent that served as the game's narrator. As the player explored the world of the game, the interface projected on the wall incrementally displayed a map showing where the player had been.

The game was written to appear as natural as possible. The game dialogs spoken by the agent were scripted in such a way so that even with limited commands, users could answer with natural language restricted to a specific context. Table 1 presents an excerpt of a player's interaction, showing the simple form of the player's verbal production.

Agent:	You enter the courtyard. A fountain rests in the room's center. A chapel lies in the distant west and the lobby to the south.
Player:	Examine the fountain.
Agent:	The water is green with algae and unsafe to drink. You notice a sign that reads: No magic allowed.
Player:	Use the magic potion.
Agent:	You dump the magic potion into the overgrown fountain. The vines within the fountain begin to smoke and hiss, finally burning away until you can see a dirty red gem lying over the drain in the fountain's base.
Player:	Pick up the red gem.
Agent:	You pick up a dirty red gem that looks like an eye.
Player:	Go to the chapel.

Table 1. An interaction transcript from the first session.

The vampire game comprises 26 different rooms, each with its own items, secret passages, points of interest, descriptions, and vampires. The agent is voiced by a text-to-speech engine that responds to several versions of four available commands (e.g., take the potion, pick up the potion, grab the potion). The commands are *move*, *take*, *use*, and *examine*; these commands can be applied to locations or items. For its part, the agent can respond to misunderstandings or unknown commands in five different ways.

3. Corpus Collection

We developed a first version of the system before the version from which the corpus reported here was collected. With Version 1, players were given two sheets, one with a printed set of commands and their respective examples and a second with a template for drawing a map to mark the player's progress. We found that in Version 1, players would concentrate their gaze on the sheets rather than on the agent. For the rapport study to be effective, we needed the players to be looking at the agent so that the players would perceive differences in the agent's behaviors, our independent variable. So to immerse the players and fix their gaze towards the agent, we developed Version 2 of the game, which featured a small help box in the upper-left corner of the projection and a map displayed behind the agent that was automatically updated as the user progressed through the game. This also reduced the cognitive load required to play the game, as memorizing every place that players visited and every item they carried at any point in time would make the game impractical and effectively unplayable.

The game play took place in the Immersion Lab of UTEP's Interactive Systems Group. A full-body life-sized ECA was projected on a wall, roughly 18 feet diagonal, with a displayed background that resembles other walls of the Immersion Lab, which we intended to suggest that both the player and the agent were co-located in the same physical space. Figure 1 shows one of the authors conversing with the ECA during a game.

In each session the agent displayed nonverbal behaviors that reflected the study's independent variable of familiarity vs. non-familiarity. Although it is possible to slowly transition from the non-familiar to the familiar animations in a single session, we opted to include only

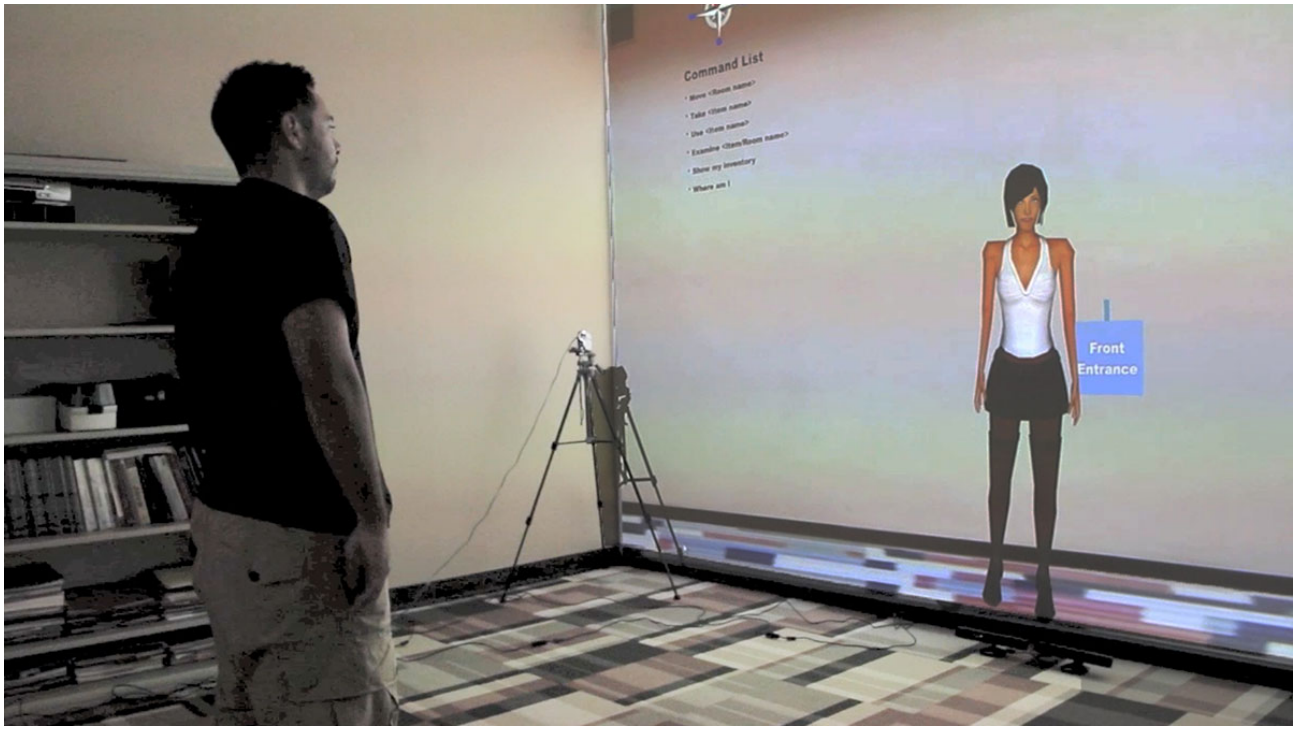

Figure 1. A conversation in the Immersion Lab between a player and the ECA

one type of behavior (i.e., low or high amplitude) per session to make a clear distinction between them and to ensure that subjects find differences in each behavior. The initial conversations exhibited non-familiar behaviors (low amplitude). The second sessions alternated between the behaviors (half with non-familiar and half familiar).

We recruited 20 undergraduate students to play with the agent over two days, in thirty-minute sessions; the subjects were assigned randomly to the familiar or non-familiar condition in the second session. We recorded both video and audio in each session from two angles, one from a Microsoft Kinect and one from a regular digital camcorder. The Kinect recorded the locations and angles of twenty user joints (see Figure 2). A normal stance and crossed arms were automatically detected and annotated on the log file; however the agent did not react to any position. We tested and recorded a total of 40 conversations, two for each of the 20 participants.

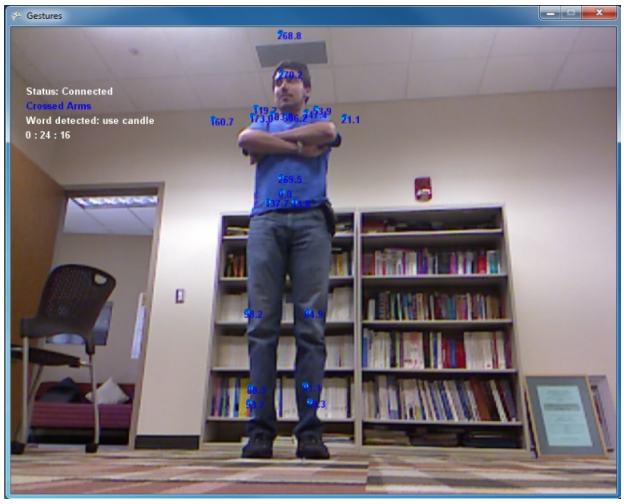

Figure 2. Pose recognizer detecting crossed arms

4. Annotations

Each episode was automatically annotated using two different methods, a game-save file and a log file.

With the game-save-file method, subjects were asked to save their game after their first session so that the agent would remember their previous interactions when they returned for the second session. These save files contain a list of all the valid interactions that led to a state change in the game; these valid interactions are immediately and silently recreated when the game is loaded.

The log-file method created a log file with a time stamp. These log files contain the current pose (normal stance or arms crossed) and what was understood by the agent. The log file was updated after every utterance that was heard by the ECA. Figure 3 presents examples of both a log file and a save file.

5. Limitations

Pose annotations were limited, and while the ECA logged them it did not react to particular positions. Because the Kinects were visible to the players, most players were aware that their pose might be recognized, and some even consciously attempted to make the agent react to their movements. In addition, the game task required the players to remember a considerable amount of game information, even when we displayed the map. Consequently, there were extended periods of silence or inactivity while players attempted to recall something.

Finally, the physical position of the Kinect sensor was not optimal. Because the wall served as a projection screen, the Kinect had to be placed on the floor close to the wall. The Kinect does not have high-resolution cameras, so images at this distance were difficult to analyze. In particular, even though we dedicated one of the Kinects specifically to the subjects' facial expressions, we failed at effectively recording and annotating facial gestures. Figure 4 shows the Kinect tracking the face of a person playing the game in the Immersion Lab.

6. Future Work

We expect to improve and expand the system by using annotations from unrecognized arbitrary poses to create new detectors. In particular, we want to collect additional data from the pose detector. As it is, we can recognize and annotate particular poses with their timestamp; however, creating the detection for these poses is a lengthy process. Figure 5 shows a manually coded detector of frustration gestures based on our corpus. It includes the angles between joints per participants and

```
1  Log File                                                                                                    | Save File
2  ------------------------------------------------------------------------------------------------------------------------
3  Time: 0 : 23 : 850 -- Word recognized: load game player one   -- Pose: No pose detected    | load game
4  Time: 0 : 33 : 816 -- Word recognized: take rusty sword       -- Pose: Normal Stance       | take rusty sword
5  Time: 0 : 43 : 416 -- Word recognized: go to the dinning hall -- Pose: Normal Stance       | move dinning hall
6  Time: 0 : 48 : 916 -- Word recognized: go to the clock tower  -- Pose: Normal Stance       | move clock tower
7  Time: 0 : 53 : 933 -- Word recognized: go to the courtyard    -- Pose: No pose detected    | move courtyard
8  Time: 1 : 7  : 200 -- Word recognized: use rusty sword        -- Pose: No pose detected    | use rusty sword
9  Time: 1 : 31 : 466 -- Word recognized: go to the library      -- Pose: Normal Stance       | move library
10 Time: 1 : 44 : 200 -- Word recognized: use holy water         -- Pose: Normal Stance       | use holy water
11 Time: 1 : 58 : 783 -- Word recognized: pick up ancient book   -- Pose: Normal Stance       | take ancient book
```

Figure 3. Example of a log file (left) and save file (right)

several statistical measures to calculate efficient margins of error. The next step is to collect the information related to the angles between joints and create new poses from them. We also hope to improve the illumination, camera, microphone and sensor location, and file compression to attain portable, high quality media that automatically provides additional information to improve the behavior of our agents in real time.

A corpus for Version 3 of the Escape from the Castle of the Vampire King game will be forthcoming. The new corpus will differ primarily with respect to improved game-play, including using recorded speech for the ECA and having backgrounds that represent the world of the game rather than the virtual reality of the Immersion Lab. For the longer run, we are building a new game, based on a jungle survival scenario, that is designed to support a more conversational style of dialog, advanced gesture recognition, longer-term interaction, and, at least to a limited extent, the mutual-understanding dimension of rapport.

7. Acknowledgments

The authors thank Guillaume Adoneth and David Manuel for their contributions to the design of this study, Jonathan Daggerhart for permission to adapt his original text-based adventure game into "Escape from the Castle of the Vampire King," and Alex Rayon, Adriana Camacho, Baltazar Santaella, Juan Vicario, Joel Quintana and Anuar Jauregui for their help in developing the game.

8. References

Anderson, T., Galley, S.: The history of Zork. *The New Zork Times*, 4(1-3) (1985).

Cafaro, A., Vilhjálmsson, H. H., Bickmore, T., Heylen, D., Jóhannsdóttir, K. R., Valgarðsson, G. S.: First impressions: Users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In *Intelligent Virtual Agents* (pp. 67-80). Springer Berlin Heidelberg (2012).

Clausen-Bruun, M., Ek, T., Haake, M.: Size certainly matters–at least if you are a gesticulating digital character: The impact of gesture amplitude on addressees' information uptake. In *Intelligent Virtual Agents* (pp. 446-447), Springer Berlin Heidelberg (2013).

Crowther, W., Woods, D., Black, K.: Colossal cave adventure. *Computer Game* (1976).

Neff, M., Wang, Y., Abbott, R., Walker, M.: Evaluating the effect of gesture and language on personality perception in conversational agents. In *Intelligent Virtual Agents* (pp. 222-235). Springer Berlin Heidelberg (2010).

Novick, D., Gris, I.: Building rapport between human and ECA: A pilot study. In *Proceedings of HCI International 2014* (in press).

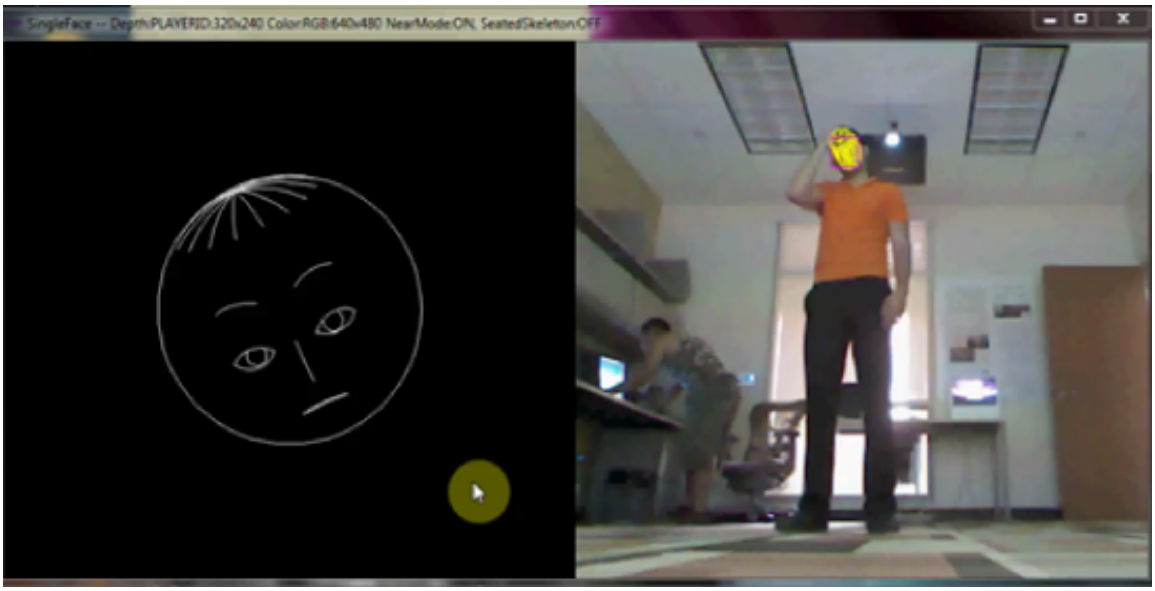

Figure 4. Facial gesture recognition

Frustration 2	P1	P2	P3	P4	P5	P6	MAX	MIN	AVG	DIF	DIF2	DIF3	RANGE
Shoulder Center - Shoulder Left	52	35	36	33	43	6	65	33	44.0	21.0	11.0	32	10.0
Shoulder Left - Elbow Left	338		135	316		32	338	135	277.8	60.3	142.8	203	-82.5
Elbow Left - Wrist Left	213	213	211	219	250	23	250	211	224.0	26.0	13.0	39	13.0
Wrist Left - Hand Left	260	251	241	225	249	26	268	225	249.0	19.0	24.0	43	-5.0
Shoulder Center - Shoulder Right	146	152	141	149	157	15	157	141	149.7	7.3	8.7	16	-1.3
Shoulder Right - Elbow Right	112	95	98	112	103	11	118	95	106.3	11.7	11.3	23	0.3
Elbow Right - Wrist Right	88	92	92	102	89		102	88	92.6	9.4	4.6	14	4.8
Wrist Right - Hand Right	143	93	96	102	108		143	93	108.4	34.6	15.4	50	19.2

Figure 5. Coding of frustration gestures

Multimodal Annotation of Synchrony in Longitudinal Parent–Child Interaction

Kristina Nilsson Björkenstam and Mats Wirén

Section for Computational Linguistics, Department of Linguistics, Stockholm University
SE-106 91 Stockholm, Sweden
kristina.nilsson@ling.su.se, mats.wiren@ling.su.se

Abstract

This paper describes the multimodal annotation of speech, gaze and hand movement in a corpus of longitudinal parent–child interaction, and reports results on *synchrony*, structural regularities which appear to be a key means for parents to facilitate learning of new concepts to children. The results provide additional support for our previous finding that parents display decreasing synchrony as a function of the age of the child.

Keywords: language acquisition, parent–child interaction, multimodal annotation, discourse annotation, synchrony

1. Introduction

A cognitive model of language learning ultimately needs to be dialogue-driven and multimodal to reflect the interaction of parents and children, involving devices such as words, gaze and object manipulation, and to reduce the complexity of the grammar induction problem (Clark and Lappin, 2011, p. 207). As a basis for such a model, we have developed a scheme for annotation of video and sound recordings of longitudinal parent–child dyads. A specific phenomenon that we are studying within these dyads is *synchrony*, "relatively stable patterns or structural regularities" (Gogate and Hollich, 2010, p. 496), which appear to be a key means for parents to facilitate learning of new concepts to children. As a vehicle for studying this, we use two target objects in the form of cuddly toys, *Kucka* (a yellow rabbit) and *Siffu* (a black monkey), set in an otherwise free-play scenario of the parent and child.

To get a handle on synchrony, we annotate dialogue segments involving uni- or multimodal reference (speech, gaze and/or hand movement) to either of the two target objects. The results strengthen the support for our previous finding that parents display *decreasing* synchrony as a function of the age of the child (Björkenstam and Wirén, 2012). Furthermore, we have obtained results using a more fine-grained annotation scheme for discourse (Björkenstam and Wirén, 2013). This makes it possible to look separately at instances of synchrony that occur during the initial mention of an object in a focus shift and in subsequent dialogue.

2. MINGLE: A longitudinal corpus of parent–child interaction with multimodal annotation

The audio and video recordings have been made from naturalistic parent–child interactions in a recording studio at the Phonetics Laboratory at Stockholm University, using two cameras (Lacerda, 2009). The speech signals from the the child and parent were recorded in separate channels via wireless lavalier microphones. One was attached to a vest that the child wore during the session, and the other was clip-mounted on the shirt of the parent. The child and parent were thus free to move around in the studio, and were provided with toys, including the two target objects *Kucka* and *Siffu*. The scenario was free play, but the parent was instructed to use the toys. The free play sessions were typically followed by a session when the parent and the experiment leader chat informally while working through the Swedish Early Communicative Development Inventory (SECDI, a version of the MacArthur Communicative Development Inventory) with the child in the room. These sessions have also been transcribed, and provide a valuable comparison for the parent–child interaction (Björkenstam et al., 2013).

The tool used for synchronization of video and audio files as well as multimodal annotation is ELAN[1] (Wittenburg et al., 2006).

2.1. Verbal annotation

All utterances by parents and children are transcribed, and we also keep an additional record of each spoken mention of one of the target objects by either parent or child.

2.1.1. Transcription

All utterances by the parent have been orthographically transcribed, with labels for features like laughter and onomatopoeia. The following disfluency categories are annotated: truncated words and phrases, prolongations, and filled pauses. Utterances interpreted as exclamations, appeals or orders are marked with an exclamation mark, and questions with a question mark. Utterances interpreted as adult-directed are labeled as such, while the default is child-directed speech.

In the early recordings, vocalizations by the child are phonetically transcribed, and in later recordings as a combination of orthographic words and non-word vocalizations.

2.1.2. Target object mentions

We define mentions as names (*Kucka*, *Siffu*), definite descriptions such as *den gula kaninen* ('the yellow rabbit') and *apan* ('the monkey'), or third person singular pronouns

[1] ELAN, by The Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands. URL: http://tla.mpi.nl/tools/tla-tools/elan/

den ('it'), *han* ('he'), and *hon* ('she'). The timewise extent of each verbal mention of the target object by child or parent are indicated in the annotation.

2.2. Multimodal annotation

A subset of the MINGLE corpus has also been multimodally annotated with eye gaze, gestures, and object-related actions. This subset, called MINGLE-3, consists of 18 dyads with three children; two girls and one boy recorded between the ages of 7 and 33 months, with six dyads per child. The complete duration of the 18 dyads is 7:29 hours and the multimodally annotated parts have a total duration of 1:39 hours.[2] The mean duration of a dyad is 24:58 minutes, and the mean duration of the multimodal annotation per dyad is 5:30 minutes. The children were interacting interchangeably with their mothers (10 dyads) and fathers (8 dyads).

A basic condition for annotating a segment multimodally is that the parent has mentioned (orally referred to) a target object using, e.g., the name *Kucka*. Given that such a mention exists, the segment starts at or before this point when one of the target objects is brought into focus by either parent or child by means of speech, eye gaze or hand movement, and ends whenever focus is similarly shifted to another object. The rationale for this is that what we primarily study is synchrony manifested by the parent, and that therefore a spoken mention by the parent is a confirmation that the target object is indeed in focus (at least for the parent).

2.2.1. Gaze

Eye gaze is annotated by marking whether the child (parent) is looking at a) the parent (child); b) Siffu or Kucka; or c) at any other object.

2.2.2. Gestures

Hand gestures by the parent and the child are annotated according to functional categories commonly used in multimodal annotation of adult interaction (McNeill, 1992; Allwood et al., 2007) and adult–child interaction (Parladé and Iverson, 2011; Stefanini et al., 2009).

2.2.3. Object-related actions

We categorize hand movements involving objects as object-related actions. Such actions (by both parent and child) are annotated by marking the timewise extent of the action with a description of the action and the object. There are three kinds of objects: a) the parent and the child; b) the target objects; and c) all other objects. In our data, a typical object-related action by the parent is to pick up an object and display it to the child by, e.g., holding it or shaking it within the visual field of the child. Many of these actions (e.g., holding objects up) can be categorized as manipulative forms of deixis (McNeill, 1992, p. 327). We include and describe all actions involving objects, e.g., banging a toy against the floor or dressing a doll.

[2] A previously reported study on a subset of this data was based on annotations with a duration of 40:59 minutes from five dyads with two children (Björkenstam and Wirén, 2012).

2.3. Discourse annotation

We annotate shifts in focus of attention by categorizing parent object mentions as *initial mention* or *subsequent mention*. That is, we distinguish between the first mention of an object and any follow-up mentions of that object until the next focus shift. Such focus shifts are detected by combining the verbal and non-verbal annotation.

Each instance annotated as initial mention is further annotated for initiative as either a *bring-in* when the parent introduces the object by either a vocal reference or a combined vocal and non-vocal reference to the child, or a *follow-in* when the child introduces the object (by speech, gaze, and/or hand) and the parent responds by either vocal or vocal and non-vocal means. Note that in the latter case, joint attention is already established.

3. Synchrony across modalities of parent–child interaction

In this paper, we are primarily interested in exploring synchrony across modalities as manifested in parent–child interaction. Thus, we discuss speech, eye gaze, and hand movements by parents and children with respect to the target objects, and the extent to which references to these objects are synchronized.

We have grouped the dyads based on the children's age at the time of recording, resulting in four data sets representing the ages 7–9 months, 12–14 months, 16–19 months, and 27–33 months.

3.1. Analysis of synchrony

The analysis of synchrony is based on the parent's target object mention annotation described in section 2.1.2. For each such mention (row P-Speech in table 1), we determine if the parent was synchronously looking at and/or manipulating the object.

From the eye gaze annotation described in section 2.2.1., we extract information on whether child or parent is looking at the target object in synchrony with the speech (rows P-Gaze and C-Gaze in table 1). Furthermore, from the annotation of gestures (section 2.2.2.) and object-related actions (section 2.2.3.), we extract information on whether child or parent is handling the target objects in synchrony with the speech (rows P-Hand and C-Hand in table 1), e.g., by pointing at the object, or by reaching for, holding, shaking, or offering it to the child.

By "synchronously" we mean overlap with a time interval beginning 0.5 seconds before and ending simultaneously with the parent's spoken mention of the object. Overlaps were given 1 point, and non-overlaps were given 0 points. The figures in Table 1 are means, and thus reflect the proportions of the data points exhibiting synchrony with the spoken modality of the parent. Note that the duration of the recordings vary, since the scenario of the child-parent interaction is free play.

3.1.1. Synchrony and discourse analysis

During a free play session, the (joint) focus of attention changes over time as the parent and child play with different objects. Our hypothesis was that when a new referent

Table 1: Proportions of synchrony of parent (P) and child (C) with the spoken modality of the parent (P-Speech) with respect to target objects for different modalities (Gaze, Speech, Hand) in MINGLE-3. Boldface indicates statistically significant difference to boldfaced neighbour.

Child age (months)	7–9	12–14	16–19	27–33
P-Speech	1	1	1	1
P-Gaze	**0.49**	**0.64**	0.52	0.47
P-Hand	**0.77**	**0.51**	**0.28**	0.34
C-Speech	0	0.02	0.12	0.08
C-Gaze	0.81	**0.73**	**0.54**	0.66
C-Hand	0.44	0.41	0.49	0.50
No. of dyads	5	5	5	3
No. of data points	217	240	153	38
Duration (m:s)	35:46	37:29	15:50	09:48

Table 2: Initial mentions categorized as *bring-in*: Proportions of synchrony of parent (P) and child (C) with the spoken modality of the parent (P-Speech) with respect to target objects for different modalities (Gaze, Speech, Hand) in MINGLE-3. Boldface indicates statistically significant difference to boldfaced neighbour.

Child age (months)	7–9	12–14	16–19	27–33
P-Speech	1	1	1	1
P-Gaze	0.50	0.64	0.10	0.29
P-Hand	**0.98**	**0.73**	**0.19**	0.43
C-Speech	0	0	0	0
C-Gaze	0.64	0.62	0.29	0.29
C-Hand	0.14	0.18	0.10	0.00
No. of data points	42	55	21	7

Table 3: Initial mentions categorized as *follow-in*: Proportions of synchrony of parent (P) and child (C) with the spoken modality of the parent (P-Speech) with respect to target objects for different modalities (Gaze, Speech, Hand) in MINGLE-3. Boldface indicates statistically significant difference to boldfaced neighbour.

Child age (months)	7–9	12–14	16–19	27–33
P-Speech	1	1	1	1
P-Gaze	0.67	0.76	0.69	0.63
P-Hand	0.33	0.20	0.10	0.25
C-Speech	0.00	0.08	0.24	0.13
C-Gaze	0.90	**0.92**	**0.62**	0.88
C-Hand	0.62	0.72	0.83	0.75
No. of data points	21	25	29	8

Table 4: Subsequent mentions: Proportions of synchrony of parent (P) and child (C) with the spoken modality of the parent (P-Speech) with respect to target objects for different modalities (Gaze, Speech, Hand) in MINGLE-3. Boldface indicates statistically significant difference to boldfaced neighbour.

Child age (months)	7–9	12–14	16–19	27–33
P-Speech	1	1	1	1
P-Gaze	**0.46**	**0.62**	0.56	0.48
P-Hand	**0.78**	**0.48**	0.35	0.35
C-Speech	0.00	0.02	0.11	0.09
C-Gaze	**0.84**	**0.73**	**0.56**	0.70
C-Hand	0.50	0.44	0.48	0.57
No. of data points	154	160	103	23

is established through an initial mention, this would involve more synchrony than subsequent mentions, and further, that parents make use of different strategies when aligning their speech to the child's focus of attention as compared to when trying to get the child's attention. Tables 2 (bring-in) and 3 (follow-in) show the proportions of initial mentions exhibiting synchrony with the spoken modality of the parent, and table 4 shows the proportions of synchrony of subsequent mentions.

4. Discussion

4.1. Synchrony and child age

The results corroborate our previous finding that parents display decreasing synchrony in terms of hand manipulation of target objects as a function of the age of the child (P-Hand in Table 1). The differences are statistically significant according to a z-test of sample proportions, with $z = 2.6$, $p = 0.0091$ (two-tailed) in a comparison of the second and first age group, and with $z = 4.6$, $p < 0.0001$ (two-tailed) in a comparison of the third and second age group. For the oldest age group (27–33 months), however, the tendency seen in the table is that synchrony *increases* compared to the previous age group. How should this be interpreted? The oldest age group includes fewer dyads and data points than the others, and the difference compared to the third age group is not statistically significant ($z = 0.4$, $p = 0.6773$, two-tailed). Still, one might speculate whether the lack of difference (or seeming increase) corresponds to an actual change in behaviour.

Anecdotally, when looking at the videos of this age group (27–33 months old), the children seem to have lost much of their interest in the cuddly toys compared to when they were younger. The parents, having been instructed that some part of the dialogue should be devoted to the target objects, therefore must exert additional effort to introduce the toys, and it appears that one means for doing this is increased synchrony. In this case, however, the primary function of the synchrony is hardly to facilitate language understanding as the children have already grasped how the toys are referred to, but rather to evoke the children's interest and attention.

As for the parents' gaze synchrony, the differences with respect to age groups are not statistically significant, except between the second and first age group. One factor that may distort P-Gaze is that the child was occasionally sitting in the lap of the parent, and that hence the child and parent could not see each other's faces. Still, it is interesting to see that the maximal gaze synchrony occurs at 12–14 months. At around 9–12 months of age, children begin to acquire an "understanding of other persons as intentional agents like the self whose psychological relations to outside entities may be followed into" (Tomasello, 2009, p.

21). In particular, this is when children begin to look where other persons are looking (gaze following).

The synchrony displayed by the children is somewhat secondary, since the data points investigated are all based on the parents' verbal mention of an object. Still, they illustrate some key elements of the learning process, such as the parents' ability to establish and maintain joint attention, and to align their speech with the child's focus of attention (Estigarribia and Clarke, 2007). The only statistically significant difference with respect to the children's gaze synchrony (C-Gaze) is the decrease between the second and third age group. The differences with respect to the children's hand movements (C-Hand) are very small, and none of them are statistically significant. Naturally, very little child speech (C-Speech) occurs that is synchronised with parents' speech, and again no differences are significant.

4.2. Synchrony and discourse

For both bring-in initial mentions (when the parent is trying to shift focus) and subsequent mentions (when, in most cases, joint attention has already been established), we find a similar pattern of decreasing speech–hand synchrony as the children develop (P-hand in table 2 and table 4). While the proportion of speech–hand synchrony is higher for bring-in than for subsequent mentions for the first two age groups, we find the high proportions of synchrony in subsequent mentions particularly interesting as this shows the intersensory redundancy available to infants in parent–child interaction even when joint attention has been established.

We find that parents use different strategies when trying to get the child's attention as compared to when aligning their speech to the child's focus of attention, as shown by the different patterns of synchrony for bring-in and follow-in initial mentions. In the first age group, the parents display 98% speech–hand synchrony for the bring-in initial mentions (P-hand in table 2), but only 33% for the follow-in initial mentions (P-hand table 3). There is a decrease over time for both initial mention types: in the second age group, the parents display 73% synchrony for the bring-in mentions, and 20% for the follow-in mentions. We note that for follow-in initial mentions, the parents seem to respond primarily to child eye gaze, but also to child gestures and manipulation of target objects (especially in the third age group, 16–19 months). We also find that the parents display a high proportion of gaze synchrony for this mention type, and that this does not change over time.

4.3. Conclusion

In sum, additional data and an enriched multimodal annotation provide both a more complete and a more fine-grained picture of how parents appear to facilitate language learning for children by means of synchrony. In particular, the results further strengthen the support for the finding that parents display decreasing synchrony as a function of the age of the child.

5. Acknowledgements

This research is part of the project "Modelling the emergence of linguistic structures in early childhood", funded by the Swedish Research Council as 2011-675-86010-31. We thanks our annotators: Anna Ericsson, Joel Petersson Ivre, Annika Schwitteck, and Johan Sjons. We also thank the anonymous reviewers for valuable comments.

6. References

J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3–4):273–287.

K. Nilsson Björkenstam and M. Wirén. 2012. Reference to objects in longitudinal parent–child interaction. In *Workshop on Language, Action and Perception (APL) at The Fourth Swedish Language Technology Conference (SLTC 2012)*, Lund, Sweden, October.

K. Nilsson Björkenstam and M. Wirén. 2013. Multimodal annotation of parent–child interaction in a free-play setting. In *Workshop on Multimodal Corpora: Beyond Audio and Video (MMC 2013) at The 13th International Conference on Intelligent Virtual Agents (IVA 2013)*, Edinburgh, UK, September.

K. Nilsson Björkenstam, M. Wirén, and R. Eklund. 2013. Disfluency in child-directed speech. In *Proceedings of Fonetik. The XXVIth Swedish Phonetics Conference*, Studies in Language and Culture, 21, Linköping, Sweden.

A. Clark and S. Lappin. 2011. *Linguistic Nativism and the Poverty of the Stimulus*. Wiley-Blackwell.

B. Estigarribia and E. Clarke. 2007. Getting and maintaining attention in talk to young children. *Journal of Child Language*, 34:799–814.

L.J. Gogate and G. Hollich. 2010. Invariance detection within an interactive system: A perceptual gateway to language development. *Psychological Review*, 117(2):496–516.

F. Lacerda. 2009. On the emergence of early linguistic functions: A biological and interactional perspective. In K. Alter, M. Horne, M. Lindgren, M. Roll, and J. von Koss Torkildsen, editors, *Brain Talk: Discourse with and in the brain*, pages 207–230. Media-Tryck.

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

M.V. Parladé and J.M. Iverson. 2011. The interplay between language, gesture, and affect during communicative transition: A dynamic systems approach. *Developmental Psychology*, 47(3):820–833.

S. Stefanini, A. Bello, M.C. Caselli, J.M. Iverson, and V. Volterra. 2009. Co-speech gestures in a naming task: Developmental data. *Language and Cognitive Processes*, 24(2):168–189.

M. Tomasello. 2009. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. 2006. Elan: a professional framework for multimodality research. In *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006)*.

Annotation of space and manner/path configuration in bilinguals' speech and manual gestures

Federica Cavicchio, Amanda Brown, Reyhan Furman, Shanley Allen, Asli Özyürek, Tomoko Ishizuka and Sotaro Kita

CIMeC Unitn, Syracuse University, University of Alberta, University of Kaiserslauten, Ratboud University of Nijmegen, Tama University, University of Warwick

E-mail: author1@xxx.yy, author2@zzz.uuu.edu, author3@hhh.com

Abstract (10-point Times New Roman bold, centred)

Different languages and cultures use gestures differently. The goal of this paper is describing the coding scheme used to annotate a corpus of English/Italian bilinguals and English and Italian monolinguals describing, a set of stimuli designed to elicit the description of manner and path events and the corresponding gestures ("the Tomato Man stimuli"). The first question we investigated was the relationship between clause structure for motion event expressions and gestural representation of the same event. From the seminal work of Kita and Özyürek (2003), many studies have investigated manner and path in the verbalization of motion events and the co-produced manual gestures in different languages. Following Talmy's (1985) typology, English allows verbal constructions to conflate complex meaning within a single clause as path can be expressed as a "satellite" to the verb. That is, manner and path can be expressed in within a single clause (e.g. roll down). On the other hand, Italian is more restricted in situations in which manner of motion verbs can occur with path phrases. That is, manner and path are often realized by two verbs (scende rotolando, i.e. it goes down rolling). We describe the annotation scheme we used to code speech and gesture in English/Italian bilinguals and monolinguals. Three annotators codified 403 tokens corresponding to the gesture stroke phase. We coded gestural and verbal expressions of manner and path and gesture space.

1. Introduction

Different languages and cultures use gestures differently. The goal of this paper is describing the coding scheme used to annotate a corpus of English/Italian bilinguals and English and Italian monolinguals describing a set of stimuli designed to elicit the description of manner and path events and the corresponding gestures ("the Tomato Man stimuli", Özyürek et al., 2001). Researchers have been interested in whether bilinguals transfer gestures from one language to another, that is, if some linguistic or spatial aspects of gestures that are linked to a certain language are transferred to the other language while speaking (see Nicoladis, 2007 for a review). Gesture transfer (or lack thereof) will give an insight on how gesture and language are linked in production. This project investigated whether gestural expressions of manner and path, gesture frequency and gesture space transfer from the first language to the second language. Thanks to this coding scheme, we addressed two research questions on the matter of bilingualism. The first question investigated the relationship between clause structure for motion event expressions and gestural representation of the same event. From the seminal work of Kita and Özyürek (2003) many studies have investigated manner and path in the verbalization of motion events and the co-produced manual gestures in different languages. Following Talmy's typology (1985), English allows verbal constructions to conflate complex meaning within a single clause as path can be expressed as a "satellite" to the verb. That is, manner and path can be expressed within a single clause (e.g. *roll down*). On the other hand, Italian is more restricted in situations in which manner of motion verbs can occur with path phrases. Manner and path are often realized by two verbs (*scende rotolando*, i.e. *it goes down rolling*). Nevertheless, single-clause constructions such as "*rotola giù/su*" (rolls down/up) can be used by Italian native speakers. Following Kita and Özyürek (2003), it is expected that, regardless of the language, single-clause verbal constructions will be accompanied by conflate gestures, combining the information about manner and path in one movement, whereas two-clause verbal constructions will be accompanied by 2 gestures, one expressing manner and one expressing path.

The second question we investigated was how the four groups of speakers differ in gesture frequency and gesture space. For example, Italian is reported as a high gesture frequency language (Barzini, 1964; Kendon, 1992, 1995), as opposed to (British) English, described as a low gesture frequency language (Graham and Argyle, 1975). Another gesture parameter that varies across cultures is gesture size: bigger in Mediterranean cultures than in northern European cultures. Since the seminal study of Efron (1941/1972) comparing Jews and Italian immigrants' gestures, we know that in different cultures gestures differ in how they are performed in the space. In particular, Efron observed that Italian immigrants' gestures were spatially expansive, moving the entire arm from the shoulder joint, and tended to occupy the lateral (transversal) plane. More recently, Müller (1998) compared the gesture space of native Spanish and German speakers involved in a naturalistic conversation task with a language matching confederate. She found that Spanish speakers produced more gestures in the space above their shoulder than German speakers. Gesture size is an interesting variable to consider for gesture transfer in bilinguals. In the following, we describe the annotation scheme we used to code speech and gesture in English/Italian bilinguals and monolinguals. Three annotators codified 403 tokens corresponding to the gesture stroke phase. We coded gestural and verbal expressions of manner and path and gesture space.

2. Corpus description

Data were collected from two monolingual control groups so that we can properly address the questions whether

bilinguals' gestures are different from monolinguals' gestures and/or whether parameters of gesture production transfer from a language to another. The two monolingual control groups of English and Italian speakers were matched with the bilinguals for gender, age and education background. We focused on highly proficient Italian/English early bilinguals (i.e. they learned both languages before age 6) who had a very similar fluency in both languages. Bilinguals and monolinguals described the exact same stimuli in each language to a confederate language matching speaker. Stimuli consisted of 10 single-scene cartoons depicting actions performed by red tomato and green triangle. Participants where required to describe each cartoon as accurately as possible to a language matched monolingual speaker. 20 English native speakers, 20 Italian native speakers and 20 English/Italian bilinguals were recorded while describing to a matching language listener the ten Tomato man cartoons. Bilinguals described the stimuli twice, once in English and once in Italian, to two different native speakers. Monolinguals described the stimuli twice in their native language to two different native speakers.

4.1 Transcription

A native speaker of Italian and two native speakers of English transcribed the descriptions. Disfluencies, repetitions and laughter were transcribed with special fonts. The transcriptions were checked for accuracy by a fluent speaker of Italian and English. All the transcriptions were done in Elan 4.3.3 to ensure a correct time alignment with coverbal gestures. In this study we focused on the stroke phase of each gesture performed by the speakers, as defined by Kita, van Gijn, & van der Hulst (1998). Gesture strokes were transcribed and aligned with speech.

4.2 Coding scheme

The coding scheme was implemented in Elan 4.3.3. Annotators found the speech transcription and the gesture stroke already marked and aligned. The coding scheme for expressions of manner and path was adapted from the Coding Manual: NSF: Crosslinguistic motion event project (2004), which was developed from a coding scheme for Kita and Özyürek (2003) and used in subsequent studies (Allen et al., 2007; Kita et al., 2007; Özyürek, et al., 2005, 2008). The current coding scheme for verbal description was adapted from the Coding Manual: NSF Crosslinguistic motion event project (2004). With respect to the original manual, we added gesture space annotation to capture the difference, if any, in gesture salience between languages (English and Italian) and language groups (bilingual or monolingual).

Manner and Path verbal production ("verb type"): All the cartoons had three main action events. The annotators coded the verbal production corresponding to the target event of each cartoon. For example, they did not code the verbal typology regarding the initial event (e.g., *The triangle pushed the tomato*) or the final event (e.g., *Tomato bumped into the tree*) but only the verbal typology of the target event (e.g., *Tomato rolled down the hill*). There were four categories: **IV**, **2V**, **VP** and **VM**. The speakers may describe the target event with 1 verb (e.g., *it rolls up the water*- coded as **1V**); 2 verbs (e.g., *it ascends rolling to the shore*, **2V**) or can describe only the path (e.g., *it rose up to the shore*, **VP**) or only the manner (e.g., *it rolls to the shore*, **VM**).

Manner and Path gesture production ("gesture type"): We coded all the gestures that overlap with speech that refers to the "target event".

We coded the spatial pattern for gestures expressing manner and path into seven types: M, P, C, MC, U, J. The key question is whether the gesture encodes manner (**M**), path (**P**) or Manner and Path together (Conflate, **C**). Sometimes a gesture combines more than one type, expressing for example both manner and conflation (**MC**). For example, when describing the green triangle jumping around a tree, participants might gesture the jump event (Manner) followed by a conflate gesture (jump around the tree). In particular, for jumping, one jump (up and down) or one and half (up and down and up) will be coded as a separate category, **J**, as it is unclear whether it should be C or M. If a single jump/rotation in one location is followed by a clear C gesture, then it is coded MC (see the first example in the second row "atypical examples" in fig. 1). Gestures that cannot be classified into M, P, C MC or J were coded as unclear (**U**). Note that in order for a gesture to be coded as expressing manner, the gesture must have one full rotation; otherwise, it is coded as unclear (**U**). **Typical and atypical examples** of gestures are reported in fig. 1, left panel.

Gesture space annotation ("gesture salience")*:* Coders annotated the space areas where the gesture stroke took place. Gesture saliency was coded for the target gesture performed during description of the target event (e.g. rolls up). To code saliency we followed McNeill, who divided the gesture space into sectors using a system of concentric squares (McNeill, 1992, p. 89-see fig.1, panel on the right). Our annotation coding scheme reflects this notation dividing the gesture space in 2 sectors: "centre" and "periphery" expressed respectively with *not salient* and *salient*.

3. Coding scheme reliability experiment

To ensure the reliability of the adopted coding scheme, a subset of the corpus was annotated by three independent coders. For verbal description of motion events, 91 target events were coded. For gesture depiction of motion events and gesture salience the annotators rated 156 gestures. Annotators had the following training before starting their work. They were trained in a one-to-one session going through the manual with one of the authors. In the same session they were also trained on how to use ELAN. This first session lasted about 2 hours. After that annotators worked by themselves on three files. Once they completed the files, they met the author for the second time and individually went through their files with her. The annotators finalized the three files in this session. After that, the annotators went through more files alone. A Kappa statistics (Siegel and Castellan, 1998) was computed on the resulting annotated files.

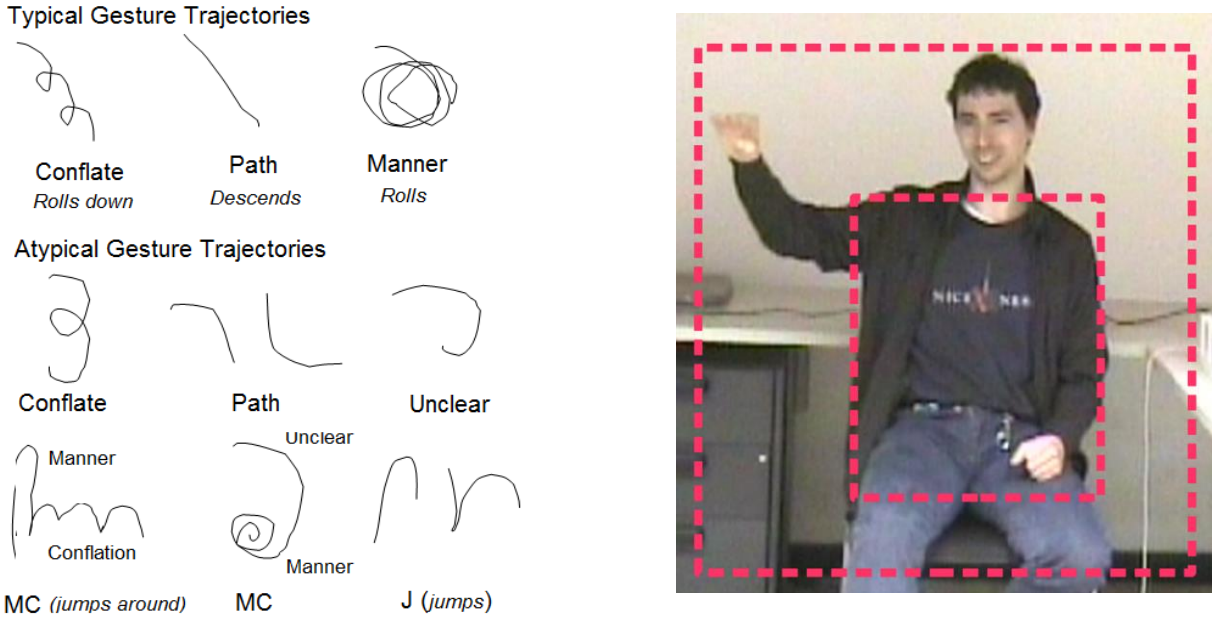

Figure 1: On the left panel, a graphical representation of typical and atypical examples of gesture trajectories ad their annotation (Coding Manual: NSF Crosslinguistic Motion Event Project, 2004, p.43). On the right, a participant describes one of the 10 cartoons. The dotted concentric squares define the gesture space: centre (the inner square, gesture is **not salient**) and periphery (the outer square, gesture is **salient**).

With regards verb type, the Kappa score was 0.81 ($p<.001$); for gesture type, Kappa was 0.78 ($p<0.001$), and for gesture salience, Kappa had the highest score, 0.89 ($p<.001$).

All the features of our coding scheme had Kappa scores above 0.75. The coding scheme is therefore highly reliable. For verb type all the coding scheme features reached a Kappa above 0.75. For gesture typology, one feature had a mild Kappa score and a low p value (Unclear: $K = 0.67$; $p < .001$). It is also worth noting that annotators agreed more on the labelling of Manner and Path gestures (Manner Kappa=0.9, $p<.001$; Path Kappa=0.9, $p<.001$) whereas Jump and Manner + Conflate gestures had intermediate scores (Kappa=0.8, $p<0.001$ for Jump and Kappa=0.79 $p<0.001$ for MC). Finally, the score for Conflate gestures was fairly high (Kappa=0.73, $p<0.001$). With regards gesture salience, both categories had a high Kappa score (Kappa=0.85, $p<.001$ for salient; Kappa=0.9, Kappa<.001).

4. Conclusion

Despite the increasing interests in gestures, there are still not many annotation coding schemes shared and by the multimodal corpora community (a notable exception includes Lausberg and Sloetjes, 2009). In this work we illustrated the annotation coding scheme adopted to investigate whether bilinguals change their gestures when switching from a language to the other. The issue has been addressed focusing on verbal and gestural expression of motion verbs (manner, path or conflation in speech and gesture) and on gesture salience. The proposed coding scheme for typology has been adopted from Coding Manual: NSF Crosslingusitic Motion Event Project (2004), whereas for gesture salience it has been applied for the first time, based on McNeill (1992). This coding scheme focuses on both gesture content and form. This is because we wanted to test both the form (gesture space and shape) and the gesture function (manner and path description) with regards to speech. With this report we make available our coding scheme to the community, hoping to contribute to the investigation of gesture/speech interaction.

5. References

Coding manual: NFS Crosslinguistic Motion Event Project. (2004). Nijmegen, the Netherlands: Max-Planck-Institute for Psycholinguistics. (Allen et al., 2007; Kita et al., 2007; Özyürek, et al., 2005, 2008)

Allen, S., Özyürek, A., Kita, S., Brown, A., Furman, R., Ishizuka, T., & Fujii M. (2007). How language specific is early syntactic packaging of Manner and Path? A comparison of English, Turkish, and Japanese. *Cognition*, *102*, 16-48.

Barzini, L. (1964). *The Italians*. New York: Atheneum.

Efron, D. (1972). *Gesture, race, and culture*. The Hague: Mouton. (Original work published as Gesture and environment, 1941.)

Graham, J. A., & Argyle, M. (1975). A cross-cultural study of the communication of extra-verbal meaning by gestures. *International Journal of Psychology*, 10, 57–67.

Kendon, A. (1992). Some recent work from Italy on quotable gestures ('emblems'). *Journal of Linguistic Anthropology*, 21, 72–93.

Kendon, A. (1995). Gestures as illocutionary and discourse structure markers in southern Italian conversation. *Journal of Pragmatics* 23: 247-279.

Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., & Ishizuka, T. (2007). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. *Language and Cognitive Processes, 22*(8), 1212-1236.

Kita, S., van Gijn, I., & van der Hulst, H. (1998). Movement Phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), *Gesture and sign language in human-computer interaction, International Gesture Workshop Bielefeld, Germany, September 17-19, 1997, Proceedings. Lecture Notes in Artificial Intelligence* (Vol. 1317, pp. 23-35). Berlin: Springer-Verlag.

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48, 16-32.

McNeill, David. 1992. *Hand and Mind: What Gestures Reveal About Thought*. Chicago: University of Chicago Press.

Muller, C. (1998). Redebegleitende Gesten: Kulturegeschichte -Theorie -Sprachvergleich. Berlin: Berlin Verlag.

Özyürek, A., Kita, S., Allen, S., Brown, A., Furman, R., & Ishizuka, T. (2008). Development of cross-linguistic variation in speech and gesture: Motion events in English and Turkish. *Developmental Psychology, 44*(4), 1040-1054.

Özyürek, A., Kita, S., Allen, S., Furman, R., & Brown, A. (2005). How does linguistic framing of events influence co-speech gestures? Insights from crosslinguistic variations and similarities. *Gesture, 5(1), 215-237*.

Özyürek, A., Kita, S., & Allen, S. (2001). Tomato Man movies: Stimulus kit designed to elicit manner, path and causal constructions in motion events with regard to speech and gestures. Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics.

Nicoladis, E. (2007). The effect of bilingualism on the use of manual gestures. *Applied Psycholinguistics*, 28, 441–454.

Siegel, S. & Castellan, J. N. (1988). *Nonparametric Statistics for the Behavioral Sciences*: McGraw-Hill, New York.

Talmy, L.(1985). Lexicalization patterns: semantic structure in lexical forms. In Language typology and syntactic description, vol. III: *Grammatical categories and the lexicon*, ed. Timothy Shopen, 57–149. Cambridge: Cambridge University Press.

Automatic Video Analysis for Annotation of Human Body Motion in Humanities Research

O. Schreer, S. Masneri
Fraunhofer Heinrich Hertz Institute
Einsteinufer 37, 10587 Berlin, Germany
E-mail: oliver.schreer@hhi.fraunhofer.de, stefano.masneri@hhi.fraunhofer.de

Abstract

The analysis of multi-modal audio-visual data is the very first step to perform research on gestural behaviour in a variety of disciplines such as psycho-linguistic, psychology, behaviour analysis or sign language. The annotation of human gestures and motion of hands is a very time consuming process and requests a lot of effort in terms of personal. Due to the large amount of available data in existing corpora, much video material has not been annotated and even not touched, i.e. valuable material cannot be exploited for research. Thanks to modern video processing algorithms some of the most time consuming annotation tasks can be performed automatically. In this paper, we present a video analysis tool that is specifically designed to perform automatic analysis and annotation of video material for the above mentioned research domains. The presented algorithm provides a large variety of annotations required for behaviour analysis without any user interaction in a fully automatic way. The proposed video analysis tool is currently designed to provide annotations according to the NEUROGES coding system for gestural behaviour, but it can provide also other means of annotations for other coding schemes.

Keywords: automatic annotation, video analysis, gesture, humanities

1. Introduction

In many different research disciplines in humanities, very large multimodal corpora are being processed and analysed in order to solve quite a large variety of different research questions. Not only in the gesture community, but also in psycho-linguistics and psychology, the way how human act with their hands and body is of interest. To solve the different research questions, a detailed annotation of multimodal data (video and speech) is performed. This annotation is basically performed manually by human raters and it is usually a very time consuming process.

In this paper, we present an automatic video analysis tool that supports the researchers in this exhaustive exercise inspecting and analysing videos. Many significant events in human body motion can be detected quite robustly by modern video analysis tools and therefore speed up the annotation process by a significant amount of time. Several experiments in the joint Max-Planck/Fraunhofer research project AVATecH proved a reduction of annotation time by more than 50% (Lenkiewicz, 2011). The presented analysis tool provides annotation in line with the NEUROGES coding system, developed at Deutsche Sporthochschule, Cologne (DSH) (Lausberg, 2013). This coding system provides a unified tree-based structure to describe gestural behaviour. It consists of three modules progressing from gesture kinetics to gesture function. However, the provided annotations by our video analysis tool can be modified and adopted to other means of gestural behaviour apart from the NEUROGES coding system.

The paper is organized as follows. In the next section, an overview of the automatic video analysis tool is given. Section 3 describes the resulting annotations related to the NEUROGES coding system. Section 4 presents experimental results based on the video material under investigation. A final summary concludes the paper.

2. Challenges and solutions for automatic video analysis

A useful video analysis tool must fulfil a number of different challenges in order to be easy to use by users in humanity research. Some of the main challenges are:

- The algorithms must be able to cope with different number of persons;
- varying background must be taken into account in terms of arbitrary colour, texture and motion as well as moving cameras;
- the algorithms should be able to extract meaningful information without any prior knowledge of the scene;
- the algorithms must be able to cope with different video quality, different spatial and temporal resolution;
- common video formats must be supported.

The overall goal is to achieve processing of videos in a fully automatic way, i.e. without the need for human interaction.

The tool presented thereafter is based on different video processing techniques aiming to achieve the above mentioned goals. It is built on top of previous developments in the context of the AVATecH project (Lenkiewicz, 2012), (Schreer, 2012).

The first step of the processing is the detection and tracking of hands and it is based on skin colour, which is a unique feature of humans. Together with motion information, the visible hands of the persons in the scene are detected and tracked. A face detection and tracking module provides necessary positional information and motion information (head rotation, eyes position) about the faces of the captured persons.

The skin-colour based hand tracking module provides a number of information for each frame such as:

- the position of the hand
- the speed of the hand movement in succeeding frames
- directional information of the hand movement in

succeeding frames

This frame-based information is then post-processed to get information for longer temporal segments e.g.:

- the start and end of a hand movement
- the temporal sub-segments in which the hand moves in the same direction
- relational information between hands and between hands and head

Furthermore, the tool also provides a number of additional important information as follows:

- If hands are touching each other, they are assigned additionally as joined hands.
- In the case a person is wearing a short-sleeves shirt (as shown in Figure 3), the tool automatically detects it and separate the arm from the hand region, to increase the accuracy of hand tracking.
- Quite often the hands of a person are not moving in space, but fingers are moved. This is very important information, which can be gathered from video analysis as well. This kind of movement is called intrinsic motion and hands are annotated respectively.
- Furthermore, the rest positions of both hands are calculated and are adapted over time. This rest position has been identified as valuable information for gesture researchers.

By using the result of the face detection module, a body part assignment is performed in order to relate the face of a person to the detected hands of the same person. In Figure 1, two examples are given, where the different hands, the head and the estimated rest position are visualized. The differently coloured ellipses at the hand position assign the left and right hand.

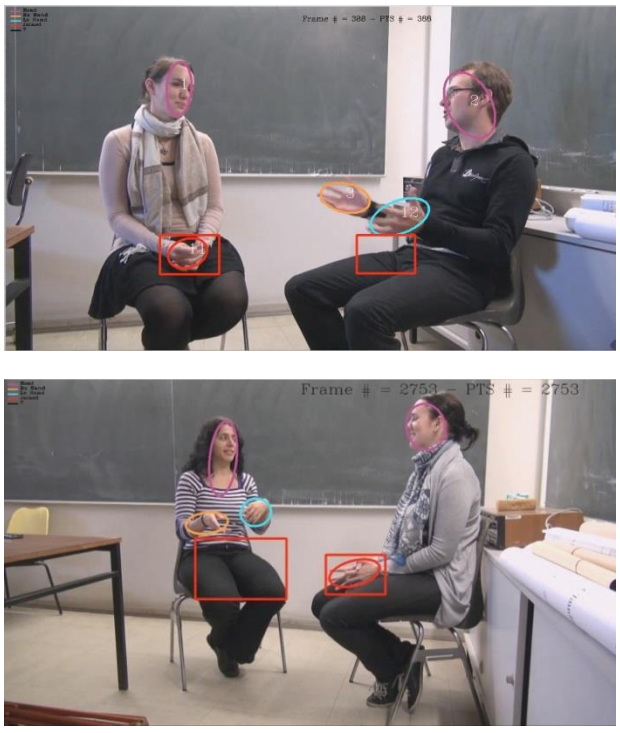

Figure 1: Results of hand and head tracking and related body part assignment

The rest position is assigned by a squared rectangle. As it can be recognized from the examples, not only frontal view scenarios can be analysed. However, the left-right hand assignment may be incorrect in side view situations.

3. Annotation of gestural behaviour

Due to collaboration with DSH in the German AUVIS project[1], the task was to provide annotations following the coding system for gestural behaviour, called NEUROGES.

The NEUROGES coding system is divided in three different modules, with increasing semantic complexity. Module I classifies the hand movement into the categories activation, structure and focus. Module II classifies the relation between both hands, while Module III considers several function and type categories. The most important information for the classification of structure in gestural behaviour in Module 1 is the detection of hand movement. It is obvious that this kind of annotation requires the biggest effort for manual annotation. Therefore, automatic video analysis can efficiently contribute to overall effort reduction, if annotations of hand movements can be provided.

In Figure 2, the different annotations for the structure of gestural behaviour of Module I are depicted.

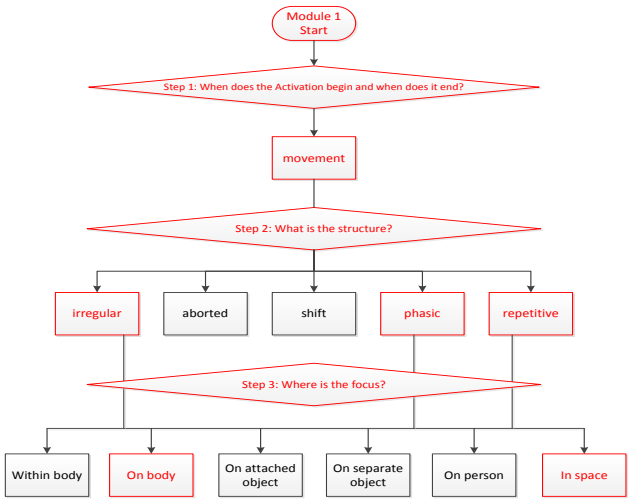

Figure 2: NEUROGES Module 1 with its categories Activation Units and Structure Units

The following definitions for the categories *phasic*, *repetitive* and *irregular* have been taken into account in order to transform the analysis results from video analysis into NEUROGES annotations:

- Phasic: „Movement with a phase structure and a static complex phase or a motion complex phase in which the movement path is one-way“
- Repetitive: „Movement with a phase structure and a motion complex phase in which the same movement path is used repetitively“
- Irregular: „small movements without distinct trajectory“

[1] http://tla.mpi.nl/projects_info/auvis/

The categories marked in red are directly derived from the available results of our automatic video analysis tool. The resulting annotations are stored in an xml-file that follows the notion of the multi-media annotation tool ELAN (Wittenburg, 2006). In Table 1, an example is given, how the results of the annotations based on automatic video analysis will look like.

```
<?xml version="1.0" encoding="utf-8"?>
<TIERS  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="./avatech-tier.xsd">
  <TIER columns="HHI_lh_Activation_P1"> ... </TIER>
  <TIER columns="HHI_lh_Focus_P1"> ... </TIER>
  <TIER columns="HHI_lh_Activation_Intrinsic_P1"> ... </TIER>
  <TIER columns="HHI_lh_Structure_P1">
     <span start="1.6800" end="1.8800">
        <v>irregular</v>
     </span>
     <span start="4.8000" end="5.0400">
        <v>irregular</v>
     </span>
     <span start="7.3200" end="9.2000">
        <v>phasic</v>
     </span>
  </TIER>
      <span start="0.0000" end="1.2000">
         <v>CENTER</v>
      </span>
      <span start="1.2000" end="1.4000">
         <v>CENTER_CENTER</v>
      </span>
      <span start="1.4000" end="3.0000">
         <v>CENTER</v>
      </span>
   </TIER>
</TIERS>
```

Table 1: Example for annotation results in xml style.

It is important to note that the available video analysis results can be used for many different kinds of higher level annotations. For example, in the gesture community, the annotation of preparation, stroke and retraction phase is important, which can be gathered from our results as well. Researchers are also interested about the position of the hands related to the body as defined in the McNeill gesture space (McNeill, 1992). As the position of the body is tracked continuously, the hands position can also be annotated relative to the body of the person according to McNeill's definition. In Figure 3, an example image is given showing rectangles that relate to the gesture space. The inner rectangle represents the centre-centre part of the gesture space.

4. Experimental results

The evaluation has been performed by comparing the automatically created annotations with the ones created by a human rater.

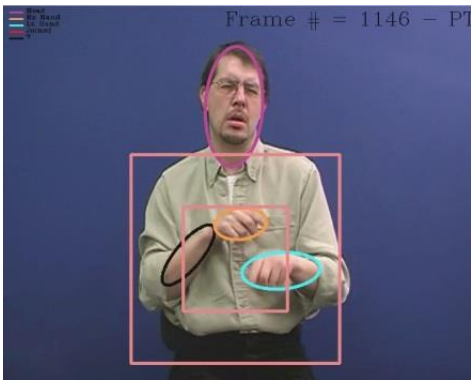

Figure 3: Visualization of gesture space

The ground truth information has been provided for two videos, for which the human rater has marked the start and end time of each movement, as well as the structure of the hand movement. Currently, we restricted the automatic analysis and annotation only to *phasic*, *repetitive* or *irregular*, since they are by far the most common occurring structures for a hand movement.

The detection of hand movements is based on the analysis of the speed of the tracked hands: when it is above a predefined threshold, the start of the movement is set; similarly, the end of the movement is set when the speed goes below threshold. The system allows the user to choose between three different thresholds, depending on the sensitivity required. The detection of intrinsic motion is different, because there is no hand movement in space. In that case the detection is based on the change of shape of the hand rather than on the change in speed. The classification of the structure of movement is based on a directional analysis: every movement is divided into various segments based on the main direction of motion. The classification of the whole movement into phasic, repetitive or irregular is then based on the analysis of the sub-segments' direction and position. A movement is classified as:

- *Phasic* if it has same start and end point and it has a unique directional pattern repeats at most once (e.g. start – left – up – down – right – end);
- *Repetitive* if it has same start and end point and its directional pattern repeats at least twice (e.g. start – up – down – up –down – end);
- *Irregular* if it doesn't have a clear directional pattern or it doesn't have the same start and end position.

The values of precision, recall and F-measure have been calculated checking if, for each frame in the two videos, there was an ongoing movement or not. Furthermore, the same procedure has been applied to evaluate the correctness of the assignment of the structure of hand movement. The results are summarized in Table 2. The algorithm achieves a very high accuracy in the detection of movement, while it is slightly less accurate when it comes to distinguishing the type of hand movement. The main source of error results from the misclassification of phasic movements.

	Precision	Recall	F1-Measure
Movement	91.2%	64.1%	75.3%
Phasic	48.5%	19.9%	28.3%
Repetitive	47.4%	40.2%	43.6%
Irregular	54.0%	64.0%	58.6%

Table 2: Performances on the detection and classification of structure of hand movement

Table 3 shows the confusion matrix that summarizes the classification errors between ground truth annotation and automatic annotation across the different classes under investigation.

	No move	Phasic	Irregular	Repetitive
No move	76.0%	5.8%	7.4%	10.8%
Phasic	33.8%	19.9%	12.4%	33.9%
Irregular	44.5%	2.5%	48.2%	4.8%
Repetitive	31.4%	14.9%	13.4%	40.3%

Table 3: Confusion matrix for hand movement classification

Another important performance measure is the accuracy of the algorithm with respect to the detection of the start and the end of a hand movement. The evaluation of this metric is based on the offset between the start and end of the hand movement as selected by the human rater compared to the one detected by the automatic video analysis. The results show high variability ranging from an almost perfect accordance (less than 0.5 seconds difference) in most of the cases, but also big differences when the algorithm fails to detect a hand movement. It is also worth noticing that the detection of the start of a movement is skewed towards a positive error, which means the start of the movement detected by the algorithm occurs most of the time after the start assigned by the human rater. No such behavior has been noticed while detecting the end of a movement. Table 4 summarizes these results, showing the median of the time offset, both when the detection occurs later and when it occurs before the ground truth value.

	Start (+)	Start (-)	End (+)	End (-)
Median	0.45 s	1.37 s	0.96 s	0.84 s
Occurence	62.4%	37.6%	47.0%	53.0%

Table 4: Median and occurrence of the difference in detection of start and end time of hand movement

5. Conclusion

A tool for automatic video analysis and annotation has been presented. The system described allows the researchers to save time by automatically detecting body parts and recognizing hand movement. The tool can be used in different research areas (i.e. gestural behaviour analysis, sign language analysis) and is capable to deal with a large variety of scenarios such as multiple persons, moving camera, short-sleeves tracking of hands and non-uniform background scenarios.

The result of the analysis consists in a series of annotations representing the movements of the hands over time and the spatial relationship between the hands and the body. The higher level semantic annotations provided are designed to follow the NEUROGES coding system developed by Deutsche Sporthochschule Cologne, Germany. The tool can run from within ELAN and the annotations it creates can also be exported as XML files for further analysis. First evaluations of the accuracy of the automatic annotations are promising, but further improvement is still required.

Future work aims to improve the classification of gestures, to make the system more robust (i.e. improve tracking in case of illumination changes or slow camera movements) and to add new types of annotations such as the ones from Module 2 and 3 from the NEUROGES coding manual.

6. Acknowledgements

This work was supported by German Ministry of Education and Research, Grant no. 01UG1240B. We gratefully thank Prof. Hedda Lausberg and Harald Skomroch, Deutsche Sporthochschule Köln, Germany for their valuable feedback and support with respect to the NEUROGES coding system.

7. References

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. Proc. of LREC 2006, Fifth International Conference on Language Resources and Evaluation.

Lausberg, H. (2013). Understanding Body Movement: A Guide to Empirical Research on Nonverbal Behaviour: With an Introduction to the NEUROGES Coding System. Publisher: Peter Lang, Frankfurt/Main, Editor: Hedda Lausberg, ISBN: 978-3-631-58249-7.

Lenkiewicz P., Wittenburg P., Masneri S., Schreer O., Gebre B.G., Lenkiewicz A. (2011): Application of Video Processing Methods for Language Preservation, 5th Language & Technology Conference (LTC), Poznan, Poland, November 25-27, 2011.

Lenkiewicz P., Auer E., Schreer O., Masneri S., Schneider D., Tschöpel S. (2012): AVATecH - automated annotation through audio and video analysis, In N. Calzolari (Ed.), Proceedings of the Eighth Int. Conference on Language Resources and Evaluation (LREC 2012), Istanbul, pp. 209-214. European Language Resources Association, May 23-25, 2012.

McNeill, D. (1992). Hand and Mind: What Gestures Reveal about Thought, Chicago: The University of Chicago Press, 1992.

Schreer O., Schneider D. (2012): Supporting linguistic research using generic automatic audio/video analysis, Language Documentation & Conservation Special Publication No. 6 (2012): Potentials of Language Documentation: Methods, Analyses, and Utilization, ed. by Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek pp. 39–46.

A Multimodal Corpus of Communicative Behaviors of Disabled Individuals during HRI

Trine J. Eilersen, Costanza Navarretta

University of Copenhagen
Njalsgade 140, build. 25
2300 Copenhagen S - Denmark
E-mail: cgt315@alumni.ku.dk, costanza@hum.ku.dk

Abstract

This paper describes the collection of a multimodal corpus of video- and audio-recorded interactions between an anthropomorphic robot and normal and cognitively disabled individuals. The aim of the work is to provide data for the study of the multimodal behaviors of the two groups of test participants during the conversational interactions. The study of the communicative multimodal behaviors of possible future users of assistive social robotics is expected to provide useful information for the development of human-robot interfaces. The data-collection was conducted using an anthropomorphic robot, NAO, which was interacting with the participants via a Wizard of Oz setting and a modular dialogue-script. Test participants were recruited from 2 municipal care centers for adult individuals with multiple disabilities located in Copenhagen, Denmark. The corpus was collected over 2 weeks and consists of recordings of 17 dyadic interactions. Each interaction comprises chat-based conversation units and a small cooperative game with a ball. A first analysis of the video-recordings shows that both disabled and non-disabled participants interacted multimodally (speech and body behaviors) with the robot. Furthermore, the answers of the participants to a questionnaire about their feelings towards the communicative situation show that they were unaffected by the experimental set-up while they were very affected by the meeting with the robot.

Keywords: Multimodal Corpora, Human-Robot Interaction, Wizard of Oz

1. Introduction

The entry of social robotics into the field of welfare technological services has created the need for studying and modeling how users with different needs and cognitive capabilities react to this emerging technology. For a robot to successfully become part of the everyday life of a user, it must be able to interact with him adapting to his particular reactions. This is especially the case with cognitively disabled users.

While much work has been conducted studying the behaviors in HRI of normal users, less emphasis has been put on analyzing the behaviors of cognitively disabled users. This paper, addresses this issue. More specifically, we describe the construction of a multimodal corpus of interactions between disabled and non-disabled subjects with an anthropomorphic robot.

To our best knowledge, there are no other corpora for studying the communicative body behaviors of cognitively disabled people with post-stroke and developmental disorders in interaction with an anthropomorphic robot like NAO, at least in Danish.

The paper describes how a Wizard of Oz (WoZ) based approach was taken to collect the conversational corpus, which consists of seventeen interactions in Danish between humans and a "talking" robot. Thirteen of the test participants are cognitively disabled and were recruited from a municipal centre for multiple disabilities in Copenhagen (CMF), and four were not disabled. The NAO robot, lent to the project by the robotics department of the Danish Technological Institute (DTI), was used. The interactions with the robot were realized using a modular dialogue script whose components were appropriately activated by the wizard. The resulting interactions consist of a mixture of chat-based conversations and physical interactions during a cooperative game.

2. Background Literature

Face-to-face communication involves both speech and body behaviors and the importance of these in the development of more human-like interaction systems is recognized in research and the temporal and semantic relation between body behaviors and speech indicates that they have a common cognitive base, as it is suggested inter alia by McNeill (2005). Allwood (2002) focuses on the communication management functions of multimodal behaviors which are fundamental for the success of conversational interactions and in particular on feedback giving (backchannelling) and eliciting. Paggio and Navarretta (2013) have exploited feedback head movements, facial expressions and speech in a Danish corpus of first encounters which was formally annotated according to the MUMIN scheme (Allwood et al. 2007).

The use of communicative body behaviors has also been investigated in HRI. For example, Rehm et al (2009) show how a cross-cultural corpus of human-human interactions can provide empirical data for implementing multicultural agents. Han et al. (2012) collect a multimodal corpus for the study of timing in chat-based conversations and Al Moubayed et al. (2009) study the development of robots and virtual agents that communicate multimodally with humans. The research focuses on head-movements, and tracking technologies are used to detect the use of nods and shake by the human interlocutor. In "Digital chameleons", Baileson and Yee (2005) study the implementation of mirroring behaviors in artificial agents.

3. The Method

The collection of the corpus was done using the WoZ technique, which is broadly used in HCI and HRI. The data-collection of the present study was inspired by Han et al.'s (2012) work on using WoZ in the corpus collection for the Herme Database, as well as the corpus collection and annotation of the Nordic NOMCO corpora of dyadic first encounters (Paggio et al. 2010, Navarretta et al. 2011). In the present work, the WoZ method was used to simulate conversational behavior in the robot NAO when communicating with test participants.

3.1 The Wizard-of-Oz Method (WoZ)

The WoZ method is a technique for studying user behavior in HCI and HRI. It is a technique by which intelligence of a system is simulated allowing for testing of applications that have not yet been developed (Kelley, 1984). The user is usually not aware of the simulated status of the system. The method is not only used for incremented development of interaction systems but also for studying the behaviors of users in the field of HRI. Han et al. (2012) use the WoZ approach for collecting the Herme corpus, and Delaborde et al. (2009) use the method for building a corpus of interactions between children and a NAO robot.

Also the alignment of behaviors between a robot and human users has been investigated through WoZ experiments. For example, Xu et al. (2007) study human alignment behaviors with the purpose of creating cooperative robotic systems for the industries while Kouluri and Lauria (2009) investigate miscommunication in HRI. A more comprehensive review of HRI related use of the WoZ method in later years is provided by Riek (2012). In the review, Riek draws attention to the lack of a unified methodology in the use of the WoZ method and proposes a set of reporting guidelines. These guidelines were taken into account in the present work.

4. The Participants

Thirteen cognitively disabled and four cognitively normal adults aged from 33 to 60 years interacted with the robot. Eight of the disabled test participants were diagnosed with acquired deficiencies, while five were diagnosed with congenital or developmental disorders.

The recruitment of the disabled test participants was based on a minimum set of perceptive and expressive abilities as well as signs of interest expressed in a first short encounter with the robot and the two collectors of the corpus at the institution (one of the collectors is the first author of this article). These pre-visits were arranged to ethically ensure that only the residents who showed interest in meeting the robot would participate in the data collection. The pre-meetings consisted of a short demonstration of the robot, through a short monologue, a walk and a tai chi performance. The observed reactions of the residents also served as an indicator of what to expect from the interactions. It was clear, that possible test participants expected some level of physical act from the robot. This information inspired the design of the dialogue script with regard to the variation and timing of the dialogue components. Thus, the Tai chi performance and a small cooperative game with a red ball were implemented in the final script besides the sub-dialogues.

5. The NAO Robot and the Dialogue Script

The NAO robot (Figure 1) used in the data collection was loaned to the project by the Danish Technological Institute (DTI). Current studies in Denmark focus on the use of NAO as a didactic tool in schools and as a therapeutic tool in work with autistic children.

Figure 1: The NAO robot

The NAO robot was made available to the project for a two weeks period. During the interactions the robot was wirelessly connected to a laptop running *Choreographe*, which was used as the controlling interface by the Wizard.

A dialogue script was implemented for the interactions. The manuscript of the dialogue was inspired by that proposed in Han et al. (2012) for collecting the Herme Database involving Lego Mindstorms robots. The Herme manuscript comprises a mixture of chat- and task based elements embedded in a fixed-sequence script, and the Wizard must step through the utterances one by one, awaiting the response of the participants. This enables the Wizard to control the timing of the robot utterances, but not their sequencing. For the present project, it was also desirable to control the sequencing of the sub-dialogues partly because of the unknown nature of the reactions and abilities of the test participants, and partly because we wished to be able of resuming the interactions in case of technical failures.

The dialogue-script constructed for the present work consists of the following thematic sequences, and chat-based sub-dialogues: 1) "I'm awake", 2) "Hello", 3) "My name is..", 4a+4b) "Miscellaneous single utterances", 5) "Do you live here?", 6) "Tai chi", 7) "Red ball game", 8) "Tell me something", 9) "Goodbye". Differing from the script by Han et al. (2012) which only consisted of spoken utterances, the present dialogue script contains utterances of both speech and body behaviors. The dialogue language is Danish.

The dialogue script and other actions of the robot were implemented in the visual scripting interface of Choreographe, which also functioned as the interface for the remote control of the NAO robot. Script boxes in Choreographe may contain simple functions or larger scripts; furthermore scripts may be modified in Python or C++. The scripting boxes can be arranged in conjoined sequences, which can be automatically launched or can be arranged independently for manual activation.

6. Set-up and Collection Procedure

The WoZ-setting used for the corpus collection is shown in Figure 2.

Figure 2: WoZ setup with NAO robot

As the recordings of the data-collection were conducted in either the private apartments of the test participants or in a common room at the institution, it required mobility of the setup. This consisted of the NAO robot, a wireless network router, two iPads on tripods, and a laptop. The robot was placed on a table and in front of it was a chair for the test participants. The Wizard sat next to the NAO robot and controlled the predefined actions real time from the laptop. The two iPads on tripods served as cameras and were placed to capture the actions of the test participant. Two student-experimenters were present during the recordings. One experimenter (the first author of this paper) managed robot related settings and technical issues. She also played the secret role of the Wizard controlling the behaviors of the robot from the laptop. The other experimenter set up and controlled the iPad cameras and handed out written instructions and questionnaires to the test participants and their companions. The presence of the experimenters was necessary due to protection of both test- and robotic hardware in case of unforeseen events and insurance requirements. The known role of the experimenters during the interactions was that of being natural bystanders of the test situation.

Figure 3: A snapshot from the corpus: waving goodbye

Before each meeting the robot was arranged on a table in a sitting position and the test participant was then led to sit in front of it. A snapshot from the corpus is shown in Figure 3.

Once the test participant was seated, the Wizard padded the robot on the shoulder and told it to "wake up". The robot then "woke up" saying "Ah, I'm awake now". The interaction continued through the thematic sequences of the dialogue script. The basic structure of the interaction was: greetings, chat, the robot's "Tai Chi" demonstration, a small cooperative game with a small red ball, chat and farewell greetings.

This order was followed by the Wizard unless the intervention of the test participant caused it to deviate. The interaction had another course, for instance, if the test participant took the initiative and played the cooperative game more than once throughout the interaction. The Wizard made sure that each interaction came around all sequences at least once. Finally the Wizard told the robot to say goodbye, and the robot responded by waving to the test participant and telling him/her goodbye.

After the interaction the test participants were handed a questionnaire enquiring their experience with respect to the presence of the cameras, the experimenters, the companion, and the robot. The test participants were asked to specify the perceived influence of the robot, the experimenters and cameras respectively on a five-point rating scale ranging from "very affected" to "completely unaffected". Prior to each session, it was judged by the associated care worker whether the test participant was capable of assessing his own experiences and answering a questionnaire. If a test participant was not judged able to answer, the companion filled out the questionnaire for him/her.

7. The corpus

The resulting corpus contains 17 interactions consisting of a mix of thematic sequences of chat-based conversation and a cooperative game. The length and sequencing of the thematic sequences differ between subjects, yet all interactions contain at least one instance of each of the thematic sequences. The duration of the interactions varies between 5 and 15 minutes, for a total of approximately two hours and fifty minutes.

The analysis of the questionnaires shows that the participants had a neutral reaction to the presence of the experimenters in the majority of the cases. They were also mostly neutral or unaffected by the cameras (iPads) while they were affected or very affected by the interaction with the robot.

A preliminary qualitative analysis of body behaviors in the corpus shows that both cognitively disabled and cognitively normal test participants interacted with the robot with both verbal and non-verbal behaviors. We are now annotating multimodal feedback and mirroring behaviors with an extension of the MUMIN scheme.

8. Conclusion

In the paper, we described the collection of a multimodal corpus of the conversational interactions of cognitively disabled and normal participants with the anthropomorphic robot NAO focusing on the collection methodology. A WoZ technique was used in combination with a modular dialogue script which allowed the wizard to adapt the behavior of the NAO robot to that of the test participants. This approach

proved successful in engaging both groups of test participants in chat-based conversational interactions. The modular arrangement of the dialogue script, with flexible sequencing of sub-dialogues, tai chi sequences and a ball game, was efficient for resuming interactions when interrupted and ensuring the caption of user behaviors during specific thematic sequences. Interruptions were mainly caused by technical problems. Deviations from the sequencing of the original script were mainly due to the varying behaviors of the test participants.

The answers to a questionnaire about the reaction of the test participants to the presence of the experimenters, the cameras and the robot indicate that they were neutral with respect to the presence of the experimenters and the camera, but they were strongly affected by the meeting with the robot. A first analysis of the video-recordings shows that both disabled and non-disabled participants interacted actively and multimodally (speech and body behaviors) with the robot. Further analysis must be done to extract the actual patterns of communicative behaviors in the HRI relation, and possible differences between the various groups of participants. Such analysis could provide information on the nature of the differences in communicative abilities of the disabled participants and could and be used to model multimodal behaviors in dialogic interactions of social assistive robots. We have thus started annotating specific feedback behaviors following an extension of the MUMIN annotation scheme.

9. Acknowledgements

The corpus was built up as part of a Master thesis in the IT & Cognition program at the University of Copenhagen. We want to acknowledge Mikael Lockert who was also involved in the collection of the corpus (the second experimenter). Finally, we want to thank the residents and staff of the Centre for Multiple Disabilities of Copenhagen (CMF) and the robotics department at the Danish Technological Institute (DTI) for their collaboration.

10. References

Allwood, J. 2002. Bodily Communication - Dimensions of Expression and Content. *Multimodality in Language and Speech Systems*. Björn Granström, David House and Inger Karlsson (Eds.). Dordrecht: Kluwer Academic Publishers, pp. 7-26.

Allwood,J., Cerrato, L. Jokinen, K., Navarretta, C. and Paggio, P. 2007. The MUMIN coding scheme for the annotation of feedback in multimodal corpora: a prerequisite for behavior simulation. In Language Resources and Evaluation. Special Issue. J.-C. Martin et al. (Eds.) *Multimodal Corpora for Modeling Human Multimodal Behavior,* Volume 41, Nr. 3-4:273-287, Springer.

Al Moubayed, S., Baklouti, M., Chetouani, M., Dutoit, T., Mahdhaoui, A., Martin, J.-C., Ondas, S., Pelachaud, C., Urbain, J., Yilmaz, M., 2009. Generating robot/agent backchannels during a storytelling experiment, in: Robotics and Automation, 2009. ICRA'09. IEEE International Conference on. pp. 3749–3754.

Bailenson, J.N., Yee, N. 2005. Digital chameleons automatic assimilation of nonverbal gestures in immersive virtual environments. Psychological science 16, 814–819.

Delaborde, A., Tahon,M. Barras, C. and Devillers, L. A . 2009 Wizard-of-Oz Game for Collecting Emotional Audio Data in a Children-Robot Interaction. In Proceedings of the International Workshop on Affective-aware Virtual Agents and Social Robots, ICMI-MLMI, Boston, USA.

Han, J. G., Gilmartin, E., De Looze, C., Vaughan, B., & Campbell, N. 2012. Speech and multimodal resources-The Herme database of spontaneous multimodal human-robot dialogues. In *Proceedings of LREC* (pp. 1328-1331).

Kelley, J.F., 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst*. TOIS 2, 26–41.

Koulouri, T., Lauria, S., 2009. A WOz framework for exploring miscommunication in HRI, in: *Procs. of the AISB Symposium on New Frontiers in Human-Robot Interaction*. pp. 1–8.

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

Navarretta, C., Ahlsén, E., Allwood, J., Jokinen, K., Paggio, P. 2011. Creating Comparable Multimodal Corpora for Nordic Languages. In *Proceedings of the 18th NODALIDA*. Riga, Latvia, May 11-1, pp. 153-160.

Paggio, P., Allwood, J. ,Ahlsén, E. Jokinen, J., Navarretta, C. 2010. The NOMCO multimodal Nordic resource - goals and characteristics. In *Proceedings of LREC 2010*, Malta, May 17-23, pp. 2968-2973.

Paggio, P. and Navarretta, C. 2013. Head movements, facial expressions and feedback in conversations - Empirical evidence from Danish multimodal data. *Journal on Multimodal User Interfaces - Special Issue on Multimodal Corpora*, Springer Verlag, Volume 7, Issue 1-2, pp. 29-37.

Rehm, M., Nakano, Y., André, E., Nishida, T., Bee, N., Endrass, B., Wissner, M., Lipi, A.A., Huang, H.-H. 2009. From Observation to Simulation — Generating Culture Specific Behavior for Interactive Systems. *AI & Society* 24, 267–280 .

Riek, L., 2012. Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. J. Hum.-Robot Interact. 119–136.

Xu, Y., Ueda, K., Komatsu, T., Okadome, T., Hattori, T., Sumi, Y., Nishida, T., 2007. WOZ experiments for understanding mutual adaptation. AI Soc. 23, 201–212.

Catching wind of multiparty conversation

Jens Edlund*, Mattias Heldner†, Marcin Włodarczak†

*KTH Speech, Music and Hearing, †Department of Linguistics, Stockholm University, Sweden
edlund@speech.kth.se, {heldner, wlodarczak}@ling.su.se

Abstract

The paper describes the design of a novel corpus of respiratory activity in spontaneous multiparty face-to-face conversations in Swedish. The corpus is collected with the primary goal of investigating the role of breathing for interactive control of interaction. Physiological correlates of breathing are captured by means of respiratory belts, which measure changes in cross sectional area of the rib cage and the abdomen. Additionally, auditory and visual cues of breathing are recorded in parallel to the actual conversations. The corpus allows studying respiratory mechanisms underlying organisation of spontaneous communication, especially in connection with turn management. As such, it is a valuable resource both for fundamental research and speech techonology applications.

Keywords: breathing, multiparty conversation, turn-taking, respiratory inductance plethysmography, physiological measurements

1. Introduction

Even though we may not be aware of it, much breathing in dialogue is both clearly audible and visible. Consequently, it has been suggested that respiration is used in coordination of dialogue flow (Schegloff, 1996; Local and Kelly, 1986), e.g. by indicating intention to take or release a speaking turn. As a result, breathing is likely to provide a more direct access to speaker's communicative intentions than is otherwise available. However, few studies addressed interactional aspects of breathing. While notable exceptions exist, for instance (McFarland, 2001; Winkworth et al., 1995), even those studies were based on interactions which were not entirely spontaneous. In addition, no account exists of breathing in dialogue between more than two speakers, which is likely to show a greater range of respiratory patterns due to increased turn management complexity.

These omissions are particularly glaring given the potential relevance of breathing to speech technology application. As dialogue turns are normally preceded by deep and easily perceivable inhalations and followed by marked exhalations, presence of breathing noises could be used to improve turn management strategies implemented in the state-of-the-art dialogue systems. For instance, loud inhalations during system output could be used to detect user interruptions *prior to* the actual speech onset. Likewise, identification of post-completion exhalations should reduce the number of pause interruptions, which are a major problem in current speech technology applications.

Consequently, studying breathing in conversation is highly relevant from the point of view of both fundamental and applied research. On the one hand, it contributes significantly to the understanding of physiological constraints driving speech production and organisation of human interaction. On the other hand, it informs computational models of human interaction and paves the way towards more human-like embodied conversational agents capable of using previously unavailable cues.

Motivated by these goals, we have begun collection of a multimodal corpus of spontaneous multiparty conversations which includes physiological measurements relevant to breathing. Below we outline the recording setup and briefly discuss possible applications of the corpus.

2. Data acquisition setup

The recordings take place at the Phonetics Laboratory, Stockholm University in a quiet, sound-treated room. As it was observed that a standing position minimises noise in the respiratory signal due to body movement, subjects are recorded standing at a table 95 cm in height. The recording setup is shown in Figure 1.

Respiratory activity is measured using using respiratory inductance plethysmography (Watson, 1980), which quantifies changes in rib cage and abdominal cross sectional area by means of two elastic transducer belts (Ambu RIPmate) placed at the level of the armpits and the navel, respectively. Contributions of individual belts to the net lung volume change are estimated using isovolume manoeuvres (Konno and Mead, 1967).

The belts are connected to a dedicated respiratory belt processor (RespTrack, Figure 2) designed and built in the Phonetics laboratory at Stockholm University. The RespTrack

Figure 1: Recording setup. The white boxes are earlier prototypes of our respiratory belt processors.

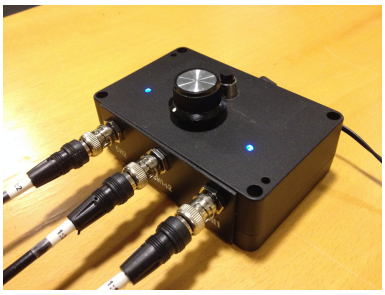

Figure 2: The second prototype of the RespTrack processor.

processor was designed for ease of use, and optimised for low noise recordings of respiratory movements in speech and singing. In particular, DC offset can be corrected simultaneously for the rib cage and abdomen belts using a ”zero” button. Unlike in the processors supplied with the belts, there is no high-pass filter, thus the amplitude will not decay during for example breath-holding. A potentiometer allows the signals from the rib cage and abdomen belts to be weighted so that they give the same output for a given volume of air, as well as for the summed signal, enabling direct estimation of lung volume change (see Figure 3).

The signal is collected by an integrated physiological data acquisition system (PowerLab by ADInstruments), which also allows connecting other measuring instruments, such as air-flow masks or electroglottographs. A sample signal is presented in Figure 3.

High-quality audio is recorded with close-talking directional microphones (Sennheiser HSP 4), and video is captured by GoPro Hero3+ cameras.

We plan to expand the setup by including contact microphones attached to speakers' necks (throat microphones) with a view to obtaining clearer recordings of inhalation and exhalation noises. Additionally, we will use thermal probes placed in the nostril to be able to distinguish nasal and mouth breathing. All these extensions are fully compatible with our current recording setup and will be presented during the workshop.

Minimally, the corpus will be annotated with interactional events derived from voice activity detection, as well as (semi-)automatically detected inhalation and exhalation events in the respiratory data.

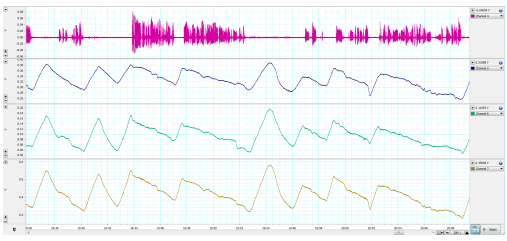

Figure 3: Speech recording (channel 1) and respiratory measurements from rib-cage and abdomen belts (channels 2-3) for one speaker. The bottom channel shows the weighted sum of the two belts.

3. Possible applications of the corpus

Our primary rationale for the corpus collection is studying the role of breathing in turn taking. Above all, it will allow a stringent quantitative investigation of previously untested claims made in literature, for instance about the role of inhalations as an interactionally salient cue to speech initiation, exhalations as a turn yielding device and breath holding as a marker of turn incompleteness. Furthermore, detection of pre-speech inhalations should allow to infer speaker's *intention* to initiate a new turn, whether or not this intention is realised or abandoned. Thus, respiratory data will also shed light on “hidden” events in dialogue, which are otherwise unavailable for analysis.

Furthermore, the corpus could serve as a test bed for computational models of turn-taking. In particular, the combination of physiological measurements with audio recordings of respiratory noises will provide valuable training data for automatic detection and classification of interactionally salient breathing.

Last but not least, given scarcity of corpora of spontaneous multiparty interactions, it is expected that the corpus will be a valuable resource for many other dialogue studies not necessarily related to studying respiration. We plan to make the corpus available for research use.

Acknowledgements

The research presented here was funded in part by the Swedish Research Council project 2009-1766 *Samtalets rytm* (*The Rhythm of Conversation*).

4. References

Kimio Konno and Jere Mead. 1967. Measurement of the separate volume changes of rib cage and abdomen during breathing. *Journal of Applied Physiology*, 22(3):407–422.

John Local and John Kelly. 1986. Projection and ‘silences’: Notes on phonetic and conversational structure. *Human studies*, 9(2):185–204.

David H. McFarland. 2001. Respiratory markers of conversational interaction. *Journal of Speech, Language and Hearing Research*, 44(1):128–143.

Emanuel A. Schegloff. 1996. Turn organization: One intersection of grammar and interaction. *Studies in Interactional Sociolinguistics*, 13:52–133.

H. Watson. 1980. The technology of respiratory inductive plethysmography. In F. D. Stott, E. B. Raftery, and L. Goulding, editors, *Proceeding of the Second International Symposium on Ambulatory Monitoring (ISAM 1979)*, London. Academic Press.

Alison L. Winkworth, Pamela J Davis, Roger D. Adams, and Elizabeth Ellis. 1995. Breathing patterns during spontaneous speech. *Journal of Speech, Language and Hearing Research*, 38(1):124–144.

Resources for Analyzing Productivity in Group Interactions

Gabriel Murray

University of the Fraser Valley
Abbotsford, BC, Canada
gabriel.murray@ufv.ca

Abstract

Productivity can vary both within and across meetings. In this work, we consider the question of how to measure productivity, and survey some of the available and potential resources that correspond to productivity. We then describe an initial experiment in which we define productivity in terms of the percentage of sentences from a meeting that are considered summary-worthy. Given that simple definition of productivity, we fit a logistic regression model to predict productivity levels of meetings using linguistic and structural features.

Keywords: productivity, multimodal interaction, extractive summarization

1. Introduction

How can we measure *productivity* in group interactions? In the absence of gold-standard annotations for productivity, we can begin by defining productivity within the context of an automatic summarization task. If we employ *extractive* techniques to summarize a meeting by labeling a subset of dialogue acts from the meeting as important, then productive meetings would seem to be ones that have a high percentage of important, summary-worthy dialogue acts, while unproductive meetings would have a low percentage of such important dialogue acts.

Starting with that simple definition of productivity, we can see that productivity is indeed a critical issue in meetings, and that meetings differ in how productive they are. Using gold-standard extractive summaries of the AMI and ICSI corpora (to be described later), we can index the extracted dialogue acts by their position in the meeting and see from Figure 1 that important dialogue acts are more likely to occur at the beginning of meetings and are less likely at the end of meetings. This suggests that many meetings decrease in productivity as they go on. Figure 2 shows that productivity also varies *between* meetings, e.g. longer meetings tend to have a smaller percentage of summary dialogue acts.

This paper discusses our corpora and initial experiments for analyzing meeting productivity. In Section 2. we discuss related work. In Section 3. we describe the two corpora we are currently using in terms of their available resources that relate to productivity, as well as potential new resources. In Section 4. we describe an experiment where productivity is defined in relation to an extractive summarization task. Section 5. gives the results of that first experiment, and we conclude in Section 6.

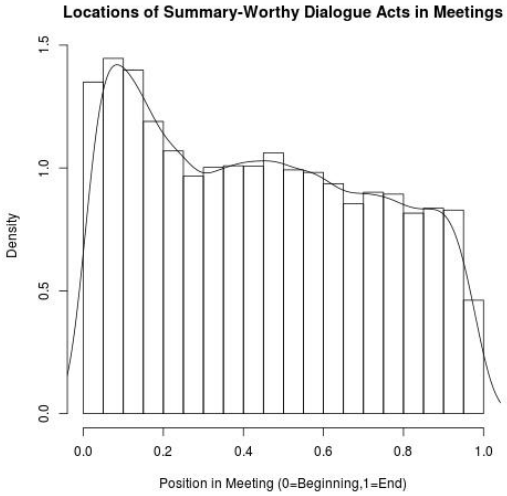

Figure 1: Histogram/KDE of Extractive Locations

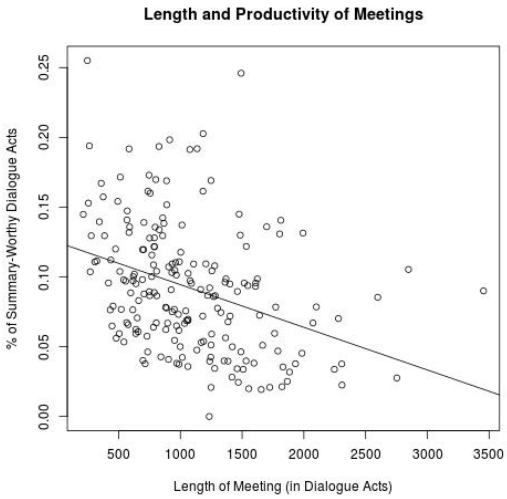

Figure 2: Length and Productivity of Meetings

2. Related Work

This work closely relates to meeting summarization, including *extractive* (Zechner, 2002; Murray et al., 2005; Galley, 2006) and *abstractive* (Kleinbauer et al., 2007; Murray et al., 2010) approaches. Carenini et al (2011) provide a survey of techniques for summarizing conversational data. This work also relates to the task of identify-

ing action items in meetings (Purver et al., 2007; Murray and Renals, 2008; Morgan et al., 2006), detecting decision points (Hsueh et al., 2007; Fernández et al., 2008; Bui et al., 2009), and detecting speaker dominance (Rienks and Heylen, 2005; Rienks et al., 2006; Jayagopi et al., 2009). Renals et al (2012) provide a survey of various work that has been done analyzing multimodal interactions, while op den Akker et al (2012) give a survey of research investigating conversational dynamics in meetings.

3. Corpora

In analyzing meeting productivity, we use both the AMI (Carletta et al., 2005; Carletta, 2006) and ICSI (Janin et al., 2003) meeting corpora. These corpora each include audio-video records of multi-party meetings, as well as both manual and speech recognition transcripts of the meeting discussions. The main difference between the two corpora is that the AMI meetings are scenario-based, with participants who are role-playing as members of a fictitious company, while the ICSI corpora features natural meetings of real research groups.

3.1. Available Resources

As part of the AMI project on studying multi-modal interaction (Renals et al., 2012), both meeting corpora were annotated with extractive and abstractive summaries, including many-to-many links between abstractive sentences and extractive dialogue acts. We use these gold-standard summary annotations in the following experiment.

Available resources that are *not* used in the experiment described here, but which may end up being useful in follow-up work, include:

- **Decision items:** Some dialogue acts are linked to the "decisions" portion of the abstractive summary; these constitute a specific subset of the summary dialogue acts we use. For example, a meeting may be considered more productive if it contains more decision items.
- **Action items:** Similarly, some dialogue acts are linked to the "action items" portion of the abstractive summary.
- **Dominance** annotation of Rienks and Heylen (2005). As described in Section 4., we do utilize simple features relating to dominance.
- **Participant Summaries:** In the AMI corpus, meeting participants individually authored short summaries after each meeting. These may yield clues on how productive or efficient they perceived the meetings to be.
- **Subjectivity Annotation:** AMI meetings have been annotated for various subjective criteria (Wilson, 2008). This could be useful if it turned out that less productive meetings are more associated with negative subjectivity, to give one example.

3.2. Proposed Resources

While the experiment described below in Section 4. shows promising results using extractive annotations as a proxy for productivity, we may need to more directly address the issue by directly annotating for the phenomenon. This could include:

- **Meeting-Level Annotation:** Meetings could be either categorized (e.g. unproductive, productive) or rated on a scale of productivity (e.g. 1-10). This would be the least costly and time-consuming annotation, and least difficult for the human judges.
- **Dialogue Act-Level Annotation:** Individual dialogue acts could be rated on how much they contributed to meeting productivity. We expect that this could be difficult for many ambiguous dialogue acts, and would likely be a challenging, time-consuming task for human judges.
- **Turn-Level Annotation:** If we define a turn as a sequence of dialogue acts by the same speaker, then each turn could be rated on how much it contributed to meeting productivity. This would be less costly and time-consuming than labeling of every dialogue act.
- **Participant-Level Annotation:** Each participant in the meeting could be rated on how much of a contribution they made to meeting productivity. In conjunction with one or more of the other proposed annotations above, we could learn how individual participants affect the productivity outcome of a meeting.

4. Predicting Productivity of Meetings

In this initial experiment, the task is to predict the overall productivity of a meeting, given some linguistic and structural features of the meeting. The productivity is measured as the percentage of meeting dialogue acts labeled as summary-worthy. That is, we are predicting a value between 0 and 1. For that reason, we employ logistic regression for this task.

Logistic regression is well-known in natural language processing, but is usually used in cases where there are dichotomous (0/1) outcomes, e.g. in classifying dialogue acts as extractive or non-extractive (Murray and Carenini, 2008). Unfortunately, we do not have gold-standard labeling of meetings indicating that they were productive or non-productive. However, logistic regression can also be used in cases where each record has some associated numbers of successes and failures, and the dependent variable is then a proportion or percentage of successes. That is our case here, where each meeting has some number of extractive dialogue acts ("successes") and some remaining non-extractive dialogue acts ("failures").

For this task, the meeting-level features we use are described below, with abbreviations for later reference. We group them into feature categories, beginning with **term-weight (tf.idf)** features:

- **tfidfSum** The sum of $tf.idf$ term scores in the meeting.

- **tfidfAve** The average of $tf.idf$ term scores in the meeting.
- **conCoh** The conversation cohesion, as measured by calculating the cosine similarity between all adjacent pairs of dialogue acts, and averaging. Each dialogue act is represented as a vector of $tf.idf$ scores.

Next are the features relating to meeting and dialogue act **length**:

- **aveDALength** The average length of dialogue acts in the meeting.
- **shortDAs** The number of dialogue acts in the meeting shorter than 6 words.
- **longDAs** The number of dialogue acts in the meeting longer than 15 words.
- **countDA** The number of dialogue acts in the meeting.
- **wordTypes** The number of unique word types in the meeting (as opposed to word tokens).

There are several **entropy** features. If s is a string of words, and N is the number of words types in s, M is the number of word tokens in s, and x_i is a word type in s, then the word entropy $went$ of s is:

$$went(s) = \frac{\sum_{i=1}^{N} p(x_i) \cdot -\log(p(x_i))}{(\frac{1}{N} \cdot -\log(\frac{1}{N})) \cdot M}$$

where $p(x_i)$ is the probability of the word based on its normalized frequency in the string. Note that word entropy essentially captures information about type-token ratios. For example, if each word token in the string was a unique type then the word entropy score would be 1. Given that definition of entropy, the derived **entropy** features are:

- **docEnt** The word entropy of the entire meeting.
- **speakEnt** This is the speaker entropy, essentially using speaker ID's instead of words. The speaker entropy would be 1 if every dialogue act were uttered by a unique speaker. It would be close to 0 if one speaker were very dominant.
- **speakEntF100** The speaker entropy for the first 100 dialogue acts of the meeting, measuring whether one person was dominant at the start of the meeting.
- **speakEntL100** The speaker entropy for the last 100 dialogue acts of the meeting, measuring whether one person was dominant at the end of the meeting.
- **domSpeak** Another measure of speaker dominance, this is calculated as the percentage of total meeting DA's uttered by the most dominant speaker.

We have one feature relating to **disfluencies**:

- **filledPauses** The number of filled pauses in the meeting, as a percentage of the total word tokens. A filled pause is a word such as um, uh, erm or $mm - hmm$.

Finally, we use two features relating to **subjectivity / sentiment**. These features rely on a sentiment lexicon provided by the SO-Cal sentiment tool (Taboada et al., 2011).

- **posWords** The number of positive words in the meeting.
- **negWords** The number of negative words in the meeting.

5. Experimental Results

For this experiment, we evaluate the fitted model primarily in terms of the *deviance*. The deviance is -2 times the log likelihood:

$$Deviance(\theta) = -2\ log[\ p(y|\theta)\ ]$$

A lower deviance indicates a better-fitting model. Adding a random noise predictor should decrease the deviance by about 1, on average, and so adding an informative predictor should decrease the deviance by more than 1. And adding k informative predictors should decrease the deviance by more than k.

Feature	Deviance
null (intercept)	4029.7
tfidfSum	3680.3
tfidfAve	3792.8
conCoh	3825.1
aveDALength	4029.7
shortDAs	3690.7
longDAs	3705.9
countDA	3637.8
wordTypes	3599.4
docEnt	3652.3
domSpeak	3575.2
speakEnt	3882.6
speakEntF100	3758.9
speakEntL100	3825.8
filledPauses	3986.9
posWords	3679.2
negWords	3612.5
COMBINED-FEAS	**2843.7**

Table 1: Deviance Using Single and Combined Predictors

Table 1 shows the deviance scores when using a baseline model (the "null" deviance, using just a constant intercept term), when using individual predictor models, and when using a combined predictor model. We see that the combined model has a much lower deviance (2843.7) compared with the null deviance (4029.7). Using 16 predictors, we expected a decrease of greater than 16 in the deviance, and in fact the decrease is 1186. We can see that the individual predictors with the largest decreases in deviance are *wordTypes*, *domSpeak*, and *negWords*.

6. Conclusion

Using the percentage of extracted dialogue acts as a proxy for a meeting's productivity, we have shown that a logistic regression model can predict productivity effectively based on linguistic and structural features. In our ensuing work, we plan to leverage available resources such as *dominance* and *sentiment* annotations, as well as participant summaries. We will also begin meeting-level annotations of productivity in order to more directly study this phenomenon.

7. References

T. Bui, M. Frampton, J. Dowding, and S. Peters. 2009. Extracting decisions from multi-party dialogue using directed graphical models and semantic similarity. In *Proceedings of the SIGDIAL 2009, London, UK*.

G. Carenini, G. Murray, and R. Ng. 2011. *Methods for Mining and Summarizing Text Conversations*. Morgan Claypool, San Rafael, CA, USA, 1st edition.

J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. 2005. The AMI meeting corpus: A pre-announcement. In *Proc. of MLMI 2005, Edinburgh, UK*, pages 28–39.

J. Carletta. 2006. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. In *Proc. of LREC 2006, Genoa, Italy*, pages 181–190.

R. Fernández, M. Frampton, P. Ehlen, M. Purver, and S. Peters. 2008. Modelling and detecting decisions in multi-party dialogue. In *Proc. of the 2008 SIGdial Workshop on Discourse and Dialogue, Columbus, OH, USA*.

M. Galley. 2006. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proc. of EMNLP 2006, Sydney, Australia*, pages 364–372.

P-Y. Hsueh, J. Kilgour, J. Carletta, J. Moore, and S. Renals. 2007. Automatic decision detection in meeting speech. In *Proc. of MLMI 2007, Brno, Czech Republic*.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI meeting corpus. In *Proc. of IEEE ICASSP 2003, Hong Kong, China*, pages 364–367.

D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. 2009. Modeling dominance in group conversations from non-verbal activity cues. *IEEE Transactions on Audio, Speech and Language Processing*, 17(3):501–513.

T. Kleinbauer, S. Becker, and T. Becker. 2007. Indicative abstractive summaries of meetings. In *Proc. of MLMI 2007, Brno, Czech Republic*, page poster.

W. Morgan, P-C. Chang, S. Gupta, and J. Brenier. 2006. Automatically detecting action items in audio meeting recordings. In *Proc. of the 7th SIGdial Workshop on Discourse and Dialogue*.

G. Murray and G. Carenini. 2008. Summarizing spoken and written conversations. In *Proc. of EMNLP 2008, Honolulu, HI, USA*.

G. Murray and S. Renals. 2008. Detecting action items in meetings. In *Proc. of MLMI 2008, Utrecht, the Netherlands*.

G. Murray, S. Renals, and J. Carletta. 2005. Extractive summarization of meeting recordings. In *Proc. of Interspeech 2005, Lisbon, Portugal*, pages 593–596.

G. Murray, G. Carenini, and R. Ng. 2010. Generating and validating abstracts of meeting conversations: a user study. In *Proc. of INLG 2010, Dublin, Ireland*.

R. op den Akker, D. Gatica-Perez, and D. Heylen. 2012. Multi-modal analysis of small-group conversational dynamics. In S. Renals, H. Bourlard, J. Carletta, and A. Popescu-Belis, editors, *Multimodal Signal Processing*, pages 155–169. Cambridge University Press, New York, June.

M. Purver, J. Dowding, J. Niekrasz, P. Ehlen, and S. Noorbaloochi. 2007. Detecting and summarizing action items in multi-party dialogue. In *Proc. of the 9th SIGdial Workshop on Discourse and Dialogue, Antwerp, Belgium*.

S. Renals, H. Bourlard, J. Carletta, and A. Popescu-Belis. 2012. *Multimodal Signal Processing: Human Interactions in Meetings*. Cambridge University Press, New York, NY, USA, 1st edition.

R. Rienks and D. Heylen. 2005. Automatic dominance detection in meetings using easily obtainable features. In *Proc. of MLMI 2005, Edinburgh, UK*.

R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post. 2006. Detection and application of influence rankings in small group meetings. In *Proc. of ICMI 2006, Banff, Canada*.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, June.

T. Wilson. 2008. Annotating subjective content in meetings. In *Proc. of LREC 2008, Marrakech, Morocco*.

K. Zechner. 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485.

Dynamic stimuli visualization for experimental studies of body language

Nesrine Fourati, Jing Huang, Catherine Pelachaud

Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI
37/39 rue Dareau, 75014 Paris
nesrine.fourati@telecom-paristech.fr, jing.huang@telecom-paristech.fr, catherine.pelachaud@telecom-paristech.fr

Abstract

Understanding human body behavior have relied on perceptive studies. Lately experimental studies have been conducted with virtual avatars that reproduce human body movements. The visualization of human body behaviors stimuli using avatars may introduce bias for human perception comprehension. Indeed, the choice of the camera trajectory and orientation affects the display of the stimuli. In this paper, we propose control functions for the virtual camera.

Keywords: virtual camera, camera motion, camera position, perception

1. Introduction

The studies of the human perception of body language and motion patterns received a wide range of interest since a long time for different fields of research like the recognition of affect in body movement (Kleinsmith et al., 2011) and the identification of body cues that contribute to the attribution of emotions and affects (Meijer, 1989; Dahl and Friberg, 2007).

The content of the stimuli that the observers are asked to judge depends on the research question that needed to be answered from the results the perception based study. One can use the raw videos (videotaped) that depict the real visual content of body movement of the "actors" (Meijer, 1989). Digital modifications can also be done on the original movies or pictures to abstract some bodily information (Dahl and Friberg, 2007; Atkinson et al., 2004). For many other purposes, it is required or preferable to use computer avatar as the content of the stimulus to be shown to the observers for the perception based study.

To visualize stimuli of an avatar wandering around an environment (walking, turning, etc) we can choose to have a static or a dynamic camera that follows the avatar in its displacement. However when the camera is static the distance between the avatar and the camera varies and this may affect the perception of the avatar body movement. Identically if the orientation of the camera and the avatar body varies, it may affect how body movement is perceived. To overcome such biases in perception studies we propose tools to parameterize the camera movement and orientation. For example we can control the trajectory of the camera and its orientation so that it maintains an equal distance and orientation with the avatar.

2. Related work

The use of computer avatars in body movement perception based studies based on body movement has widely emerged recently (Coulson, 2004; Hicheur et al., 2013; Kleinsmith et al., 2011; Roether et al., 2009). Depending on the goal of the study, the movements reproduced with the avatar may be the result of motion capture data (Kleinsmith et al., 2011) or the results of a model that provides the synthesize of new body movements (Hicheur et al., 2013).

Previous discussions related to the body movement perception based studies were mostly around the body model of the computer avatar. Point-light display of body movements was primarily used for the studies related to the perception of biological motion (Johansson, 1973; Dekeyser et al., 2002). Other body models were used for the studies that rely on the perception of both body posture and the dynamic of movement. Those models are mostly based on body skeleton model through specific geometric shape primitives (Griffin et al., 2013; Kleinsmith et al., 2011; Kleinsmith et al., 2006; Roether et al., 2009) or a virtual animated character (Hicheur et al., 2013).

As body posture involves a three-dimentional presence, human perception of body postures and body movements reproduced on a three-dimentional avatar may be depending on the viewing angle (Coulson, 2004; Daems and Verfaillie, 1999), especially for viewpoint that result in occlusion of some body parts by others. In the studies based on the perception of body movement, the viewpoint is defined according to the goal of the study. Kleinsmith et al. (Kleinsmith et al., 2011) reproduced expressive postures on computed avatar rotated to simulate a frontal view for the perception of emotion from body posture. Hicheur et al (Hicheur et al., 2013) chose a side viewpoint to create the videos depicting walking behaviors reproduced on an animated character. Roether et al. (Roether et al., 2009) used movies of a animated virtual avatars turned 20 degrees from the frontal view. However, it could be interesting to study the effect of different viewpoint on the perception of body behavior (Coulson, 2004).

3. The description of the proposed approaches

Two different types of virtual camera must be distinguished: free camera and target camera. While the orientation of free camera requires the definition of the 3D rotation, target camera is automatically oriented toward its target. Most often, the target refers to the center of interest of the object to be followed (the avatar). We assign the target of the camera to the pelvis in order to perceive the whole body posture, but the choice of the joint associated with the target could change from one study to another.

3.1. The position and the orientation of the camera

The definition of the viewpoint of the avatar refers to the determination of the position of a virtual camera that looks toward the avatar.
The viewpoint determined by the virtual camera has to be defined based on the orientation of the object (here the avatar body). The orientation of the whole body refers to the direction of the boy displacement in the space . We define the orientation of the whole body based on the orientation of the pelvis.

3.1.1. The position of the camera

The desired viewpoint of the avatar may differ from one study to another. Our goal is to provide a solution that can be controlled through a set of parameters. The set of the position of the camera regarding the avatar involves: the distance between the camera and the target, the height of the camera, and the angle that defines the viewpoint of the avatar. By default the height of the camera and the distance between the camera and the target could be proportional to the height of the pelvis. As a result, the attribution of the desired viewpoint relies on the determination of X and Y components of the camera position.
The determination of the camera position turn out to be a geometric problem that involves both the vector orthogonal to the direction of the whole body and the vector between the target and the camera position. When considering the pelvis posture as the indication of the body orientation, the geometric problem involves the vector defined with the Left Hip Position and the Right Hip Position and the vector defined with the Pelvis Position and the Camera Position. Knowing the positions of Right Hip and Pelvis, the distance between the pelvis and the camera, and the angle that defines the angle of viewpoint, we are able to determine the position of the camera.

3.1.2. The orientation of the camera

The target of the camera is used to define the orientation of the camera toward an object. Assigning the target to the pelvis position makes the camera pointing to the center of the body structure. However, defining the target as the pelvis itself could affect the perception of pelvis motion. In fact, a target camera will not only be oriented toward its target, but it will follow also (without changing the position) all the motions performed by its target, including the more subtle motions. For instance, if the avatar is jumping up and down, the camera motion will follow the same motion (up and down). As a result, in the related video, we will perceive the floor as a moving object and the pelvis as a static object, which is the opposite of the result that we are expecting. For this reason, we define the target as an approximation of the pelvis position. For body movements that involve small body displacement (where the avatar can be still visible to the camera), the target position can be set to the first position of the pelvis, and still static for the whole animation. However, for body movements that involve considerable body displacement in the space, the target has to move according to the pelvis motion. In the next section, we introduce some solutions for the motion of the camera as well the target following the avatar motion.

3.2. The control of virtual camera motion

Up to our knowledge, previous perception based studies that rely on the perception of body movements tend to use movies where the viewpoint as well as the position of the camera is static while the avatar is moving in the 3D space. While this approach could be a good solution when the whole body movement is relatively small, it has the limitation of loosing the details of body motion during the perception if the animation involves turning behavior or walking along long distance. In this section, we introduce some solutions for the control of the camera trajectory and the target motion when the animation involves a considerable displacement of the whole body in space.
One principal issue that could affect our perception of body movements is the desynchronisation of the camera motion with the avatar displacement, which creates an effect of zoom in and out. Another issue that can create the same effect is the change of the distance between the camera and the target. So the first motivation for the solutions that we propose to control the camera path is the non-uniform motion of the camera following as much as possible the same change of velocity and acceleration in the avatar displacement. And the first motivation for the solution proposed to control the path of the target is to keep as much as possible the same distance between the camera and the avatar.

3.2.1. The path of the target

As we explained previously, the target position has to follow an approximation of the targeted joint. We project the pelvis positions along straight lines defined through the positions of the pelvis in two successive time steps. In this way, the target motion pattern is the same as the camera motion pattern shown in Figure 2 (1) but translated to the real positions of the pelvis.

3.2.2. The path of the camera

For perception based studies, there is a lack of discussions on the control of the path of virtual camera. Thus, we based our work on the assumption that the virtual camera motion can influence the perception of body movements. Our aim is to create camera with less potential influence on the perception of body movements.
A first intuitive solution is the update of the camera position in each frame based on the algorithm described in the previous section (See Figure 1 (1)). This method results in a perfect synchronisation between the motion of the avatar and the camera, while handling the same viewpoint during the whole animation (based on the angle between Left Hip - Right Hip vector and Pelvis-camera vector). The algorithm that controls the path of the camera is; for each frame, we get the positions of Right Hip and Pelvis and we update the position of the camera according to their current positions. However, one should bear in mind that walking motion give rise to non linear pattern of body segments, including the pelvis (Fourati and Pelachaud, 2013; Olivier et al., 2009). Hence, this camera motion may affect the perception of body movements since the camera is shaking from the left to the right due to the non linear motion of the pelvis.
In the following, we propose some different solutions for

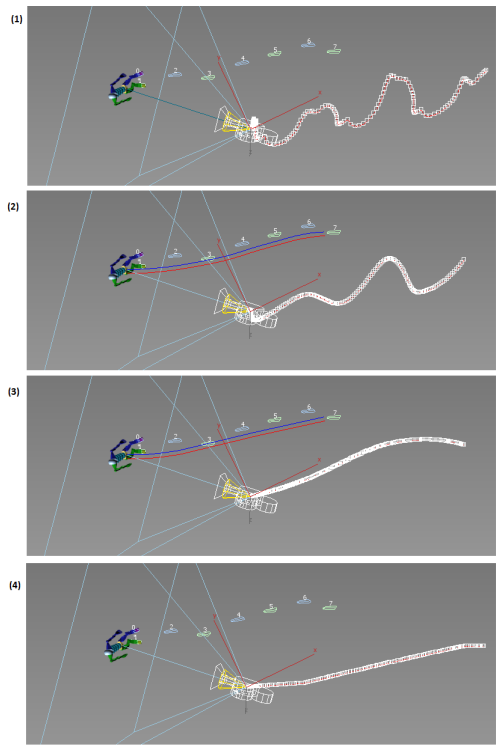

Figure 1: Camera trajectories when 1) Updating the camera position each frame based on the new pelvis and right hip positions, 2) Updating the camera position each frame based on the approximation of Pelvis and Right Hip motion through spline curve where the control points are the time steps, 3) Updating the camera position each frame based on the approximation of Pelvis and Right Hip motion through spline curve where the control points are the timing of all the right steps, 4) Defining the trajectory of the camera by the translation of the projection of the pelvis position along the windows of time defining through time steps

the control of virtual camera path according to the avatar movement.

- Synchronized non uniform non linear style: Update the camera position in each frame based on the approximation of pelvis motion

  One solution to control the motion of the camera is to approximate first the trajectory of pelvis and the right hip (or in more general way the joints that determine the position of the camera) and then to update the position of the camera (as explained in section 3.1.2.) at each frame. In this way, the viewpoint is updated at each frame according to the approximation of pelvis posture for straight walk and turning behavior (See Figure 2).

  The approximation of pelvis and right hip positions is set on a spline curve (the red and blue curves in Figure 1 (2) and (3)). However, this approximation is strongly based on the control points. Defining the control points along a fixed time window (for example each 30 frames) or along the time step results in a sinusodal form of the camera motion (See Figure 1 (2)). This is due to the opposite posture of the pelvis in two successive steps (left step and right step). This problem can be resolved by defining the control points on the steps of one side (all the right steps or all the left steps), which result in a more linear camera motion (see Figure 1 (3)). This solution provide a good compromise between the smoothness of the camera motion and the conservation of the same viewpoint during the animation.

- Synchronized non uniform semi-linear style: Have the camera follow the avatar motion without conserving the same viewpoint.

  Another solution for the control of the camera motion is to maintain a perfect synchronization between the camera motion and the avatar displacement without updating the viewpoint (See Figure 1 (4) and Figure 2 (1)). Comparing to the results in Figure 2 (2), the camera position in Figure 2 (1) does not provide the same viewpoint during the whole animation, but this might be interesting for some turning behavior perception based studies. The camera motion is obtained by the translation of the pelvis trajectory estimation. The latter is based on the projection of pelvis positions along the window defined with two successive steps on the straight line defining with the position of the pelvis in those two successive steps timing. This is why the camera motion looks like a succession of small straight lines according to the successive steps.

- Walking steps based style: Update the camera motion differently for straight walking steps and turning steps

  Finally another solution that aims to maintain the same viewpoint on the avatar and a good synchronization between the avatar and the camera motion is to combine the synchronized non uniform semi-linear style for straight walking steps and the update of the camera position in each frame for turning steps. This approach requires the annotation of walking steps into straight walking steps and turning steps. However, this solution needs smoothing the camera path during the transition between straight and turning steps.

4. Conclusion and future work

In this paper, we propose some solutions to control the position and the trajectory of the virtual camera used to visualize stimuli for experimental studies. Our solutions allow to automatically convert a database of body movement animation files into a database of movies for the use in a perception based study. Furthermore, we propose automatic control of the virtual camera position and motion in perceptive studies.
For future work, we aim to compare the visualization of stimuli using moving virtual camera with those created using static virtual camera through a perception based study. We aim also to compare the stimuli displayed with the different solutions that we proposed through a perception based studies for different body movements (walking, turning, sitting down...).

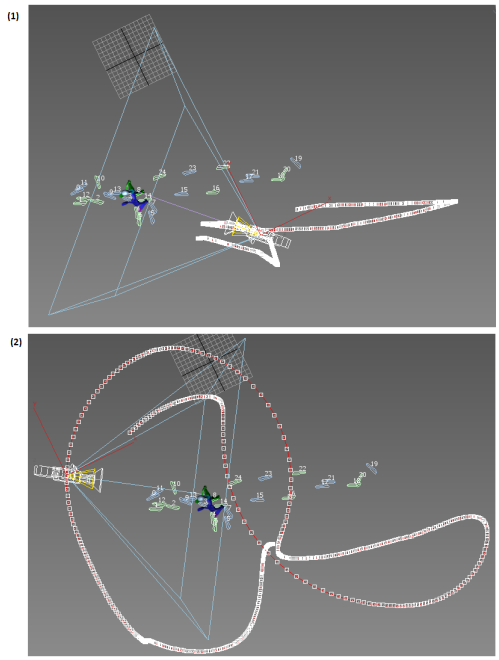

Figure 2: Camera trajectories in turning behavior; 1) without conserving the same viewpoint (Synchronized non uniform semi-linear style), 2) while conserving the same viewpoint (Synchronized non uniform non linear style). The screen shot corresponds to the same frame in the animation.

5. References

Anthony P Atkinson, Winand H Dittrich, Andrew J Gemmell, and Andrew W Young. 2004. Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception*, 33(6):717–746.

Mark Coulson. 2004. Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior*, 28(2):117–139.

Anja Daems and Karl Verfaillie. 1999. Viewpoint-dependent Priming Effects in the Perception of Human Actions and Body Postures. *Visual Cognition*, 6(6):665–693.

Sofia Dahl and Anders Friberg. 2007. Visual Perception of Expressiveness in Musicians' Body Movements. *Music Perception*, 24(5):433–454.

Mathias Dekeyser, Karl Verfaillie, and Jan Vanrie. 2002. Creating stimuli for the study of biological-motion perception. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, 34(3):375–82, August.

Nesrine Fourati and Catherine Pelachaud. 2013. Head, shoulders and hips behaviors during turning. *4th international workshop on Human Behavior Understanding, In conjunction with ACM Multimedia 2013*, 8212:223–234.

Harry J. Griffin, Min S.H. Aung, Bernardino Romera-Paredes, Ciaran McLoughlin, Gary McKeown, William Curran, and Nadia Bianchi-Berthouze. 2013. Laughter Type Recognition from Whole Body Motion. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 349–355, September.

Halim Hicheur, Hideki Kadone, Julie Grèzes, and Alain Berthoz. 2013. Perception of emotional gaits using avatar animation of real and artificially synthesized gaits. *Humaine Association Conference on Affective Computing and Intelligent Interaction*.

Gunnar Johansson. 1973. biological motion. *Perception & Psychophysics*, 14(2):201–211.

Andrea Kleinsmith, P. Ravindra De Silva, and Nadia Bianchi-Berthouze. 2006. Cross-cultural differences in recognizing affect from body posture. *Interacting with Computers*, 18(6):1371–1389, December.

Andrea Kleinsmith, Nadia Bianchi-Berthouze, and Anthony Steed. 2011. Automatic Recognition of Non-Acted Affective Postures. *IEEE Transactions on Systems Man and Cybernetics Part B Cybernetics*, 41(4):1027–1038.

Marco Meijer. 1989. The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behavior*, 13(4):247–268.

Anne-Helene Olivier, Richard Kulpa, Julien Pettre, and Cretual Armel. 2009. A Velocity-Curvature Space Approach for Walking Motions Analysis. *MIG 2009*, pages 104–115.

Claire L Roether, Lars Omlor, Andrea Christensen, and Martin A Giese. 2009. Critical features for the perception of emotion from gait. *Journal of Vision*, 9(6):1–32.