Multimodal Corpora:
How Should Multimodal Corpora Deal with the Situation?

Workshop Programme

09:15 – 09:30 – Welcome

09:30 – 10:30 – Session 1

Stylianos Asteriadis, Noor Shaker, Kostas Karpouzis and Georgios N. Yannakakis: *Towards player's affective and behavioral visual cues as drives to game adaptation*

Elena Grishina and Svetlana Savchuk: *Multimodal clusters in spoken Russian*

10:30 – 11:00 Coffee break

11:00 – 12:30 – Session 2

Mary Swift, George Ferguson, Lucian Galescu, Yi Chu, Craig Harman, Hyuckchul Jung, Ian Perera, Young Chol Song, James Allen and Henry Kautz: *A multimodal corpus for integrated language and action*

Masashi Inoue, Ryoko Hanada, Nobuhiro Furuyama, Toshio Irino, Takako Ichinomiya and Hiroyasu Massaki: *Multimodal corpus for psychotherapeutic situation*

Samer Al Moubayed, Jonas Beskow, Björn Granström, Joakim Gustafson, Nicole Mirning, Gabriel Skantze and Manfred Tscheligi: *Furhat goes to Robotville: A large-scale human-robot interaction data collection in a public space*

12:30 – 14:00 Lunch break

14:00 – 16:00 Session 3

Anders Grove: *Automatic analysis of hand movement phases in video speech*

Yasuharu Den and Tomoko Kowaki: *Annotation and preliminary analysis of eating activity in multi-party table talk*

Patrizia Paggio and Costanza Navarretta: *Classifying the feedback function of head movements and face expressions*

Jens Edlund, Mattias Heldner and Joakim Gustafson: *Who am I speaking at? Perceiving the head orientation of speakers from acoustic cues alone*

16:00 – 16:30 Coffee break

16:30 – 17:30 Session 4

Jens Allwood and Elisabeth Ahlsén: *Incremental collection of activity-based multimodal corpora and their use in activity-based studies*

Johannes Wienke, David Klotz and Sebastian Wrede: *A framework for the acquisition of multimodal human-robot interaction data sets with a whole-system perspective*

17:30 – 18:00 Discussion and closing

Editors

Jens Edlund	KTH, Sweden
Dirk Heylen	Univ. of Twente, The Netherlands
Patrizia Paggio	Univ. of Copenhagen, Denmark/ Univ. of Malta, Malta

Workshop Organizers/Organizing Committee

Jens Edlund	KTH, Sweden
Dirk Heylen	Univ. of Twente, The Netherlands
Patrizia Paggio	Univ. of Copenhagen, Denmark/ Univ. of Malta, Malta

Workshop Programme Committee

Hamid Aghajan	Stanford University, USA
Elisabeth Ahlsen	Univ. of Göteborg, Sweden
Jan Alexandersson	DFKI, Germany
Jens Alwood	Univ. of Göteborg, Sweden
Philippe Blache	Univ. de Provence, France
Susanne Burger	Carnegie Mellon Univ., USA
Kristiina Jokinen	Univ. of Helsinki, Finland
Stefan Kopp	Univ. Bielefeld, Germany
Costanza Navaretta	Univ. Of Copenhagen, Denmark
Karola Pitsch	Univ. Bielefeld, Germany
Andrei Popescu-Belis	Idiap Research Inst., Switzerland
Ronald Poppe	Univ. of Twente, The Netherlands
Albert Ali Salah	Bogazici University, Turkey
Bjoern Schuller	TU Munich, Germany
Alessandro Vinciarelli	Univ. Glasgow, UK

Table of contents

Author Index

Introduction

Currently, the creation of a multimodal corpus involves the recording, annotation and analysis of a selection of many possible communication modalities such as speech, hand gesture, facial expression, and body posture. Simultaneously, an increasing number of research areas are transgressing from focused single modality research to full-fledged multimodality research. Multimodal corpora are becoming a core research asset and they provide an opportunity for interdisciplinary exchange of ideas, concepts and data.

The 8th Workshop on Multimodal Corpora is again collocated with LREC, which has selected *Speech and Multimodal Resources* as its special topic. This points to the significance of the workshop's general scope, and the fact that the main conference special topic largely covers the broad scope of our workshop provides us with a unique opportunity to step outside the boundaries and look further into the future, and emphasize the fact that a growing segment of research takes a view of spoken language as situated action, where linguistic and non-linguistic actions are intertwined with the dynamic conditions given by the situation and the place in which the actions occur. As a result, the 2012 Workshop on Multimodal Corpora holds a number contributions which share a focus on the acquisition, description, and analysis of situated multimodal corpora.

Towards Player's Affective and Behavioral Visual Cues as drives to Game Adaptation

Stylianos Asteriadis[1], Noor Shaker[2], Kostas Karpouzis[1], Georgios N. Yannakakis[2]

[1]National Technical University of Athens, 157 80 Zographou, Athens, Greece
[2]IT University of Copenhagen, Rued Langaards Vej 7, 2300 Copenhagen, Denmark
{stiast, kkarpou}@image.ntua.gr,{nosh, yannakakis}@itu.dk

Abstract

Recent advances in emotion and affect recognition can play a crucial role in game technology. Moving from the typical game controls to controls generated from free gestures is already in the market. Higher level controls, however, can also be motivated by player's affective and cognitive behavior itself, during gameplay. In this paper, we explore player's behavior, as captured by computer vision techniques, and player's details regarding his own experience and profile. The objective of the current research is game adaptation aiming at maximizing player enjoyment. To this aim, the ability to infer player engagement and frustration, along with the degree of challenge imposed by the game is explored. The estimated levels of the induced metrics can feed an engine's artificial intelligence, allowing for game adaptation.

1. Introduction

Playing computer games is an activity enjoyed by millions of people worldwide, for thirty or even more years now. It is an industry in which huge amounts are invested, and even slight changes of the technological components of a game engine are accepted with enthusiasm from the fanatics of computer games. Since the basic controls of interaction in the 80's, a lot has changed today, with one of the latest achievements of today's technology that of gesture recognition platforms (Microsoft Kinect). The player can just interact with the game, in a completely non-intrusive way, while his body itself plays the role of the game controls. Within this view, the path to affective computing (Picard, 1997) in game-playing has opened, showing the way to using, not only one's gestural movements as input to game control, but his behavioral and affective state.

Human-Computer interaction (e.g. human-agent communication), within in this view, is beginning to take advantage of systems consisting of sensors capturing affective and physiological data (Picard, 1997; Castellano et al., 2009; Kapoor et al., 2007). Player behavior towards particular game events or during whole sessions of gameplay can become a useful source of information for the game engine's Artificial Intelligence (AI), so that it adapts itself to player's affective state. Within this frame, heart rate measurements, respiration, pressure on the mouse, posture in a chair, blood or brain oxygen levels have been shown to be valuable behavioral indicators used as inputs to the AI of a game, so that player's enjoyment is optimized.

In search for features correlated with the notion of engagement, frustration and challenge in games, a lot of works have been proposed in bibliography (Boone and Cunningham, 1998; Wallbott, 1998; van den Hoogen et al., 2008; Sanghvi et al., 2011) using expressive body and facial movements, as well as a multitude of sensorial cues (Kapoor et al., 2007; Sykes and Brown, 2003) to inform an immersive game environment about player's actual cognitive and affective state. Estimating moments of particular behavioral cues (see Figure 1) using non intrusive means can be a valuable source of information for the game experience: First, the player is not disrupted by intrusive mechanisms which might interfere with the whole experience. Furthermore, cognitive and affective features can be transferred automatically, not necessitating that the player interrupts gameplay in order to report these data nor that he has to recall his perception on each separate gameplay experience. The advances on computer vision techniques under non-pretending conditions have allowed the proposal of a few techniques incorporating notions such as body movements (Castellano et al., 2009; Sanghvi et al., 2011) head motion and eye gaze (with eye gaze still necessitating specialized hardware (Jennett et al., 2008)).

(a) (b)

Figure 1: Players' visual reactions towards certain events taking place during gameplay

In this paper, we address the issue of estimating those game events that, in conjunction with specific player characteristics and behavioral cues, could trigger specific affective states towards a gameplay session. The proposed research is in line with Csikszentmihalyi's flow theory, i.e. game features that would characterize a game as challenging are combined with player's expressed arousal (during whole game sessions or when specific events occur) and self reported skill level, in order to infer engagement. Taking the above as input to a clustering algorithm, the system attempts to define possible moments of high engagement, frustration, and challenge, extending the model of Csikszentmihalyi. The results of the proposed system are promising, in the sense that they could contribute to the design of a self-adaptive game, aiming at maximising the feeling of engagement during gameplay.

The structure of the paper is organized as follows: Sections 2. and 3. present the game platform and the data acquisition procedure, respectively, while Section 4. gives an analytical description and discussion on experiments conducted under personalized and generalized protocols. Section 5. concludes the paper.

2. Testbed Platform Game

The testbed platform game used for our study is a modified version of Markus Perssons Infinite Mario Bros (see Fig. 2), which is a public domain clone of Nintendo's classic platform game Super Mario Bros. The original Infinite Mario Bros and its source code is available on the web [1].

Figure 2: Snapshot from Super Mario Bros game.

The gameplay in Super Mario Bros consists of moving the player-controlled character, Mario, through two dimensional levels. Mario can walk, run, duck, jump, and shoot fireballs. The main goal of each level is to get to the end of the level. Auxiliary goals include collecting as many coins as possible, and clearing the level as fast as possible.
While implementing most features of Super Mario Bros, the stand out feature of Infinite Mario Bros is the automatic generation of levels. Every time a new game is started, levels are randomly generated. In our modified version, we concentrated on a few selected parameters that affect gameplay experience.

3. Dataset Acquisition

Volunteer players in Greece and Denmark were asked to play a series of game sessions. Players were between 23 and 39 years old, while conditions were typical of those of an office environment (see Figure 3). After each game, players were asked to assess the degree of engagement, frustration and challenge associated with the gameplay. The selection of these states is based on earlier game survey studies (Pedersen et al., 2010) and the intention to capture both affective and cognitive/behavioral components of gameplay experience (Yannakakis and Togelius, 2011). Furthermore, self-reporting had to be kept as limited as possible, so that experience disruption was minimized. The assessments were given in the form of ratings from 0 to 4. The analysis presented in this paper is based on 36 players playing 240 games. A more analytical description of the experimental procedure and data collection protocol can be found in (Shaker et al., 2011).
Players' recorded video sequences were analyzed using the methodology reported in (Asteriadis et al., 2009). This algorithm offers real-time estimates of head rotation. In this paper, we used the first derivative of head rotation vector norm, as an indicate of head motion "quantity".

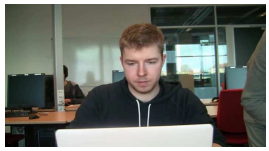

Figure 3: Typical example frame of the collected dataset.

While playing the game, different player and game-content actions, as well as their corresponding time-stamps were recorded. Player's visual behavior was estimated in the following cases: *Average head motion per game, Head motion when player loses, Head motion at stomping on an enemy to kill him, Head motion when player is about to make a critical move*.
Furthermore, profile characteristics considered here were the following: *Whether player is a frequent gamer, How much time they spend playing games on a weekly basis, Age,* and *Whether they had played Super Mario before.*
The above parameters are used as inputs for predicting user affective and cognitive state (engagement, frustration, challenge) experienced after each game session.

4. Experiments

4.1. Player independent training

For estimating user state (engagement, challenge, frustration), different combinations of the above features were tried. Each player's annotations were averaged on a per game basis, normalized from 0 to 1 and further classified to labels (challenged-not challenged, engaged-not engaged, frustrated-not frustrated). Table 1 gives an overview of F-measures and overall accuracies achieved for different combinations of features, for all game sessions, following a leave-one-player-out protocol, utilizing Fuzzy 3-NN clustering (Keller et al., 1985). Mean Head Motion is the average head movement (expressed as the first derivative of head rotation) throughout all sessions for every person, while Mean Lose-Events Head Motion, Mean Head Motion at killstomp (killing enemies by stomping them), Mean Head Motion at Move Start are the corresponding average values per person for a period of 10 frames before and after the corresponding events. Before using the algorithm all data were normalized from 0 to 1. Typical player reactions when losing can be seen in figure 4.
The above results indicate that visual motion behavior can be a strong indicate for all three affective and cognitive states. More specifically, average head motion appears to be an indicate for distinguishing between challenging and non-challenging games. Challenging interactions increase the levels of arousal (Gross, 1993) and the player externalizes this experience by high levels of overall motion. Head Expressivity when a critical move is about to take place appeared to be low when players felt challenged by the game, probably due to the fact that they were trying to concentrate on the critical move. This characteristic would be mainly associated with games provoking high levels of challenge, which usually implies that the player felt at risk of losing

[1] http://www.mojang.com/notch/mario/

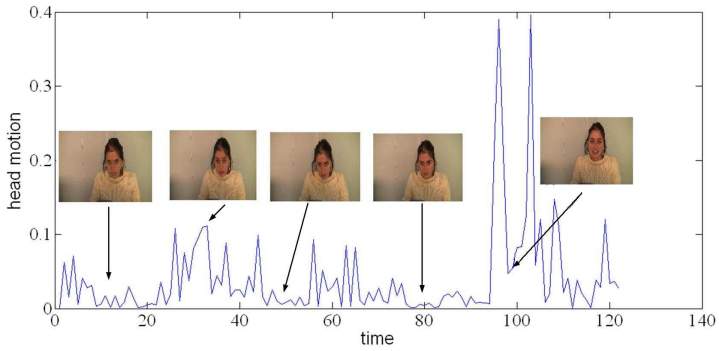

Figure 4: Player visual behavior during gameplay. In this session, Super Mario was killed in seconds $\simeq 32$ and $\simeq 100$.

Table 1: F-measures and accuracy achieved for different combinations of behavioral features and player details. 1's correspond to feature used for estimating the behavioral state of the corresponding column, and 0's mean that the corresponding feature has not been used.

	Challenge	Frustration	Engagement
Mean Head Motion per Session	1	0	0
Mean Lose-Events Head Motion	0	1	0
Mean Head Motion at Killstomp	1	1	1
Mean Head Motion at Move Start	1	1	0
Played Before	1	0	0
Time of playing per week	0	0	1
Playing Games	0	1	1
Age	0	1	1
F-measure / Accuracy	0.73/69%	0.71/74%	0.70/71%

and momentary increased levels of concentration were vital. On the contrary, frustrating games would mainly be associated with high motion expressivity at the start of critical moves. High expressivity when stomping to kill an enemy appears to be positively correlated with high levels of challenge and frustration, although engaging games showed the contrary.

Having prior experience in Super Mario also appears to play a role for the cases of frustration and engagement. Our results indicate that general gamers would, more frequently, declare that no engagement or frustration was experienced, and that may be attributed, probably, to their game habits. Similar is the case for younger players, probably due to their exposure to different kind of games (see Fig. 5). However, those players declaring that they had never played Super Mario before had more chances of saying that they felt challenged by the game, than the experienced ones.

4.2. Player dependent training

Estimating player state based on his or her *own* behavioral characteristics is of primary importance for game adaptation. Different players pose different expressions, motion patterns and expressivity characteristics when reacting to the same stimuli. This idea triggers experimentation on building on individual profiles with aim at a personalized, profile-aware game, capable of discriminating between individual behavioral and affective cues. We used a subset of

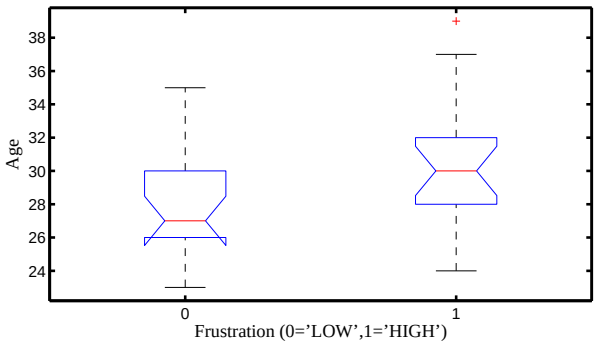

Figure 5: Frustration levels as function of age.

the players of the dataset described above, so that each of them played at least 8 games. We tested for each gameplay of each player, separately, using as knowledge-base only data from his own games, and we considered as input, behavioral cues from head expressivity. It was noticed that classifying between games provoking high and low levels of engagement gave the best results (F-measure=0.61, accuracy=82%).

5. Discussion and Conclusions

This paper has explored the possibility of using visual behavior during certain game events, as well as player's pro-

file information, as predicates of behavioral, cognitive and emotional states. Our preliminary results show that subsets of features can be utilized during gameplay, in order to elicit hidden information regarding user state and, thus, use it for game adaptation. Experimentation on a personalized level reveals that there is also potential for individualized game adaptation. However, these experiments need to be further expanded with more data, in order to be able to generalize across a much richer set of subjects. Moreover, ideally, the number of men and women in the dataset should be balanced (in this paper, out of 36 participants, only 8 of them were women). Furthermore, parameters related to game difficulty should also be taken into account in conjunction with visual and profile characteristics, as a metric for game challenge. It is also worth to point out that the moderate prediction accuracies obtained can be most likely due to the limitations of the rating reporting scheme considered in this paper. Self-reported ratings are affected by a number of effects including culture, personality and several types of scaling biases. Moreover, recent findings suggest that rating reporting schemes yield higher order and inconsistency effects when compared to ranking reporting schemes (such as pairwise preferences) (Yannakakis and Hallam, 2011). Future work will, therefore, focus on predicting ranking self-reports of the players — which are existent in the dataset but not used in this study — via the use of preference learning (Shaker et al., 2010).

Future research will focus on evaluating a closed-loop system, i.e. perform game adaptation based on the inferred player state during gameplay, in order to explore the practical usability of the above findings for minimizing frustration and maximizing player engagement. Special focus will also be placed on analyzing cultural and gender differences, as components of player's personal profile.

6. Acknowledgements

This research was supported by the FP7 ICT project SIREN (project no: 258453)

7. References

Stylianos Asteriadis, Paraskevi Tzouveli, Kostas Karpouzis, and Stefanos Kollias. 2009. Estimation of behavioral user state based on eye gaze and head pose - application in an e-learning environment. *Multimedia Tools and Applications, Springer*, 41(3):469 – 493.

R. Thomas Boone and Joseph G. Cunningham. 1998. Children's decoding of emotion in expressive body movement: The development of cue attunement. *Developmental psychology*, 34(5):1007–1016.

Ginevra Castellano, André Pereira, Iolanda Leite, Ana Paiva, and Peter W. McOwan. 2009. Detecting user engagement with a robot companion using task and social interaction-based features. In *Proceedings of the 2009 international conference on Multimodal interfaces*, ICMI-MLMI '09, pages 119–126, New York, NY, USA. ACM.

Levenson R Gross, J. 1993. Emotional suppression: Physiology, self-report, and expressive behaviour. *Journal of Personality and Social Psychology*, 64(6):970–986.

Charlene Jennett, Anna Louise Cox, Paul A. Cairns, Samira Dhoparee, Andrew Epps, Tim Tijs, and Alison Walton. 2008. Measuring and defining the experience of immersion in games. *International Journal of Human Computer Studies*, 66(9):641–661.

Ashish Kapoor, Winslow Burleson, and Rosalind W. Picard. 2007. Automatic prediction of frustration. *International Journal of Man-Machine Studies*, pages 724–736.

James M. Keller, Michael R. Gray, and James A. Givens. 1985. A fuzzy k-nearest neighbor algorithm. *IEEE Transactions On Systems Man And Cybernetics*, 15(4):580–585.

Chris Pedersen, Julian Togelius, and Georgios N. Yannakakis. 2010. Modeling player experience for content creation. *IEEE Transactions on Computational Intelligence and AI in Games*, 2(1):54–67.

Rosalind W. Picard. 1997. *Affective computing*. MIT Press, Cambridge, MA, USA.

Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W. McOwan, and Ana Paiva. 2011. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th international conference on Human-robot interaction*, HRI '11, pages 305–312, New York, NY, USA. ACM.

Noor Shaker, Julian Togelius, and Georgios N. Yannakakis. 2010. Towards Automatic Personalized Content Generation for Platform Games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*. AAAI Press, October.

Noor Shaker, Stylianos Asteriadis, Georgios N. Yannakakis, and Kostas Karpouzis. 2011. A game-based corpus for analysing the interplay between game context and player experience. In *Proceedings of the EmoGames workshop, International Conference on Affective Computing and Intelligent Interaction (ACII2011), October 9, Memphis, USA.*

Jonathan Sykes and Simon Brown. 2003. Affective gaming: measuring emotion through the gamepad. In *CHI 2003: New Horizons*, pages 732–733. ACM Press.

Wouter M. van den Hoogen, Wijnand A. IJsselsteijn, and Yvonne A.W. de Kort. 2008. Exploring behavioral expressions of player experience in digital games. In *Proceedings of the workshop for facial and bodily expressions for control and adaptation of games (ECAG 2008)*, pages 11 – 19, Amsterdam.

Harald G. Wallbott. 1998. Bodily expression of emotion. *European Journal of Social Psychology*, 28(6):879–896.

Georgios N. Yannakakis and John Hallam. 2011. Rating vs. Preference: A comparative study of self-reporting. In *Proceedings of the Int. Conf. on Affective Computing and Intelligent Interaction*, pages 437–446, Memphis, USA. Springer.

Georgios N. Yannakakis and Julian Togelius. 2011. Experience-driven Procedural Content Generation. *IEEE Transactions on Affective Computing*, pages –. (in print).

Multimodal clusters in spoken Russian

Elena Grishina, Svetlana Savchuk
Institute of Russian Language RAS
121019, Volkhonka, 18/2, Moscow, Russia
E-mail: rudi2007@yandex.ru, savsvetlana@mail.ru

Abstract

The paper introduces the notion of multimodal cluster (MMC). MMC is a multicomponent spoken unit, which includes diads "meaning + gesture", "meaning + phonetic phenomenon" (double MMC) or triad "meaning + gesture + phonetic phenomenon" (triple MMC). All components of the same MMC are synchronized in the speech, gestural and phonetic components conveying the same idea as the semantic component (naturally, with available means). To put it another way, MMC is a combination of speech phenomena of different modi (semantic, visual, sound), which are tightly connected in the spoken language, and roughly speaking mean the same, i.e. convey the same idea by their own means. The paper describes some examples of double and triple MMCs, which are specific for the Spoken Russian.

1. Introduction

1.1 Multimodal Russian Corpus (MURCO)

The Multimodal Russian Corpus (MURCO; http://ruscorpora.ru/search-murco.html) has been functioning in the framework of the Russian National Corpus (RNC) since April 2010, when the pilot version of the MURCO was released.

Since the project was described in our papers in detail (Grishina, 2009b; 2010b; 2010c), we don't intend to characterize the MURCO at great length. The paper is planned to illustrate the research resources and capabilities of this corpus.

The MURCO is the result of the further development of the Spoken Sobcorpus of the RNC. The Spoken Subcorpus includes circa 10 million tokens and contains the texts of 3 types: public spoken Russian, private spoken Russian, and movie speech (the volume of the last is circa 4 million tokens).

The Spoken Subcorpus does not include the oral speech proper; it includes only transcripts of the spoken texts (Grishina 2006). To improve it and to supplement its searching capacity, we decided to supply the transcripts with the sound and video tracks. To avoid the problem related to the copyright offence and the privacy invasion we have mainly used the cinematographic material in the MURCO. But the MURCO also includes the patterns of the public and private spoken Russian.

The MURCO is the collection of the clixts. A **clixt** is the pair of a *clip* and the corresponding *text* (i.e. the corresponding part of a textual transcript). A user has the opportunity to download not only the text component of a clixt (=marked up transcript), but also its sound/video component, so after downloading a user may employ any program to analyze it. The duration of a clip is within the interval of 5-20 sec.

As we have mentioned above, the total volume of the movie subcorpus is about 4 million tokens. This token volume corresponds to circa 300 hours of sound- and video track. Therefore, being fulfilled the MURCO presents one of the largest open multimodal sources.

1.2 What is multimodal cluster?

The multimodality of spoken discourse means the combination of different informational modi in utterance to convey information. The idea of multimodality has become quite popular in recent studies which concern the everyday natural communication (Cienki, 2005; Cienki & Müller, 2008; Mittelberg, 2007, i.a.).

The Multimodal Russian Corpus shows that in the same semantic zones of spoken utterances we quite often may find the same combination of the phenomena of different modi (phonetics and gestural). To describe the combinations of the kind we use the term **multimodal cluster, MMC** (Müller, 2008: 236–237).

So, we state that in an utterance the MMC takes place if the same semantic event (semantic proper, syntactic, pragmatic, stylistic one) is accompanied with the same phonetic or/and gestural events.

The double semantic-gestural MMCs are regular topics in gestural studies (Richter, 2010; Grishina, 2009a, 2010a). The double semantic-phonetic MMCs are studied in a less degree (see, e.g., Krivnova, 2007, where the meaning of defiance, which is specific to the glottal stop in Russian, has been analyzed).

Still, the most interesting are the triple MMCs, regularly combining the semantic, phonetic, and gestural events, which convey the same idea in aggregate. In this paper we want to describe some MMCs, which are regular enough in spoken Russian.

2. Demostrative particle *O*

In spoken Russian the demonstrative particle *O* is used quite often (Grishina, 2009a). It fixes the presence and availability of some object in a speaker's zone of attention. The object may be material or abstract one; in any case the particle implies that a speaker realizes that the object exists for the first time. So, the **meaning** of the particle may be roughly formulated like 'here it is!'.

Some examples:

(1) *O! Sergej prishol.*
Serge is here.
(2) *O! Imenno tak! Ty sovershenno prav.*
Just so! You're right, absolutely.

From the **phonetic** point of view, the particle *O* is characterized with the glottal stop at the beginning of the vowel. This glottal stop results in the pointed and very brief mode of the pronunciation of the particle.

On the other hand, the standard, the most frequent and practically obligatory **gesture**, which accompanies the particle, is the pointing with the index finger. In the 1st example the index finger would be directed to the newcomer, *Serge*; in the 2nd example the index finger would point upward, and it means that the speaker evaluates the previous cue as important one and expresses his/her full agreement with the previous speaker.

So, we can see that the meaning of the particle includes the idea of the *point* fixation of some event or object; at the same time, the pronunciation of the particle *O* (the glottal stop) and its gestural accompaniment (the index pointing, which metaphorically fixes the object with the tip of the finger) also include the idea of *point*. It gives us the possibility to conclude that in the case of the Russian particle *O* the regular coincidence of 1) meaning, 2) pronunciation mode, and 3) type of gesticulation is not accidental, but is a result of deep semantic coincidence of three modalities. Therefore, in this case we may speak about the triple MMC.

3. Summarizing *Da* 'yes'

In (Grishina, 2011) we have analyzed the semantic structure and the different types of usage of Russian *Da* 'yes'. One of the numerous meanings of *Da* is so called summarizing *Da*. A speaker uses the summarizing *Da*, when he/she intends to summarize, to resume his/her thoughts or meditation:

(3) *Da…Zhizn' slishkom korotka.*
Well Life is too short.
(4) *Da…L'udi ne umejut obsshat's'a.*
Well People aren't good at communication.

From the **phonetic** point of view, the summarizing *Da* is obligatory accompanied with such phonetic features, as

Ph^1) lengthening of the vowel phonation [dā]
Ph^2) delayed release of the consonant [d…a]
Ph^3) additional nasal sound at the beginning of the word [^{n}da] or [^{m}da]
4) combinations of 1–3 features.

All these phonetic characteristics prolong the phonation of the summarizing *Da*.

From the point of view of **gaze behavior** of a speaker, the summarizing *Da* is characterized with two main gaze patterns:

Gz^1) non-fixed, or non-referential gaze: a speaker is watching vacuum, not the definite object (Poggi, 2002: 235-236)
Gz^2) gaze into the distance: a speaker is looking far outside the limits of the zone of communication.

And, finally, from the point of view of **gesticulation**, the summarizing *Da* is accompanied with two types of gestures:

Ge^1) the gestures of thinking or concentrated meditation (*to rub one's chin, to scratch one's head, to put one's hands behind one's back, to frown, to count small things with one's fingers, to beat a tattoo, to walk to and fro,* and so on)
Ge^2) the gestures of keeping oneself away or of distancing oneself from something (*to raise one's eyebrows, to move one's chin sideways, to move one's body backward, to screw up one's eyes,* and so on).

We can see that all above-listed specific features of the summarizing *Da* are distributed between two MMC.

1) **The triple semantic-gestural-gaze MMC.** This MMC merges *Da* as a symbol of the thinking process, Ge^1 (the gestures of concentrated meditation), and Gz^1. The reason for the inclusion of Gz^1 in this MMC is as follows: the usage of non-referential gaze means that the speaker does not want to look at any object, which may distract his/her attention and prevent him/her from thinking.

2) The quadruple semantic-phonetic-gestural-gaze MMC. The summarizing *Da* here may be characterized as the result of meditation, during which the speaker considers the subject of his/her thinking *as a whole*. To consider something as a whole one ought to have a good look at it *from far away*. We may see that all three phonetic features of the summarizing *Da* (Ph^1–Ph^3) prolong the phonation time of the word and, thereby, the length of its pronunciation imitates the distance between the speaker and the subject of his/her thinking. The same idea is conveyed with the gestures of keeping away (Ge^2); as if the speaker takes a step back to see the subject of his/her meditation better. Finally, the Gz^2 (the gaze into a distance) means that the target of the gaze is disposed far from the communication zone; so, the type of the gaze also conveys the idea of remoteness.

4. Mimicking MMCs

The analysis of MURCO data shows that the speech acts, which include the mimicking citations, contain usually following components.

1) Repetition. In the mimicking speech acts a speaker usually repeats the cited component. Mainly the repetition takes place twice.

(5) *– Eto s'uda ne vojd'ot!*
– It'll hardly go in!
–"Ne vojd'ot, ne vojd'ot"… Vs'o v por'adke!
– "Hardly, hardly"... It's OK!

The lexical reduplication is quite often accompanied with the specific intonational pattern. The contour of the stressed syllable of second part of the repetition resembles the contour of the stressed syllable of first part, but the pitch level of the second part is lower than the pitch level of the first part (Fig. 1).

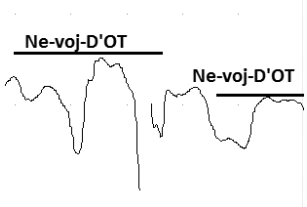

Figure 1: Lexical reduplication

But the intonational structure of the repetition may have another structure. If a speaker wants to mimic two different parts of the quoted utterance, then he/she uses the same intonational pattern, but in this case the pitch levels of two parts of the repetitions would be the same (we have called this type of the repetition intonational one, see Fig. 2).

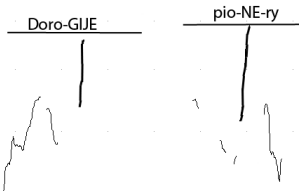

Figure 2: Intonational repetition

2) The mimicking citation is quite often accompanied with the **unnatural** (for this very speaker) **pitch** of the tone (see fig. 2). The level of the mimicking pitch is normally much higher than the pitch of the standard tone, but it is not obligatory: the level may be much lower.
3) When a person mimics somebody's utterance, he/she usually **pulls** his/her **face** and **tosses** his/her **head**.
All these mimicking features may be combined in the mimicking speech act, but also each of them may be used separately. From our point of view, they form the MMCs.

4.1 Repetition & gesture *to toss one's head*

The natural question arises, why a mimicking speech act is so closely related to repetition? To answer the question we need to formulate what mimicking is.
Mimicking is the act of disapproving citation combined with downgrading of the status of the cited utterance.
To downgrade an utterance status we may characterize it like too verbose and wordy one, i.e. we may state that the utterance contains too many words and is too short of meaning.
We suppose that the repetition of the components of a cited utterance conveys the idea of verbosity, because it symbolizes that the cited utterance includes little sense and a lot of words.
At the same time, we think that the tossing head conveys the same idea, because it imitates the forced head movements of a speaker during the act of word pronunciation.
So, the gesture in this MMC conveys the same idea as the lexical or intonational repetition, which means that here we deal with the double MMC.

4.2 Unnatural pitch & gesture *to pull one's face*

One more way to degrade the cited utterance is to state that it is not true. Besides, it may be evaluated as untimely, unsuitable or irrelevant one. In this case a mimicking person means that the cited utterance reflects real facts, but, being untrue, doesn't do justice to them. As a result, the real state of affairs seems to be reflected in a distorting mirror. The corruption of the real state of affairs in the cited utterance reflects in the mimicking utterance by means of distorted voice pitch and of disfigured and contorted face.
So, both voice pitch and distorted face in this MMC convey the idea of false meaning of the cited utterance.
At the end of the section it ought to be mentioned that mimicking is not neutral from social point of view. You may mimic somebody, but this person ought to be equal to you, or his/her position on the social scale ought to be lower than yours. If the author of a cited utterance has the privileged position on the social scale compared with yours, the only legal mode of citing is disapproving, not mimicking one. The only way to mark-up the quotation as disapproving is to use lexical or intonational repetitions in it. The other degrading means (gestures *to toss one's head*, *to pull one's face*, and distortion of voice pitch) are socially forbidden.

5. Conclusion

All MMCs listed above are based on iconicity, because the same idea in them is imitated by the means of different modalities.
1) The point in the conversation/situation, which is fixed by a speaker as the most important one, is imitated by phonetic and gestural means. The tip of an index finger, which indicates this important or newly available object, also has point character. On the other hand, the glottal stop at the beginning of the demonstrative particle *O*, also imitates the point of fixation of this object.
2) The concentrated meditation, which is specific for the meaning of the summarizing *Da* 'yes' is gesturally imitated with the non-referential gaze and with the gestures of thinking.
3) The tendency to examine the object of meditation from far away to catch and understand it as a whole, which is also specific for the summarizing *Da* 'yes', is imitated phonetically with different ways to lengthen the pronunciation of *Da* and also with the gestures, which lengthen the distance between a speaker (thinker) and the subject of his/her thinking.
4) The verbosity and dullness of a cited utterance are imitated with lexical and intonational repetitions and also with tossing/shaking head; both repetition and gesture depict the discrepancy between the quantity of words and the quality of meaning in the mimicked utterance.
5) The impropriety and incorrectness of the cited utterance are imitated with the distorted voice pitch and with the gestures to pull one's face: the unnatural tone and a grimace symbolize the faults and defects of the cited utterance.
So, the paper illustrates that in spoken speech we often meet the multimodal clusters, i.e. the concurrent and synchronized usage of the events of different conversational modi (phonetic, intonational, and gestural) to convey the same meaning (in wide sense of the word). The recent corpus methods of linguistic research and revolutionary development of multimodal corpora give the students of the spoken speech possibilities to investigate the problem purposefully.

6. Acknowledgements

The work of the MURCO group and the authors' research are supported by the program "Corpus Linguistics" of the Russian Academy of Sciences and by the RFBR (The

Russian Fund of Basic Researches) (RFFI) under the grants 10-06-00151 and 11-06-00030.

7. References

Cienki, A. (2005). Image schemas and gesture. In *From perception to meaning: Image schemas in cognitive linguistics (Vol. 29)*. Berlin, pp. 421–442.

Cienki, A.; Müller, C. (2008). Metaphor, gesture, and thought. In *The Cambridge Handbook of Metaphor and Thought*. Cambridge, pp. 483–501.

Grishina, E. (2006). Spoken Russian in Russian National Corpus (RNC). *In LREC'2006*, see bookmark.

Grishina, E. (2009a). K voprosu o sootnoshenii slova i zhesta (vokal'nyj zhest O v ustnoj rechi). In *Komp'uternaja linguistika i intellektual'nyje tekhnologii (Po materialam ezhegodnoj Mezhdunarodnoj konferencii "Dialog'2009). Vol. 8(15)*, Moscow, pp. 80–90.

Grishina, E. (2009b). Multimodal Russian Corpus (MURCO): general structure and user interface. In *NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference. Smolenice, Slovakia, 25-27 November 2009. Proceedings.* Brno, Tribun, pp. 119-131.

Grishina, E. (2010a). Vokal'nyj zhest A v ustnoj rechi. In *Komp'uternaja linguistika i intellektual'nyje tekhnologii (Po materialam ezhegodnoj Mezhdunarodnoj konferencii "Dialog'2010). Vol. 9(16)*, Moscow, pp. 102-112.

Grishina, E. (2010b). Multimodal Russian Corpus (MURCO): First Steps. In *LREC'2010*, see bookmark

Grishina, E. (2010c). Multimodal Russian Corpus (MURCO): Studying Emotions. In *LREC'2010*, see bookmark

Grishina, E. (2011). Multimodal'nyje klustery v ustnoj rechi. In *Komp'uternaja linguistika i intellektual'nyje tekhnologii (Po materialam ezhegodnoj Mezhdunarodnoj konferencii "Dialog'2011). Vol. 10(17)*. Moscow, pp. 243–257

Krivnova, O. (2007). Javlenije laringalizacii v russkoj rechi. In Russkij jazyk : istoricheskije sud'by i sovremennost' (III Mezhdunarodnyj kongress issledovatelej russkogo jazyka. Moscow, p. 348.

Mittelberg, I. (2007). Methodology for multimodality: One way of working with speech and gesture data. In *Methods in Cognitive Linguistics*. Amsterdam / Philadelphia, pp. 225–248.

Müller, C. (2008). What gestures reveal about the nature of metaphor. In A. Cienki, C. Müller (eds.). *Metaphor and gesture*. Amsterdam, pp. 219–248.

Poggi, I. (2002). The Lexicon and the Alphabet of Gesture, Gaze, and Touch Intelligent Virtual Agents. In *Third International Workshop, IVA 2001* (*Madrid, Spain, September 10–11, 2001*). *Proceedings*. Madrid, pp. 235–236

Richter, N. (2010). The gestural realization of some grammatical features in Russian. In *Gesture: evolution, brain, and linguistic structures. 4th conference of the international society for gesture studies. July 25–30, 2010*, Frankfurt/Oder, p. 245

A multimodal corpus for integrated language and action

Mary Swift[*], George Ferguson[*], Lucian Galescu[†], Craig Harman[*], Hyuckchul Jung[†], Ian Perera[*], Young Chol Song[*], James Allen[*†], Henry Kautz[*]

[*]Department of Computer Science, University of Rochester
Rochester, NY 14627
[†]Institute for Human and Machine Cognition
40 South Alcaniz Street
Pensacola, FL 32502
{swift, ferguson, charman, iperera, ysong, james, kautz@cs.rochester.edu}
{lgalescu, hjung}@ihmc.us

Abstract

We describe a corpus for research on learning everyday tasks in natural environments using the combination of natural language description and rich sensor data. We have collected audio, video, Kinect RGB-Depth video and RFID object-touch data while participants demonstrate how to make a cup of tea. The raw data are augmented with gold-standard annotations for the language representation and the actions performed. We augment activity observation with natural language instruction to assist in task learning.

1. Introduction

Much progress has been made in individual areas of artificial intelligence, including natural language understanding, visual perception, and common-sense reasoning. But to advance intelligent assistants, systems that can interact naturally with humans to help them perform real-world tasks, an integrated approach is required. To address this issue, we have created an experimental environment for learning everyday tasks in natural environments using the combination of natural language description and rich sensor data. Natural language assists in task learning by providing a useful level of abstraction from the observed actions, as well as information about task parameters and hierarchical structure (Allen et al., 2007).

We focus on structured activities of daily living that lend themselves to practical experimentation in a kitchen domain. The initial corpus we describe here consists of recorded and annotated demonstrations of making tea. Subjects verbally describe what they are doing as they make a cup of tea, as they might in an instructional video. The audio and video are annotated with gold-standard representations of the language and activities performed.

In related work, the CMU Multi-Modal Activity Database (la Torre et al., 2009) is a corpus of recorded and annotated video, audio and motion capture data of subjects cooking recipes in a kitchen. However, the subjects did not verbally describe their actions. In addition, we use the Microsoft Kinect to capture 3D point-cloud data (section 4.3.).

In the following sections we describe the data collection, equipment used and the gold-standard annotations. We conclude with a discussion of future work.

2. Corpus Overview

The corpus consists of recorded and annotated sessions of people demonstrating how to make a cup of tea. The raw data comprise audio (both speech and ambient sound), video, Microsoft Kinect RGB-Depth video, RFID object-touch data and data from other environmental sensors. The raw data are augmented with gold-standard annotations for the language representation and the actions performed. Ten sessions were selected from a larger set for the gold-standard annotation, with each of seven participants represented at least once.

3. Recording Protocol

Subjects were familiarized with the experimental kitchen setup, including the electric kettle and the location of objects they might use, such as the tea, cups and cutlery. They were instructed to make tea and at the same time verbally describe what they were doing, as if teaching someone how to do it. They were asked to speak naturally. We recorded three sessions for each of seven participants from the research community, for a total of 21 tea-making sessions. We recorded additional sessions in the event of technical failure during recording, such as no audio from the lapel mike.

Figure 1: The experimental kitchen setup.

4. Equipment Setup

The kitchen environment is shown in Figure 1. The simple kitchen has an island workspace with a working sink. There

are cabinets and a refrigerator along the back wall. The video camera and Kinect were mounted on tripods to record the activity. Ceiling-mounted microphones record the ambient noise. Subjects wore a lapel microphone to capture speech, and an RFID sensing iBracelet on each wrist.

4.1. Audio

Audio data was collected from three microphones: a lavalier microphone as an unobtrusive means of capturing speech, and two ceiling mounted microphones for recording of the environment. All equipment used was consumer-grade. The lavalier microphone was an Audio-Technica AT831cW connected to the Audio-Technica 701/L UHF Wireless System. The other two microphones were Behringer C2 condenser microphones. The interface to the computer used for data collection was via a Tascam US800 USB 2.0 audio interface. All audio was recorded at 44.1kHz in a 16-bit, linear PCM format.

4.2. Video

Video was captured using a tripod-mounted Flip Mino HD camera. Video is stored as 720p HD using H.264 (60 fps) and AAC codecs in an MPEG-4 container. The main use of the video is to support annotation and for presenting the results of recognition in context. We have not yet explored using it for computer vision (see next section on the Kinect).

4.3. Kinect

RGB-Depth cameras allow for the easy but accurate collection of synchronized depth information in addition to the RGB image. In particular, the Microsoft Kinect has been the sensor of choice for many machine vision tasks over the past year as it provides a low cost and robust alternative to video-only cameras. The Kinect is particularly suitable for indoor activity data collection due to its high depth accuracy and framerate despite its constrained field of view (of around 60 degrees). Problems of human and object segmentation and localization that are difficult for ordinary video have the potential to be solved in a more straightforward manner using RGB-Depth data.

Video was collected using the two cameras on the Kinect, RGB and depth. Using the OpenNI drivers [1] for the Kinect, the two images were aligned so that each pixel in the depth image is aligned with the RGB image. The cameras were centered and placed approximately one meter away from the kitchen table with the entire kitchen lab inside the field of view. Each collected frame was indexed, timestamped and stored in two separate image streams both 640x480 pixels in resolution, at an average of 6.5 frames per second. Figure 2 shows both the RGB and depth streams for a given frame.

4.4. Environmental Sensors

RFID tags are attached to objects in the scene that will be interacted with. The subject wears an RFID sensing iBracelet on each wrist, which records the RFID tag that is closest to the subject's wrist at any given time. Only one RFID tag is detected at any given time, and the current tag being detected is sampled every .2 seconds.

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

Figure 2: A frame from the activity recognition data showing the RGB stream (left) and depth stream (right).

The RFID detection is somewhat unreliable nearby tags can interfere with each other, and often tags go undetected for periods of time. In an attempt to overcome these limitations, we attach multiple tags to each object, improving detection rates.

Other sensors were attached to kitchen appliances and cabinets to provide additional data as the subject interacts with the environment. The electronic tea kettle was attached to a Watts up PRO ES Electricity meter. Door sensors were placed on the kitchen cabinet doors and drawers to detect when they were opened and closed.

5. Annotation

We use the ANVIL annotation tool (Kipp, 2001) to create a multi-layered representation of the session data. Our data annotations consist of the speech transcript, a logical form for each utterance, an event description extracted from the logical form, and gesture annotations for actions, objects, paths and locations. Each of these annotation layers are described in more detail below.

5.1. Speech

We transcribed the speech based on the lapel microphone recording, then we segmented the transcription into utterances. Breaking points were typically at the end of sentences. However, since the speech was spontaneous, we had many utterances that were not complete sentences (e.g., missing predicates); in such cases, we considered long pauses to mark utterance boundaries. There were some cases of sentences being uttered in a continuous sequence, with no pause between them; in such cases we considered the whole segment to be a single utterance, rather than breaking it up into sentences.

5.2. Language

5.2.1. Logical Form

The speech transcriptions were parsed with the TRIPS parser (Allen et al., 2008) to create a deep semantic representation of the language, the logical form (LF). The parser uses a semantic lexicon and ontology to create an LF that includes thematic roles, semantic types and semantic features, yielding richer representations than "sequence of words" models. Figure 3 shows a graphical representation of the logical form for the utterance *Today I'm going to make a cup of tea*. In triples of the form (:* CREATE MAKE), the second term is the semantic type in the ontology for the word, which is the third term. Nodes are con-

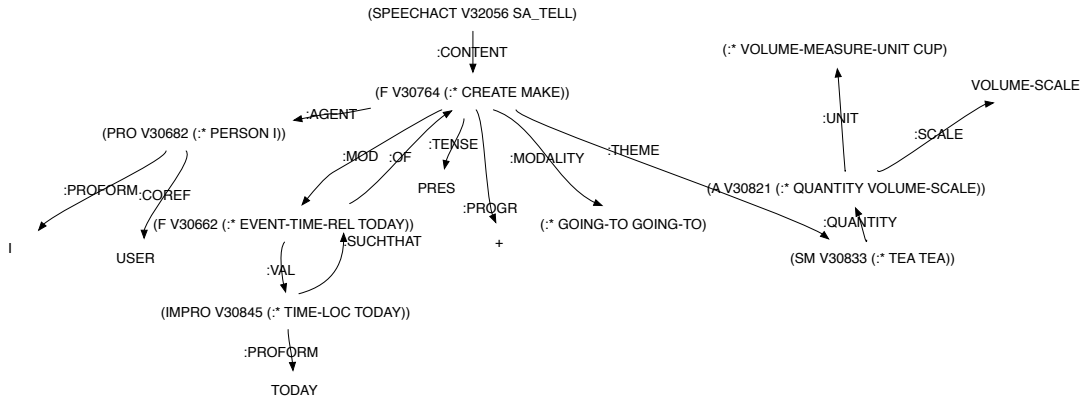

Figure 3: Logical form for *Today I'm going to make a cup of tea*.

nected by edges indicating dependencies, with labels for the relation, such as the thematic role of the argument. Tense, aspect and modality information is used as input for subsequent temporal reasoning described in (6.).

The parser-generated LFs were hand-corrected for the gold standard annotation. This was particularly necessary for utterances where multiple sentences were transcribed as a single utterance, as described in 5.1..

To facilitate using language representation as features in activity recognition models, we added new semantic types in the ontology to correspond to objects and actions in the domain, such as *tea, cup, steep*. The new labels were usually specific subtypes of already existing semantic types. For example, the word *tea* was in the ontology under the general type **TEAS-COCKTAILS-BLENDS**, so we created the specific subtype **TEA**. This extension gives us greater transparency in the surface representation, but we retain the richness of the hierarchical structure and semantic features of our language ontology.

5.2.2. Event Extraction

The LFs are input to the TRIPS Interpretation Manager (IM), which computes crucial information for reasoning in the domain, including reference resolution. The IM extracts a concise event description from each clause, derived from each main verb and its arguments. The event descriptions are formulated in terms of the more abstract semantic types in the LF, resulting in short phrases such as **CREATE TEA**, **CLOSE LID**, and **POUR WATER INTO CUP**. These phrases will be used as language features in our activity recognition models. Figure 4 shows an example of an extraction from the LF for *Place tea bag in the cup*. The objects *bag* and *cup* are identified as REFERENTIAL by the IM and it also includes the coreferential index for the first mention of the term.

5.3. Activity

For activity annotation, ground truth was created manually by observing recorded videos (Section 4.2.) and annotating actions performed by test subjects. Domain actions, their attributes (e.g., theme, associated entity, relation, etc.) and admissible attribute values were pre-defined and stored as ANVIL specification that allows annotators to easily select

```
(EXTRACTION-RESULT
 :VALUE ((EVENT V38801)
  (PUT V38801) (:THEME V38801 V38818)
  (:SHORT-DESCRIPTION V38801 (PUT (:* BAG BAG) INTO CUP))
  (:INTO V38801 V38887)
  (:TEMPORAL-RELATION V38801 >NOW) (:TENSE V38801 PRES)
  (REFERENTIAL V38818) (BAG V38818)
  (:ASSOC-WITH V38818 V38814) (:COREF V38818 V38185)
  (REFERENTIAL V38887) (CUP V38887)
  (:COREF V38887 V33594)
  (NONREFERENTIAL V38814) (TEA V38814))
  :WORDS (PUT THE TEA BAG INTO THE CUP))
```

Figure 4: Extracted from the LF for the utterance *Place tea bag in the cup*.

actions/attributes/values using GUI.

For each activity, its duration was also annotated with the start and the end time relative to the videos recording start time. The duration was made with centisecond accuracy. Each video was observed several times and some segments were observed more than ten times to produce accurate activity annotation. On average, it took about 25 minutes to annotate a video. Nevertheless, for actions that were not clearly visible in the video, the timing information can have a certain degree of error (mostly less than a second).

Simultaneous actions (e.g., opening a kettle while moving it) were also annotated with overlapping time duration. Actions (*move, put, fill, pour ...*) and their attributes, such as objects (*cup, kettle, spoon ...*), actions and paths (*to, from, into*) are annotated as separate tracks in ANVIL so that they can be accessed programmatically (e.g., using XML parser) for a compositional analysis. We annotated the composite actions (e.g., move a kettle from its kettle-base towards a cup) as a separate track for ease of viewing the overall activity annotation.

Figure 5 shows an example of the language and activity annotation for the tea-making action of pouring water into the cup. In this example we have highlighted the extraction (EX) tier for the language annotation. The concise activity description **(PUT (:* BAG BAG) INTO CUP)** represents the full extraction information, which appears the attributes window. This extraction is also shown in Figure 4.

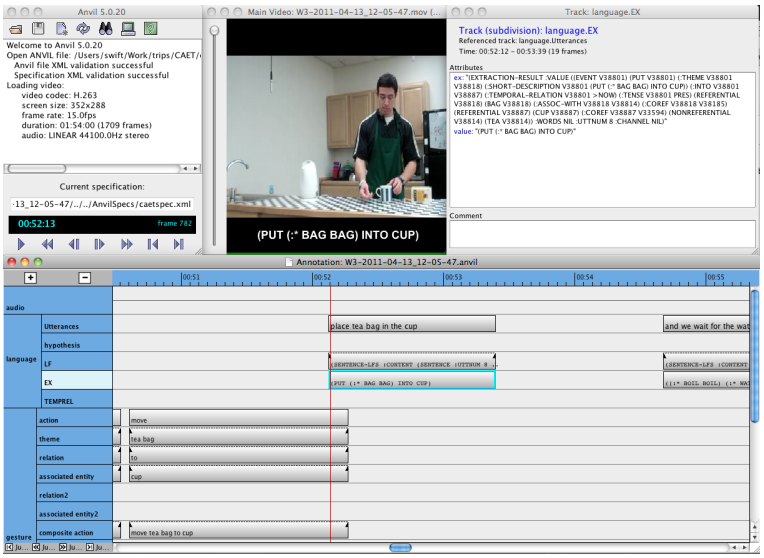

Figure 5: Language and activity annotation for *Place tea bag in the cup*.

5.4. RFID and other sensors

The sensor data annotation is in progress. While the RFID system assigns a unique ID for each tag, we must supply the name of the object (and an identifier, in the case of multiple objects of the same type) for each ID. Multiple ID's may map to the same object name. Although the dataset does not contain ground-truth annotations for the RFID, we have developed an algorithm to automatically assign names to RFID tags using the task descriptions. This statistical method, combined with an annotator's determination of object interactions from the video, provides a substitute for the ground-truth object-ID mapping.

6. Discussion and Future Work

The ten annotated sessions of our corpus comprise 184 transcribed utterances with logical form and extraction annotations, 345 annotated activities and 1.2GB (approximately 7200 frames) of Kinect data.

The use of RGB with depth data makes it possible to more easily segment and recognize human and object interactions. From the dataset we extract generally useful and "meaningful" features, such as semantic visual features (e.g. "left hand is directly above cup"). Using features at this level of abstraction, we plan on experimenting with more complex models of recognition and learning.

As mentioned above, we do not have the ground-truth data for all of the tag-object assignments, so we use an algorithm to assign the most probable object name for each RFID tag that was detected in the scene using the subject's description of the task. To learn the name of a given ID, we gather the nouns mentioned by the subject while each object is being interacted with, convert the nouns to ontological concepts using the gold-standard parse data, and determine the concept with the highest probability of being mentioned when that ID is detected. While we only have a small amount of data, the labels generated by this algorithm agreed with a human annotator, who used the video to determine the mappings, for six out of the eight tags. In future data collection, we will have access to the ground-truth data, and will not rely on our other data to discern the tag-object mappings. In the future, we hope to extend this algorithm to the Kinect data, making use of language to provide names for detected objects without the need for hand annotation.

Other future research involves computing the temporal relations between the natural language descriptions and the actions performed using Markov Logic Networks and information in the event extractions. We are in the process of adding hand-annotated gold-standard temporal relations (PAST, PRESENT, FUTURE) for each event extraction.

7. Acknowledgements

This work was supported in part by NSF grants IIS-1012205 and IIS-1012017, Activity Recognition and Learning for a Cognitive Assistant.

8. References

J. Allen, N. Chambers, G. Ferguson, L. Galescu, H. Jung, M. Swift, and W. Taysom. 2007. PLOW: A collaborative task learning agent. In *Proc. AAAI*.

J. Allen, M. Swift, and W. de Beaumont. 2008. Deep semantic analysis of text. In *Proc. Semantics in Text Processing*, STEP '08, Venice, Italy.

Michael Kipp. 2001. ANVIL - a generic annotation tool for multimodal dialogue.

F. De la Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcada, and J. Macey. 2009. Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) database. Technical Report CMU-RI-TR-08-22.

Multimodal corpus for psychotherapeutic situation

Masashi Inoue†, Ryoko Hanada, Nobuhiro Furuyama, Toshio Irino,
Takako Ichinomiya, Hiroyasu Massaki

†Yamagata University/National Institute of Informatics
3-16, 4 Jyonan, Yonezawa-shi, Yamagata, Japan
mi@yz.yamagata-u.ac.jp

Abstract

This paper presents a design principle for construction of an in-house multimodal corpus for computationally analysing and better understanding conversations during psychotherapy. We discuss some sharable information about research community data collection procedures such as recording devices, the consent form, and privacy consideration. Also, multimodal coding schema and metadata that are needed in the domain are explained. The corpus has three distinguishing properties: 1) It was constructed only for our own researches and not for public use; 2) The conversation and recording environment was in actual social situations, not controlled; 3) A multimodal coding schema that focuses on the co-construction nature of the conversation was used. Although the conversation contents are not sharable, the data collection procedure and the schema design for the psychotherapy corpus would serve as an example of an initiative to construct a multimodal corpus.

1. Introduction

To better understand the nature of psychotherapy, we are creating a micro corpus of about 20 psychotherapeutic conversation whose situation is not suitable for public sharing. In this paper, we consider the possibility to exchange information on corpus building while keeping the content of the conversation private.

Recently, there has been are growing interest in the situations where the conversation took places. The situations are either physical - such as lighting conditions, noise level, temperature and room size, or social - such as conversation domain, the relationship between speakers, and purpose of the conversation. We are interested in the social situations. In particular, we focus on domain specific characteristics of conversation. For the purpose, it is necessary to create our own corpora even though they are small compared with generic corpora. The content of our corpus will be kept private due to the sensitive nature of the counselling; however, the procedures for constructing them can be made public for validation and information sharing. In the following sections, we will describe some of the requirements for the multimodal video corpora to better understand the interactions in a specialised setting.

2. Data collection procedure

2.1. Psychotherapeutic situations

In the field of psychotherapy, in process researches rather than outcome researches, the basic form is single-case analyses. These analyses are qualitative. However, there may be hidden interactional patterns that will be discovered when we employ quantitative or mixed research approach by exploring multiple cases. For the purpose, we need to collect psychotherapy data as multimodal corpora.

The individual research project has its own target domain. Among various psychotherapeutic situations, we focus on the following two: 1) psychotherapist training situation in which participants were graduate students of psychotherapy major at Kyoto university of education (training); 2) clinical situation where participants were patients of a chronic disease and having some forms of psychological problems as well (clinical). The training situation consists of 13 psychotherapeutic interviews. In the training situation, there was a study group where therapists help each other by observing their practice sessions which last around 20 to 50 minutes. Sessions often are split into several sub-sessions by inserting of reflection periods by the observers. A restriction there was that the counselling session must be completed within a day, even if a reasonable solution is not found. In contrast, in the clinical situation, therapy may continue for several sessions. It consists of seven counsellings of various session lengths, and some of the sessions are on-going.

2.2. Privacy and motivation

The most difficult issue in the construction of psychotherapeutic counselling conversation corpora is privacy. In counselling, the participants usually talk about serious and sensitive topics. Thus, most clients do not want the content of their sessions made publicly available. In the training situation, because the participants were graduate students majoring in psychotherapy, both therapists and clients were interested in the potential for using dialogue analysis to obtain some insights on their psychotherapeutic interviews. Therefore, we are allowed to access all session data except one unavailable session where a participant refused to be video-recorded. In the clinical situation, participants were not particularly curious about the research but volunteers of good will.

2.3. Consent Form

In using therapeutic conversation for research, we have to obtain participants' agreements. We prepared a data usage consent form and asked participants to agree to allow us to disclose their counselling sessions for research purposes. From an ethical point of view, we have to maintain the privacy of the participants. From a research point of view, there should be fewer restrictions on the use of collected data to extract maximum knowledge from them. Therefore, there is a dilemma in designing a consent form. We employed the usage log approach. Participating clients are often concerned with the data accessibility, or who will see the data. Therefore, the form comes with a list of the people who participated in the research project to clarify who will have access to the data, and a list of the journals and conferences where we will present the research results on the data. However, the research group members may change after taking the consent form and it is impossible to list all potential places of presentations. Accordingly, these lists are regularly updated and can be accessed by the participants. Participants have the right to examine the list of members and presentations of all versions and to retract the use of their recorded data thereafter. There was a consent form that enhanced the utility by not specifying the places of presentation and researchers who have access to the data. For example, the following expressions were used (Clark, 2009): "The audio tape can be played at meetings of academics (e.g., conferences and presentations)" and "The written transcript can be used by other researchers". Although such option is attractive, we chose the rather restricted version that are more acceptable to many participants to increase the size of the corpus.

2.4. Non-invasive recording

Our data-recording environment was built to have minimal impact on the counselling. Luckily, the structure of the psychotherapy in the study group (training situation) already incorporated the use of a video camera for purposees of their own reviewing. Although not all of the participating members used videos in their daily counselling activities, they were accustomed to the presence of the video camera. Microphones are situated some distance away so that they did not restrict the speakers' natural movements or influence their speech. See Figure 1 to find out how the counselling was recorded. We also tested a moderate invasive recording setting as shown in Figure 2, where microphones are attached near a mouth and an accelerometer is attached on top of the head. We have encountered only one client who felt this device distracting in a clinical situation.

We did not control the topic of the counselling either. The advantage of using the conversations in the study group compared with role-playing conversations was that the clients talked about their real problems. We were able to witness conversations that were from more emotionally depressed or confused participants. Although some professionals can play a role of certain disease (simulated patients), they cannot represent clients' diversity. Sometimes, the participants hesitated to openly describe their problems, and therefore the therapists occasionally failed to grasp the clients' problems, and the counselling sometimes fails or does not produce sufficient change in the participants. These irregularities are of particular interest to us.

Figure 1: Example of minimum invasive counselling conversation.

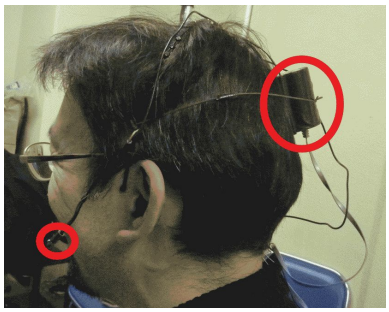

Figure 2: Example of counselling conversation with accerelometer.

3. Multimodal coding schema

3.0.1. Gesture coding

One current focus is on the gesture modality, particularly hand movements. Many researchers employ similar coding procedures for gestures based on McNeill's framework (McNeill, 1992). Basically, we also followed it but with some operational modifications. Our gesture coding schema is summarised in Table 1. An important modification lies in the distinction between gestures and non-gestures. According to McNeill, a gesture is "the movements of the hands and arms that we see when people talk". It is natural to focus on the above-defined co-speech movements if the goal of the research is psycholinguistic nature of speech and gesture production. For analysing psychotherapeutic conversations, the focus is more on the interaction toward the solution of the problem. In such cases, some silent hand movements that are categorised as non-gestures by McNeill may be included in the gesture categories as they play certain roles in the construction of the con-

Table 1: Summary of gesture coding schema

gesture	communicative	prep, stroke, hold, retract
	non-communicative	beat
		adaptor
non-gesture		(not coded)

versation. We included self-touching hand movements called adaptors in our schema. They may correspond to the mental status of the person; clearly, this is an important concept when dealing with psychotherapy. Another modification is the simplification of communicative gesture categories. Communicative gestures convey some meaning to the receivers. McNeill has come up with some sub-categories for it. Although these sub-categories are informative, due to cost restrictions, we cannot incorporate them at this moment. In addition, in our gesture-coding schema, we omit the spatial aspect of a gesture. That is, we do not look at the positions of the hands or the direction and speed of movements. There is the possibility of including these factors in our future studies. This hand gesture coding schema had been used for analysing the relationship between the frequency of gestures and semantic miscommunication (Inoue et al., 2011b).

3.1. Head gesture

The movements of heads are considered important when we want to figure out the characteristics of the conversation. They convey the feelings of speakers and listeners often in the form of head nodding. We annotated the head nodding based on the vertical head movements. Also, a series of up and down movements is considered as a single nodding. Together with the accerelometer signals, the manual nodding annotations were used in analysing the head nodding frequency and synchrony associated with the progress of therapeutic conversation (Inoue et al., 2011a).

3.2. Emotion flow

We are currently investigating the assignment of emotional scores to the video sequences. For the purpose, we developed a scoring interface named EMO. It is for the continous measurement of emotion in the conversation similar to the EMuJoy that was developed for music emotoin measurement(Nagel et al., 2007). Videos of different views are displayed with an evaluation pane. The pane is two-dimensional with axes such as pleasant versus unpleasant. An evaluator moves the mouse in the area so that the emotional state of the conversation segment they are watching can be described by the score. The six axes had been used in (Mori et al., 2010). By repeating three times, six axes were used in total.

4. Therapy-specific metadata

4.1. Therapeutic outcome

In addition to the process descriptions, the information on the entire session is of interest. The outcome of a psychotherapeutic session is a controversial issue due to its importance and ambiguity. There are several measure for the outcome(Smock, 2011). We collected the subjective evaluation for the entire session both from the therapists and clients. The criteria are whether therapists can listen to the clients well and whether therapists can solved the client's problem. We have not used them in the research so far.

4.2. Therapeutic stages

We assumed five stages in the process of therapy: introduction, elaboration, resistance, intervention, and solution. They have aliases similar to the flight of an airplane: take-off, cruising, turbulence, landing preparation, and landing. The meaning of the stages are summarised in Table 2. These codes were assigned by psychotherapy experts who have more than three years of experience after acquiring their counselling qualifications. The codes are mutually exclusive and any time slot during the recorded sessions falls into one of these five stages. The description of each label in Table 2 represents a typical event during the counselling stages. Other different but related activities could result in the same labels by the experts. Note that not all of the counselling follows the same path of the stages. For example, it may take too long to understand a client's problem and the stage remains at elaboration for the entire session. Another example may be that the therapist found a way towards a solution but the client was reluctant to accept it, and thus these intervention and resistance stages may be repeated many times. The concept of therapeutic stages is similar to the stage definitions used in the micro-counselling (Ivey and Ivey, 2002). They consist of initiating the session, gathering data, setting a mutual goal, working, and terminating the interview. They represent the process of successful interviews and are useful for teaching; students can check if their interviews follow the stage flow. However, to describe varieties of interviews, we believe that there should be a stage definition such as turbulence, since many interviews do not go straight toward the solution in reality.

4.3. Participants' background

There are two types of background information: therapists' and clients'. We keep the information as metadata for each session. Therapists' background information includes their sex and age. As for therapists' background, we are most interested in their expertise. Because the skill levels of therapists are hard to measure, we used an objective scale: the length of service after acquiring a certificate as a psychotherapist. This measure is a rough approximation of expertise but we

Table 2: List of therapeutic stage codes

Stage name	Alias	Description
introduction	take-off	Initiating session, constructing rapport. The client introduces oneself with the problem.
elaboration	cruising	Exploring the situation and searching for the solution. The therapist tries to find some clues for the solution.
resistance	turbulence	Struggling due to the miscommunication or resistance by clients. The client feels uncomfortable or resists the therapy
intervention	landing prep	Determine the candidate action toward solution. A route to a solution entered the therapist's mind.
solution	landing	Conclude interview. The client could rethink their problem in a better way.

find it reasonable enough to distinguish novice therapists from experts. Alternatively, we can use the number of cases the therapist had been working on. However, not all therapists keep a record of their works and we do not collect the information currently. Another important, yet difficult to obtain, fact is the school of psychotherapy. The techniques they use and the goal of the sessions depend on the school. However, some of them do not follow particular school or mix the knowledge of different schools into their own ways. Therefore, in our corpus, the school information serves as reference material.

Clients' background information includes their sex and age. Also, their education and family structures are important to understand the context of the therapy sessions. However, due to the sensitive nature of the information, we do not record the latter information in any files and exchange the information orally. Since the number of cases is limited and at least one of our project members is involved in any sessions, we can recall the background information when needed. When publishing or presenting the research results, we sometimes alter the clients' background information for the purpose of anonymising their identity following the convention of the psychotherapy field.

5. Conclusion

In this paper, we explained our procedure for constructing a multimodal video corpus to better understand conversations during psychotherapy. Empirical understanding of psychotherapeutic conversation is needed and the corpus we are building can be an initial step toward the corpus-based study of psychotherapy. We illustrated the importance of paying particular attention to the nature of psychotherapy regarding the following three aspects: the sensitive nature of a conversation and the privacy and motivation issue, the special data collection environment for reducing the disturbance of conversation, and the need for a particularly tailored coding scheme and metadata. These issues have been addressed in this paper.

Although our strategy can be further improved, we believe the information provided here can be useful reference information for researchers who are going to construct multimodal corpora in similar social situations for better understanding conversations.

6. Acknowledgements

This research was partially supported by Grants-in-Aid for Scientific Research 19530620 and 21500266, and a research grant from Kayamori Foundation of Informational Science Advancement.

7. References

Shannon Jay Clark. 2009. Getting personal: Talking the psychotherapy session into being. Ph.D. thesis, The Australian National University, July.

Masashi Inoue, Toshio Irino, Nobuhiro Furuyama, Ryoko Hanada, Takako Ichinomiya, and Hiroyasu Massaki. 2011a. Manual and accelerometer analysis of head nodding patterns in goal-oriented dialogues. In Human-Computer Interaction, Part II, HCII 2011, volume LNCS 6762, pages 259–267.

Masashi Inoue, Mitsunori Ogihara, Ryoko Hanada, and Nobuhiro Furuyama. 2011b. Gestural cue analysis in automated semantic miscommunication annotation. Multimedia Tools and Applications, pages 1–14, January.

Allen E. Ivey and Mary Bradford Ivey. 2002. Intentional Interviewing and Counseling. Brooks/Cole, 5th edition.

David McNeill. 1992. Hand and mind. The University of Chicago Press.

Hiroki Mori, Tomoyuki Satake, Makoto Nakamura, and Hideki Kasuya. 2010. Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics. Speech Communication.

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.

Sara A. Smock. 2011. A review of solution-focused, standardized outcome measures and other strengths-oriented outcome measures. In Solution-focused brief therapy: A handbook of evidence-based practice. Oxford University Press.

Furhat goes to Robotville:
A large-scale human computer data collection in a public space

Samer Al Moubayed, Jonas Beskow, Björn Granström, Joakim Gustafson, Nicole Mirnig*, Gabriel Skantze
KTH Speech, Music and Hearing, Stockholm, Sweden, *HCI&Usability Unit ICT&S Center, University of Salzburg, Austria

E-mail: {sameram|beskow|bjorn|jocke|gabriel}@speech.kth.se, nicole.mirnig@sbg.ac.at

Abstract

In the four days of the Robotville exhibition at the London Science Museum, UK, during which the back-projected talking head Furhat in a simple yet effective situated spoken dialogue system was seen by almost 8 000 visitors, we collected a database of 16 000 utterances spoken to Furhat in situated and unrehearsed interaction. The data collection is an example of a particular kind of corpus collection of human-machine dialogues in public spaces that has several interesting and specific characteristics, both with respect to the technical details of the collection and with respect to the resulting corpus contents. In this paper, we take the Furhat data collection as a starting point for a discussion of the motives for this type of data collection, its technical peculiarities and prerequisites, and the characteristics of the resulting corpus.

1. Introduction

In December 2011, a spoken dialogue system featuring the back-projected physical talking head Furhat (Al Moubayed et al., 2012) was on display at the Robotville exhibition at the London Science Museum. During the four days of the exhibition, Furhat was seen by almost 8000 museum visitors, including many children, which took the opportunity to chat with the system. All in all, the system collected 16000 utterances of unrehearsed, unscripted interaction. The Furhat data collection in London is an example of a type of data collection where the main effort is spent capturing *large-scale* corpora of *situated* human-machine interactions that take place in *authentic public environments*. In order to achieve this, sacrifices must be made on many levels.

This paper discusses the motivation for capturing this type of corpus, its merits, and the necessary trade-offs in data collections like Furhat at Robotville.

2. Background and related work

Although collection of large-scale situated data in public spaces is a cumbersome task, several successful attempts have been made.

The multimodal spoken dialogue system August (Gustafson, 2002) was used to collect spoken data for more than half a year in 1998 at the Culture Centre in Stockholm, Sweden. August could answer questions about for example restaurants in Stockholm or about his namesake, the author August Strindberg. More than 10000 utterances were collected from 2500 visitors.

Pixie (Gustafson, 2002) collected data from museum visitors, starting in 2002 and lasting for more than two years. Pixie was part of the futuristic exhibition "Tänk om" (with the ambiguous interpretation "What if?/Think new!"), which consisted of a full-scale future apartment, in which Pixie appeared as an assistant and an example of an embodied speech interface. Pixie was introduced to the visitors in a movie portraying a future family living in the apartment. Next, the visitors were allowed to enter that same apartment, in which they interacted with Pixie in a computer game setting, helping her perform tasks in the apartment, such as changing the lighting in the apartment. The visitors were also encouraged to ask Pixie general questions about herself or the exhibition. The resulting corpus contains about 100 000 utterances.

In 2004, the life-sized multimodal dialogue system Max was displayed for several years in the Heinz Nixdorf Museums Forum, a public computer museum in Paderborn, Germany (Kopp et al., 2005). Max took written language as input and responded with synthesized speech. In its first seven weeks at the museum, Max recorded over 50000 inputs.

Finally, Ada and Grace, two multimodal spoken dialogue system designed as twins first greeted the visitors to the Museum of Science, in Boston, US in December 2009 (Swartout et al., 2010). The twins acted as museum guides, and spook both to each other and to visitors and human guides. In early 2010, the twins collected over 6000 utterances in a little over a month.

Figure 1. The previous large-scale data collection systems: From left to right: August, Pixie, Max, Ada and Grace.

3. Motivation

Just about all development in speech technology relies heavily on data these days, and the type of data we analyse and base our models on will be reflected strongly in our results and in the behaviour of our systems. When we gather human-machine interaction data, we would ideally like it to be as realistic as possible: real users with real systems in real settings performing real tasks. And we want large quantities of data as well - the more the better. In reality, this set of requirements is unrealistic, and sacrifices have to be made one place or another. In the type of data collection discussed here, the requirements that lay firm are that the dialogues be situated - that they take place in a real, public setting with real people - and that they be sizeable, capturing large quantities of data. As is the case with Wizard-of-Oz data collections, where the system is partially or wholly replaced by a human (the "Wizard"), these data collections are in a sense a window onto the future - they reveal what will happen when we have systems that can handle what our current systems cannot (such as exceedingly noisy environments or multiple speakers with diverse goals).

4. Technical considerations

Wizard-of-Oz collections are often not feasible in these settings. In order to get large quantities of data, the systems must run full-time over extended periods of time, and having a Wizard work all hours is simply too expensive. Instead, these systems work by employing every available trick to make their interlocutors feel at home and to make them continue speaking for as long as possible. The following examples from the systems cited in the background are by no means exhaustive, but serve to illustrate that spoken dialogue designers utilize a wide range of tricks.

In August, thought balloons illustrating things the system had a chance to understand appeared above the character's head at regular intervals, in an attempt to unobtrusively suggest what visitors might say. Another, trick was to not only use push-to-talk, but to place the push-to-talk button such that speakers had to lean in close to the directed microphone to reach it, giving the impression that the button was placed somewhat poorly, rather than that the system was unable to recognize speech that was not spoken straight into the microphone. The system also made use of a video-based person detection system to simulate visual awareness. It was used to trigger the system to encourage approaching users to strike up a conversation.

In the Pixie system the visitors had to register before entering the exhibition, they were then issued RFID tags. Pixie was able to appear at different places in the futuristic home, but in order for her to show up, visitors had to insert their card at each station. This allowed the developers to track the identity, gender and age of each interlocutor, as well as keeping track of their location and of the previous dialogue. The information about the age was used to make it possible to transform children's utterances, before sending them to a black-box speech recognizer, thus improving the performance (Gustafson 2002).

In the case of Max, the most obvious trick is the use of text input rather than speech. Max also made use of face detection in order to detect users to interact with. The system also simulated its emotional state, making it possible for it to appear aware of its own performance.

The twins Ada and Grace use an entire battery of sophisticated tricks to appear more able than they otherwise would. One of the simpler is to present visitors with a list of things to ask. Another is that the dialogue with them is often human mediated - visitors will tell a guide what they want to ask, and the guide - who has had experience with addressing the twins - rephrases the questions into the microphone. Another trick is that the twins talk between themselves. As they both know exactly what the other is saying, they can often insert clever and timely remarks, which give an impression of robustness and perhaps even intelligence.

Again, these examples serve merely as an illustration of techniques to keep visitors in high spirits, which is essential for getting at the futuristic and currently otherwise unavailable data we aim at in making these data collections. Clearly, tricks are used in other circumstances as well, but to date, they are essential for the large-scale data collection in public spaces.

5. The technology behind Furhat

The robot head called Furhat, uses KTH's state-of-the-art facial animation system. Using a micro projector the animated model is projected, on a three-dimensional mask that is a 3D printout of the head used in the animation software. The back-projection technique has also allowed us to mount the head on a neck (a pan-tilt unit). The mask has been painted with back-projection paint in order to improve visibility of the projection, which makes it possible to use the Furhat head under normal light conditions. Using software-based facial animation in a robot head allows for a flexible generation of advanced facial signals that are crucial for dialogue applications. It also provides the robot with real-time lip-synchronized speech, something which has been shown to increase speech intelligibility in noisy environments (for details on why and how Furhat was built, please refer to Al Moubayed et al 2012). The lip synchronized synthesized speech also lends a sense of authenticity to the head. The laboratory version the system, which was designed to handle two human interlocutors simultaneously to make experiments with the realistic gaze provided by the back-projected talking head, used a Microsoft Kinect[1], which includes a depth camera for visual tracking of people approaching Furhat and an array microphone for capturing speech. In the public space version, these technologies are niceties that, given the current state-of-the-art, must be sacrificed for the sake of simply getting-it-to-work.

[1] http://kinectforwindows.org/

Figure 2. Pictures from Furhat at Robotville.

For speech recognition, the Microsoft Speech API was used. For speech synthesis, we used the William voice from CereProc[2]. CereProc's TTS reports the timing of the phonemes in the synthesized utterance, which was used for synchronization of the lip movements in the facial animation. It also contains a number of non-verbal gestures that were used to give Furhat a more human-like appearance, e.g. grunts and laughter.

To orchestrate the whole system, a state-chart model was used. The framework is inspired by the notion of state-charts, developed by Harel (1987) and used in the UML modelling language. The state-chart model is an extension of the notion of finite-state machines (FSM), where the current state defines which effect events in the system will have. However, whereas events in an FSM simply triggers a transition to another state, state charts may allow events to also result in actions taking place. Another notable difference is that the state chart paradigm allows states to be hierarchically structured, which means that the system may be in several states at the same time, thus defining generic event handlers on one level and more specific event handlers in the sub-state the system is currently in. Also, the transition between states can be conditioned, depending on global and local variables, as well as event parameters. This relieves state charts from the problem of state and transition explosion that traditional FSMs typically leads to, when modelling more complex dialogue systems. For the exhibition scenario, the dialogue contained two major states reflecting different initiatives: one where Furhat had the initiative and asked questions to the visitors ("when do you think robots will beat humans in football?") and one where the visitors asked questions to Furhat ("where do you come from?"). In the former case, Furhat continued the dialogue ("why do you think so?"), even though he often understood very little of the actual answers, occasionally extracting important keywords. To exploit the possibilities of facial gestures that the back-projection technique allows, certain sensory events were mapped to gesture actions in the state chart. For example, when the speech recognizer detected a start of speech, the eyebrows were raised to signal that Furhat was paying attention.

[2] http://www.cereproc.com/

6. Robotville tricks

In the crowded and noisy environment of the museum, with often tens of simultaneous onlookers, a Kinect will not work. In order to cope with this, we used handheld close-range microphones with short leads, forcing visitors to walk up to one of the microphones whenever they wanted to speak to Furhat. Close to each microphone we mounted ultrasound proximity sensors, so the system would know at all times whether someone was holding a microphone. In this way, the methods described below, that require the system to know where its interlocutors are, could be used even though the sensor technology with which they were developed could not.

The most striking feature of Furhat - his very clear gaze - was utilized to the greatest extent possible in order to raise the visitors' opinion of it. The setup with one spoken dialogue system addressing two humans was exploited in several ways:

- When nobody was present at a microphone, Furhat would look down, only to look up at each new interlocutor with a greeting as they arrived.
- Newcomers who barged in on one microphone while Furhat was already speaking with someone on the other would face a brief glance and a quick request to wait for their turn.
- When two interlocutors were involved in the same conversation with Furhat, Furhat would deflect some of the utterances he did not understand to the other interlocutor: "What do *you* think about that?"
- Furhat could pose open question to both visitors by directing the head straight in the middle then alternately seeking mutual gaze with the two visitors. By comparing the microphone levels, Furhat could then choose who to attend to and follow-up on.

Other tricks included maintaining a fairly strict control over the dialogue. The main goal of the data collection was to learn more about what happens when a system attempt to gather data - more specifically, directions - from people in public places. The dialogue type - to collect information - was kept, but the information asked for was changed to better fit the museum setting. When the system did not understand a response, it would not ask the visitor for a repetition or otherwise admit that it did not understand. Instead it would either ask a

follow-up question, or simply respond with "yeah" with positive or negative prosody, followed by “can you elaborate on that”. It could also ask the other visitor to comment on that response or ask a new question.

In order to prepare the system for initiatives from the visitors, open questions from users of August, Pixie and the twins Ada and Grace were introduced in the language model and responses to them implemented.

Both the system's ability to tell jokes and to sometimes answer with a hint of sarcasm was noted by visitors, who seemed to take it as a sign of "intelligence". Another trick that made children significantly more engaged was the possibility to tell Furhat to change his appearance (colours of his face, lips and eyebrows).

As a final trick, the developers on-site would sometimes take one of the microphones and take part in the dialogue. By doing this, they suggested to spectators what one might successfully say to the system, while they at the same time got the three-party dialogue going. In most cases, the resulting dialogue would be more successful also for the visitor speaking to the system at that time. This data, with an impromptu three-party dialogue between the system, a developer, and a visitor is interesting in several ways. It shows how naïve users can unobtrusively be guided through a dialogue, and it also allows us to model trained users of the same system under the same circumstances.

7. Robotville results

In four days, the Furhat exhibition collected around 16000 utterances - more than eight hours worth of speech and video - from people that spoke to Furhat in the presence of tens of other visitors - about 8000 all in all. The data is currently being analyzed.

The wide press coverage Furhat received often describes the system as "witty", "sarcastic" and "intelligent", statements that bear evidence of the effectiveness of the tricks exploited in the system, since the extremely noisy environment and the sheer amount of visitors resulted in the system only rarely understanding what it was being said.

86 of the visitors that interacted with Furhat also filled in a questionnaire in which they ranked the system on a number of parameters on a 5-point scale. The mean age of these visitors was 35 years, ranging from 12 to 80 years. 46 of the respondents were male, 39 female (one participant did not fill in the demographic data section of the questionnaire). All questions got mean ratings above 2.5, and questions such as "How much do you like Furhat?" and "Did you enjoy talking to Furhat?" received scores in excess of 4.

8. Conclusions

We have described an audio-visual data collection with a spoken dialogue system embodied by the animated talking head Furhat. The data collection contains situated data in a real-world public place - a museum with thousands of visitors passing by over four days. It is an example of a risky and expensive type of data collection, where great attention is paid to keeping the situation and the environment authentic and the quantities of data large, at the expense of control and system performance. A common factor for these data collections is that they collect data of human-machine dialogues that are actually more complex than what state-of-the-art technology can actually accomplish today. There are several reasons to do this - the need for data to further research and development, and the showcasing of future possibilities. Another common way of achieving this is by using Wizard-of-Oz systems, but in these massive public space collections, such systems are not feasible, both for reasons of scale and ethics.

We have described how, in this type of data collection, it is essential to exploit every trick available in order to make the conversations appear better than they actually are, if judged by the systems ability to understand and respond to what its interlocutors say. Although data collected in such setup are rich in natural interactional behaviours from naïve users, it is to some degree limited in how people interact in today’s state-of-the-art task-oriented dialogue systems. Instead, the motivation for collecting this type of data is that it is essential for us to gain insights into how people may behave when interacting with and perceive future dialogue systems and technologies. By doing this, our efforts can be guided in the right direction.

9. Acknowledgements

This work is partly supported by the European Commission project IURO (Interactive Urban Robot), grant agreement no. 248314. Also thanks to David Traum and Ron Artstein for providing transcriptions from the Ada and Grace interactions and to Jens Edlund for contributing in writing the paper.

10. References

Al Moubayed, S., Beskow, J., Skantze, G. and Granström, B. (2012). Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction, In Esposito, A. et al (Eds) Cognitive behavioural systems *Lecture Notes in Computer Science* Springer.

Gustafson, J. (2002). *Developing Multimodal Spoken Dialogue Systems. Empirical Studise of Spoken Human-Computer Interaction*. Doctoral dissertation, KTH.

Harel, D. (1987). Statecharts: A visual formalism for complex systems. Science of computer programming, 8(3), 231-274.

Kopp, S., Gesellensetter, L., Krämer, N., & Wachsmuth, I. (2005). A conversational agent as museum guide - design and evaluation of a real-world application. In *Proceedings of IVA 2005,* Berlin: Springer-Verlag.

Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., Williams, J., Leuski, A., Narayanan, S., Piepol, D., Lane, C., Morie, J., Aggarwal, P., Liewer, M., Chiang, J-Y., Gerten, J., Chu, S., & White, K. (2010). Ada and Grace: Toward Realistic and Engaging Virtual Museum Guides. In *10th International Conference on Intelligent Virtual Agents (IVA)*.

Automatic Analysis of Hand Movement Phases in Video Speech

Anders Grove
Centre for Language Technology, University of Copenhagen
Njalsgade 140, bygning 25, 2300 København S
E-mail: and.grove51@gmail.com

Abstract

I find that many video recordings of speeches show rather uncomplicated hand movements and shapes, and regarded as a corpus they would fit for analysis by primitive automation. I have implemented an automatic annotation of the hand movement phases and applied it to a political speech on a video clip. Skin colour is used to track the hands, and the boundaries of the phases are determined by changes in speed. For comparison a manual annotation has been made and a set of guidelines stated to ensure the quality. They are close to the prevailing concept of phase annotation as e.g. stated in the NOVACO scheme (Kipp, 2004), but they also use the hand shape to identify the more expressive of the phases. While the automatic annotation is simple, the comparison shows that it is plausible and could be used with caution; the kappa index is a bit above 0.5. A substantial part of the problems origins from the difficulties to distinguish between the hands when they overlap on the screen. If parameters reflecting the form of the hand could be applied it would likely remedy this, and they could also be used in an implementation of the part of the guidelines distinguishing expressive phases based upon the hand forms.

1. Introduction

1.1 Corpus of Video Clips with Speeches

The amount of video material easy to access is growing ever faster. Digital video recording is spreading through cheaper camcorders, web-cams and as a part of mobile phones. And it is easy to upload the recorded video clips to the internet. The digitalization of multimedia has made the Internet a source of an overwhelming multimodal material.
Clips of *speeches*, uninterrupted monologues, are one kind of many in the huge amount of multimodal material.

1.2 Recordings of Speeches

Browsing through examples of video clips of this kind I find that the way they are recorded makes them convenient for image processing: often they will have a view of the speaking person standing or sitting as the main object; the camera as the surroundings are rather stationary, and there are few cuts.

1.3 Hand Movements of Speeches

The hand movements accompanying recorded speeches are typically simple. The trajectories are going directly or slightly curved between two positions, repetitions are usual, the hand shape is simple, and there are frequently periods without movement. They are almost never referencing a subject through enacting, modelling or drawing it in the air.

1.4 Benefits from Video Clip of Speech

When a test person understands the purpose of the investigation, the bias is hard to consider. Considered as a corpus, in such clips the speaker is "naive" in this respect. But in any speech, especially when it is well prepared for a communicative purpose, the speaker may be highly aware of his way of appearance, or even instructed by others. In such cases he may also have lost naivety, and the value of an unwitting and spontaneous speech will be missing. Still, such a corpus does provide a considerable amount of material, and because of the simplicity of the scene and the take of the recording as well as the hand movements made by the speaker, it appears plain to deal with automatically.

1.5 Analysis of Hand Movements

Automatic hand tracking for analysis of hand movements of acceptable quality often involves a laboratory, but some have evolved methods for hand movement analysis on video. This has been done to recognize hand shapes in British Sign Language (Liwicki & Everingham, 2009) from video recording, tracking the hands by skin colour, and also using features to reveal their form, but on material made dedicated for this. Here I will make an attempt to use the skin colour on an example with a speech in the multimodal material from the web to analyze the *phases* of the hand movements,.

2. Material

The example for the investigation here is a video clip with a political speech of President Obama made June 15, 2010, on the BP oil spill disaster in the Gulf of Mexico (Obama, 2011).
It shares the traits listed above for speeches in general, but the use of hand movements is performed much more massively and consequently throughout the speech than usually seen. His hand movements will only exceptionally disappear outside the frame.
The analysis does not cover 15 s. in the beginning and again in the end because zooming of the camera obstructs the image processing. This leaves 16 m. and 15. s.

3. Method

To track the hands, I will detect identify each of them on the frames of the video clip by the colour of their skin, to get their area and the position of their centre.
When their course with the speed and direction of their movements is obtained, the phases that they go through shall be determined automatically. A prevailing conception for phase analysis is described by McNeill (1992) and Kendon (2004), and more formal and detailed by Kita, Gijn & Hulst, (1998), and adopted in the NOVACO scheme (Kipp, 2004), but I will make a very simplified version to implement here.

To assess the automatic phase detection, a manual phase coding will be made for comparison. This will be done more strictly according to the NOVACO scheme. Further I have stated and followed some rules to guide the manual annotation to improve the consistency and quality of the manual annotation, and to give an idea of what the automatic annotation is held up against.
The comparison is then done to discuss the automatic annotation, and conclude on the use of it on this example of digital speech clips.

4. Detection of Hand Movements

4.1 Representing All Colours of the Image by Few Shades

The hands are recognized by skin colour. It varies, and all these shades must be recognized. The program processing the images here has been set to generate a list of specific shades to represent all the colours in the image while minimizing the deviation, i.e.: the sum of differences between the actual colours on the image and the shades representing them. It is checked manually which on the list are used for skin colour. The image processing program maps the colour of each pixel on every frame on the list, and those of them with skin colour are identified.

4.2 Other Objects of Skin Colour

But there will also be other things than the two hands which happen to have skin colour, but are not skin. To sort out these, certain limits above, to the left and the right are defined on the image, bordering the area in which the centre of hands may be recognized, and a minimum size limit is set to get rid of the rest, too small to be hands.

4.3 Overlapping Hands

Sometimes the hands come across each other and overlap. The image processing cannot discern each hand by skin colour alone, and erroneously one of the hands is not recognized, while the other is assigned the centre position and the area of their common region. So if the image processing reports a change in number of hands, then the analysis shall find out whether the reason is that they overlap by looking at the change of area of each hand.
The analysis here always identifies overlapping as resting state. This is sometimes wrong: there is an exception when the hands are vertical and joined above the desk to make a horizontal move outwards like wiping.

5. Hand Movement Phases

5.1 Phases of Expressive Hands

When the position of the hands have been identified on each frame in a sequence, their trajectory, speed and orientation can be determined.
In this speech all the hand movements are what is commonly designated as *gestures* (Kipp, 2004). *Gestures are closely linked with the accompanying speech through timing, meaning and communicative function* (Kipp, 2004).
When the hands do not move, but are lying supported somewhere, they are in the *resting state*. When they have left the resting place and are not supported anymore, whether moving or not, they are in *activity*, and until the next resting state this period is a *gesture unit* (Kendon 2004), and can be analyzed into phases.
By pointing, drawing, enacting, or by its shape, a hand makes the signs that relate to the meaning of the speech. This is done in the significant part of the gesture phrase, the *stroke phase*, when the hand moves, and may include a following *hold phase*, if it has stopped moving, and that is what Kendon calls a "*nucleus*" (2004). Sometimes right after the stroke the hand immediately returns a little bit in the opposite direction. This is called a *recoil phase* (Kipp, 2004), and I also take this as a part of the nucleus. Just before the nucleus a hold may also occur. All this is the *expressive phase*, … *the semiotically active phase* (Kita, Gijn & Hulst, 1998). Alternatively, in some cases the expressive phase may just consist of one single hold. The expressive phase is the essential part of a gesturing hand movement and always present in it.
Often the hand must go from the resting place or the end position of the previous expressive phase to reach the position to begin the next expressive phase: this is a *preparation phase*. And if it takes some way to reach the resting position when the last expression phase before the end of the activity of the gesture unit is finished, it is the *retraction phase*. If the retraction is interrupted, it is a *partial retraction phase*.
The expressive phase together with the optional preparation phase and partial retraction phase is the *gesture phrase*; this is a common understanding of the term *gesture* (Kipp, 2004). A gesture unit may comprehend a series of gesture phrases.
Beats are beating strikes with a hand. McNeill (1992) notes that beats may be understood as an overlay on the hand movements. In this speech there are many beats, often repeated. Typically you will see the same hand shape continuously used in a sequence of beats, and in line with McNeill's conception I will here regard the beats in this speech as hold phases keeping a certain hand shape and overlaid with beats. In this analysis they will be subsumed under the stroke phase category.
Formally the conception of the phases presented above may be stated by a syntax with these definitions:

HandTrack	::=	(rest activity)* (rest)
activity	::=	phraseSeries (retraction)
phraseSeries	::=	phrase (partialRetraction) phraseSeries
phrase	::=	(preparation) expressivePhase
expressivePhase	::=	(hold) nucleus | hold
nucleus	::=	stroke (recoil) (hold)

5.2 Phase Boundaries

The NOVACO scheme observes two criteria adopted from Kita, Gijn & Hulst (1998) to settle the boundaries of a phase: change of direction and shift in velocity at the change. If only the first criterion applies it is not a phase boundary, but a segment boundary within a single phase. This distinction shall keep a stroke within the same phase while it may change the direction.

6. Automatic Determination of Hand Movement Phases

6.1 Handedness

The automatic coding of the phases here is made for each hand, and not both together. Usually this does not make much difference, since if both are used, they will usually be in the same phase

6.2 Phase Boundaries

The boundaries of hold and rest phases are placed at the moment where the speed goes below or above a certain limit. It is resting state if the hand is at the *resting place*: on the desk in the middle at the bottom of the picture; otherwise it is a hold phase.

When the hand is moving, the borders between the phases are located where the *local minima* in speed are found, i.e.: where the speed is lower than the moment before and the moment after. Phases that only last a few frames, are removed.

When the image processing stops bringing data of the hand, it may be because the hands overlap. If the area of the one hand is close to the change of area of the other at this moment, I conclude that this is the reason. It always happens at the resting place, and a boundary of a resting state is placed at the moment where the number of hands reported shift.

6.3 Phase Type Assignment

The automatic phase type assignment is much simpler than a scheme such as NOVACO:

HandTrack	::=	(rest activity)* (rest)
activity	::=	preparation expressivePhase* retraction
expressivePhase	::=	stroke | hold

7. Manual Determination of Hand Movement Phases

At the comparison to assess the automatic coding, the "gold standard" is supposed to be a manual annotation. It has only been made by myself. For this reason I have stated some guidelines for it to help to ensure consistency and make the principles explicit for discussion; they cannot guarantee against (my) human failures trying to follow them, of course.

As with the automatic annotation, a local minimum in velocity is chosen as the criterion for boundaries of the phases. Rests and holds are delimited by significant change in speed around periods where it is slow (Kita, Gijn & Hulst, 1998). If there is no movement at all, the boundary is set as soon as the slightest movement can be seen. In case a sequence of hand movements are repeated, I consider them as a single stroke phase.

The definitions of the phase types based upon the NOVACO scheme as I have stated them above is followed in the manual annotation here. But many of the phases are optional, so more criteria must be added to determine the phase.

When the hands are not moving it is the resting state if the hand is supported at the resting place. Otherwise it is a hold phase.

Basically, if it is an expressive phase, it is generally easy to decide when it is a stroke, hold or recoil; if not, when it is a preparation or (partial) retraction phase. Therefore, a crucial point in the phase analysis is to be able to distinguish the expressive phases as from others, i.e.: the preparation and (partial) retraction phases.

I determine that a phase is expressive when the hand shape is presented evident and salient. Such a presentation can be done in several ways:

- It may be that the same hand shape is maintained throughout the phase.
- Or it may be that it is accentuated through the way the movement is made,
  - the stroke is rapid and suddenly stopped,
  - even more powerful if the velocity of the stroke accelerates before the stop, or
  - by having a recoil or a hold phase following, while keeping its shape through these following phases.

The hand shape may at one moment in a sense transform either directly or indirectly to another shape at a later moment,. Basically the direct way is supposed to mean that every joint in the hand going from a specific size of angle at the initial position to another at the final position all the way through will have a size of angle that lies between the previous and the following sizes.

When this is not the case, there is a form where at least one of the joints in the hand is not on a direct transformation between the initial and the final size of angle, so I take this form to have a purpose beyond being part of the transformation: it is a sign of its own, and this indicates that the phase in which it is found is an expressive phase.

8. Agreement between the Annotations.

8.1 Confusion Matrices

Left Hand	**Hold**	**Prepar**	**Rest**	**Retract**	**Stroke**
PartRetrac	9	0	0	0	60
Hold	2731	49	248	10	459
Prepar	666	356	1186	306	1680
Recoil	11	73	14	11	233
Rest	116	19	7282	16	83
Retract	47	51	1126	856	406
Stroke	765	786	1433	223	7826
Total	4357	1334	11289	1422	10749

Table 1

Right Hand	**Hold**	**Prepar**	**Rest**	**Retract**	**Stroke**
PartRetrac	24	3	6	8	69
Hold	2184	47	206	9	437
Prepar	394	336	885	210	1109
Recoil	10	45	8	3	251
Rest	1293	149	10634	93	306
Retract	166	98	749	801	892
Stroke	732	656	987	181	5166
Total	4803	1334	13480	1305	8230

Table 2

The automatic and the manual annotations are imported into a video annotation tool, *Anvil*, to make a confusion matrix (Kipp, 2004). The time is cut into slices, and for each of these the phase type of the one annotation is compared with that of the other.
The confusion matrices can be seen in Table 1 and Table 2 for left and right hand, respectively. The phases of the automatic annotation are found in the columns and those of the manual in the rows. The highest numbers of mismatch of the slices are shown bordered by a square.

8.3 Intercoder Agreement

A comprehensive category will by chance give a high percentage of agreement because of its size. In the kappa index (Cohen, 1960) the agreement that would be provided by chance alone is subtracted from the actual agreement. The kappa index is calculated from the confusion matrix. A value of 1 shows absolute agreement, 0 that the agreement only has an extent as by chance.
The kappa index for the segmentation appears to be very high, 99.95 for the left hand and 99.98 for the right. The segmentation kappa only indicates how many of the slices are annotated by both coders at all.
To check the extent to which the slices are assigned the same phase type values another kappa index is calculated, 0.5325 for the left hand, and 0.5220 for the right. Generally the range 0.4 .. 0.6 is taken to indicate plausibility of the annotation, but it should only be used with precaution when the index is below 0.7.

8.2 Main Problems

The big parts of the errors origin from what has been automatically annotated as resting state. One reason is that the manual annotation delimits the resting state as soon as the slightest movement can be seen, while the automatic annotation also includes movements as long as the hands still overlap. Another reason is that the overlap is not always rest, but sometimes a preparation or stroke.
Many right hand resting states are erroneously annotated automatically as holds. The reason is that when the automatic annotation of rest is identified by an overlap, if the overlapping hands then begin to move a little in this period, the image processing reports that one of the hands is active, generally his right hand, as the rest position is a little left of the middle of the image. When they stop moving again, it is then recognized as a hold until the overlap is finished when the hands begin to split again.
The automatic annotation always sets respectively the preparation and retraction phase as the one just following or before the resting state, but this may be wrong. This is a confusion of expressive states from others. The confusion of stroke with preparation phase totally amounts to 7 % of all the time, and with retractions 3 % of the time.

9. Conclusion

The usual simplicity of hand movements in speeches owes much to the fact that they rarely go into complex patterns of shapes and trajectories of moving because they rarely refer to an object through similarity, and thereby sophisticated modelling, drawing or enacting. On the other hand there are lots of repetitive beats.
This paves the way for very simple requirements to the automatic annotation of the phase types. An implementation was tried on a political speech on a very suitable video clip, and a manual annotation was made to be used as a "gold standard" in comparison with the automatic one. A set of guidelines were stated for the manual annotation to make it more consistent, and to describe it.
Both annotations were based upon the generally accepted conception of hand movement phases, e.g. such as it is stated in the NOVACO scheme. For the automatic annotation the phase type assignment was done following a simplified adaptation of this.
The guidelines for the manual annotation was equipped with criteria to discern the phases of the movement that were the most expressive. These criteria were based on the hand shapes and changes of them, and their position in the concurrent movement.
Use of skin colour to recognize the hands has a basic problem to track each hand when they are crossing each other and overlap on the screen. The solution here takes advantage of the fact that most of the time both hands are in a resting state when the hands overlap, but still this is a major source of errors; the boundaries of the rest state are not precise, and sometimes the hands overlap when not at rest, and sometimes they are not found to overlap when they do.
Basically these problems are rooted in the fact that skin colour with the contour is the only criterion used here to track them. Features relating to their form could help the tracking through overlap, but also better recognition of the phases.
However, the automatic and manual annotation applied to practically all the speech and still gave a kappa value a bit above 0.5 for either hand, which generally indicates a plausible result in spite of this simple automatic analysis.

10. References

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

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

Kipp, M. (2004). *Gesture Generation by Imitation: From Human Behavior to Computer Character Animation*. Boca Raton, Florida: Dissertation.com.

Kita, S., I. van Gijn & H. van der Hulst. (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*. Lecture Notes in Artificial Intelligence, vol. 1317, Berlin: Springer, pp. 23--35.

Liwicki, S. & Everingham, M. (2009). Automatic recognition of fingerspelled words in British sign language. In *Proceedings of the 2nd IEEE Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB'09), in conjunction with CVPR2009*. pp. 50—57.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about thought.* Chicago: U. Chicago Press.

Obama, B. H. (2011). *Remarks by the President to the Nation on the BP Oil Spill.* Localised Aug. 8 2011 on http://www.whitehouse.gov/the-press-office/remarks-president-nation-bp-oil-spill

Annotation and preliminary analysis of eating activity in multi-party table talk

Yasuharu Den, Tomoko Kowaki

Faculty of Letters, Chiba University
1-33 Yayoicho, Inage-ku, Chiba 263-8522, Japan
den@cogsci.L.chiba-u.ac.jp

Abstract

In this paper, we develop a scheme for annotating eating activity in multi-party table talk, and conduct an initial investigation of how participants coordinate eating and speaking in table talk. In the proposed scheme, eating actions are classified into four categories, i) picking up chopsticks, ii) holding chopsticks, iii) catching food with chopsticks, and iv) taking in food. Each action, then, is sub-divided into a sequence of phases, i.e., preparation, stroke, retraction, and hold, in a similar way as Kendon's gesture annotation scheme. We annotated three 10-minute excerpts from a three-party table-talk corpus in Japanese, and examined the relationship between the time devoted to each type of eating action and participant's engagement in speech activity. Preliminary results showed i) that active speakers tended to spend more time for the "taking-in-food" action even when they were speaking, and ii) that the hold phase occupied the majority of the time in these "taking-in-food" actions while speaking. These results suggest that, instead of compensating for the lack of time for eating when they were not speaking, active speakers locally coordinated their eating actions with the speech by halting the movement, and retaining the location, of the hand before putting food in the mouth.

1. Introduction

In our daily life, we often converse with other persons while being engaged in an activity with a personal aim or a common aim that is shared with the others. For instance, one may talk with a friend sitting in a passenger seat while driving a car. Or, one may discuss with his/her colleagues at a business meeting while using a laptop. Understanding and modeling of such conversations in daily situations are challenging research that would contribute significantly to both scientific and engineering approaches to human communication and human-computer interaction. Yet, most of the previous resources for human conversations have concentrated on conversations with an artificial task, i.e., task-oriented dialogs, or conversations with no external activity, or they have ignored aspects of non-conversational activity (Anderson et al., 1991; Godfrey et al., 1992; Du Bois and Englebretson, 2005).

Table talk is one of the most typical examples of conversation embedded in a daily activity, which we all experience routinely. In table talk, we coordinate eating with speaking. What is peculiar to this situation is that eating and speaking are performed by the same device, i.e., the mouth. We usually speak while not eating and eat while not speaking, although some people talk with their mouths full. (In some cultures, this is thought to be bad-mannered.) Therefore, participants in table talk should solve a dual problem; they should coordinate their speech with each other and, at the same time, coordinate their eating actions with the speech. As a consequence, eating actions may influence interactional aspect of speaking such as turn-taking.

Very few studies have been conducted in this direction. Even in disciplines in humanities and social sciences, such as Conversation Analysis, in which recordings of naturally occurring spoken interaction are at the heart of the research methodology, the interplay between speech activity and non-conversational activity has not been a primary target of the study. One noticeable exception is Goodwin (1984)'s work, in which he analyzed a four-party table talk and demonstrated how listener's accessing to foods is coordinated with speaker's speech. Although pioneering, his study was restricted and only qualitative.

More recently, Mukawa et al. (2011) investigated structures of table-talk communication by analyzing three-party conversations in a simulated table-talk setting, focusing on the interplay among speech, gaze, and eating actions. They found that participants often start a new turn even when their mouths contain foods, suggesting that in table talk participants give high priority to smooth turn-taking over eating.

In this study, based on an analysis of the same corpus, we extend Mukawa et al. (2011)'s work in two ways. First, we adopt the gesture annotation scheme developed by Kendon (2004) and McNeill (2005) to precisely describe the temporal structure of eating actions, which Mukawa et al. (2011) did not consider; individual eating actions, such as "holding chopsticks" and "catching food," are sub-divided into several phases, i.e., preparation, stroke, retraction, and hold. Second, we focus on the relationship between the time devoted to each type of eating action and participant's engagement in speech activity, showing that active speakers locally coordinate their eating actions with the speech by halting an ongoing eating action. These steps would indubitably advance the methodology for analyzing non-conversational actions, and contribute significantly to understanding of conversations embedded in a daily activity.

2. The data and annotation

2.1. The TDU table-talk corpus 2007

The date used in this study was a corpus of three-party conversations in Japanese, recorded in a simulated table-talk setting, which had been developed by Mukawa et al. (2011) at Tokyo Denki University. Among the entire data, three conversations, produced by three different triads, were selected for the current analysis. All the participants were female, and they were triads of high-school students, university students, and married women. They sat across a round table, sharing a platter and a large bowl on the table. In the platter and the bowl served were Chinjao Rosu (Chi-

Clause	Eating rice									Eating meet					
Phrase	Picking-up		Holding		Catching			Taking-in		Catching		Taking-in			
Phase	P	S	P	S	P	H	S	P	S	P	S	P	S	H	R

Figure 2: Hierarchical representation of eating action sequence

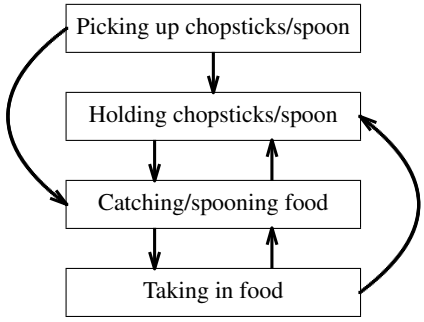

Figure 1: Transition diagram of eating actions

nese stir-fried beef and peppers) and shrimp wonton soup; rice and water were also served individually. The participants were instructed to debate on a prescribed subject, such as the pros and cons of a "pressure-free education" policy, but they could also mention other topics such as the taste of the foods they were eating. The conversation was spontaneously unfolded as in a real table talk.

Three video cameras recorded the upper bodies of the three participants and an additional camera recorded the whole scene. Each conversation lasted about 30 minutes, and a 10-minute excerpt from the middle portion of each conversation was annotated and analyzed in this study, where the participants actively speaking and eating after the dishing-out stage.

The data is not naturally occurring interaction but interaction under an experimental setting. The participants' behaviors observable in the data, however, were quite natural; the data comprises a set of perfectly natural table talks. Although some might consider that recordings from real situations would serve more reliable basis, we believe that we would gain the benefit of detailed and accurate recordings from the experimental situation, which would provide us a good starting point for the research.

2.2. Annotation

Although the original corpus had provided transcriptions and brief annotations of gaze and eating actions, they were not sufficient for the purpose of the current analysis. Thus, we reproduced transcriptions and annotations of gaze and eating actions for the 10-minute excerpt of each conversation. One of the authors performed these tasks by using annotation software ELAN.[1]

2.2.1. Utterances

Transcriptions were segmented into *long utterance-units* according to the scheme proposed by Den et al. (2010), which identifies utterances in spoken Japanese dialogs based mainly on syntactic and pragmatic criteria. Response tokens, produced by listeners as a response to the primary speaker, were also identified based on Den et al. (2011)'s scheme. They were regarded as non-substantial utterances and excluded from the subsequent analyses.

2.2.2. Gaze

Participant's gaze was classified into three categories: i) gaze at an other participant, i.e., either of the two other participants, ii) gaze at her own foods, and iii) others.

2.2.3. Eating actions

Eating actions were classified into four categories: i) picking up chopsticks or a spoon, ii) holding chopsticks or a spoon, iii) catching food with chopsticks or spooning soup, and iv) taking in food or soup. These four types of actions together constitute a chunk of eating actions, such as "eating rice," "eating meet," and "drinking soup." After eating one food, the eater may directly access to another food without putting down and re-holding her chopsticks. Thus, a chunk may also be composed of less than four actions. The diagram in Figure 1 depicts possible transitions between the four types of eating actions.

To represent the time course of an eating action more precisely, each eating action was segmented into several stages based on Kendon's gesture annotation scheme. Kendon (2004) proposed a scheme for annotating the temporal structure of gestures, in which gesture units are sub-divided into a sequence of three phases, i.e., preparation, stroke, and retraction. The stroke (S) phase is the core of the gesture unit. It can be preceded by the preparation (P) phase, in which hands move from the home position towards the gesture space, and followed by the retraction (R) phase, in which hands return from the gesture space to the home position. Before or after the stroke, there may be a hold (H) phase, in which the trajectory and the shape of the gesture is halted and retained (Kendon, 2004; McNeill, 2005).

We adopted this scheme to annotate the temporal structure of eating actions (see Figure 2). For instance, in the "catching-food" action, the movement of the dominant hand, which holds chopsticks, from the stable position towards the food is the preparation, and snaring the food with the chopsticks is the stroke. If the chopsticks, having reached the food, are held there before catching the food, it is marked as hold. Similarly, in the "taking-in-food" action, bringing the food to the mouth is the preparation, and putting the food in the mouth is the stroke. If the chopsticks stay in the mouth for a while, it is marked as hold. Furthermore, putting down the chopsticks on the table is the retraction.

In eating actions, the hold phase occurs ubiquitously, not restricted to the pre-stroke and the post-stroke positions. For instance, in the "taking-in-food" action, the movement of the chopsticks from the dish to the mouth may be paused in

[1] http://www.lat-mpi.eu/tools/elan/

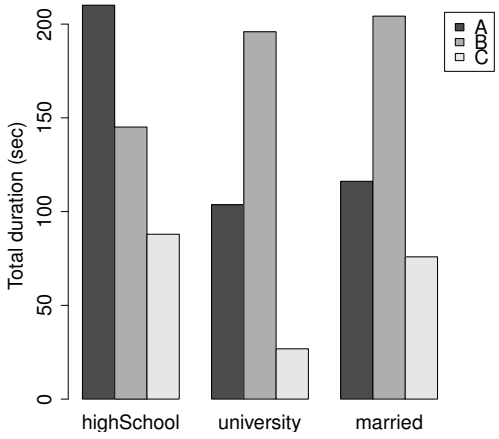

Figure 3: Total duration of utterances

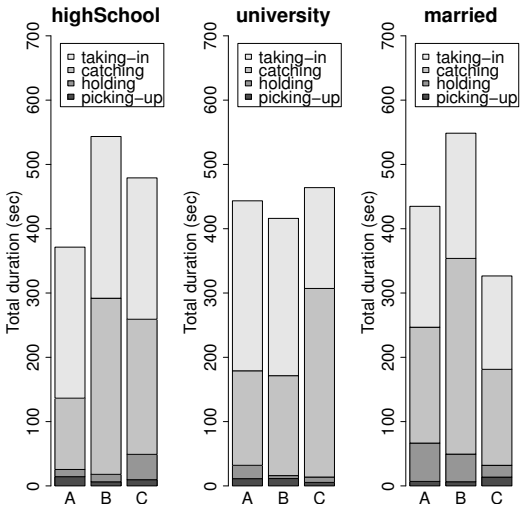

Figure 4: Total duration of eating actions

mid-course, resulting in a hold phase embedded in a preparation phase. Similarly, returning of the chopsticks from the mouth to the table may be paused, resulting in mid-retraction hold. We made no distinction between these hold phases occurring at different positions.

Now, we can represent a sequence of eating actions in a hierarchical way as in Figure 2. Eating actions, defined in Figure 1, correspond to (gesture) *phrases* in the gesture annotation scheme. Above phrases are chunks of eating actions, such as "eating rice," which we metaphorically call *clauses*; below phrases, there are *phases* of actions that compose each eating action.

In this study, we annotated only eating actions performed by the dominant hand, which is the right hand for all 9 participants. Although the non-dominant hand is also used in an eating action—it is usual for Japanese to eat rice with holding chopsticks in the dominant hand and a bowl in the non-dominant hand, in our data the majority of eating actions were performed by the dominant hand.

3. Preliminary analysis of eating actions

In this section, we analyze eating actions in our table-talk data, aiming at an initial investigation of how participants coordinate eating and speaking in table talk. Immediate assumption that comes to our mind would be that the participant controls her eating and speaking actions in a *time-sharing* manner. That is, the participant would schedule a time for eating and a time for speaking exclusively. To see if this assumption is correct, we focus on the relationship between the time devoted to each type of eating action and participant's engagement in speech activity.

3.1. Engagement in speech activity

Some participants were more actively engaged in speech activity than others; they produced more utterances than others. Figure 3 shows the total duration of substantial utterances produced by each participant in each (10-minute excerpt of) conversation.[2]

[2]Response tokens, such as *un*, *hee*, *soo* (*I think so*), and *naruhodo* (*really*), were excluded from substantial utterances. Intra-utterance pauses were included in the duration.

In every conversation, there was an active speaker, who uttered most among the three participants: A in the high-school-student group, B in the university-student group, and B in the married-woman group. In the subsequent analyses, we will pay particular attention to these participants.

3.2. Eating actions with respect to types

Figure 4 show the total duration of each participant's eating actions, classified into the four categories in Figure 1. For all the participants, most of the time was devoted to the "catching-food" and the "taking-in-food" actions, and the remaining two actions, i.e., the "picking-up-chopsticks" and the "holding-chopsticks" actions, occupied only a small amount of time.

There were no reliable correlations between the duration of utterances and the duration of "catching-food" actions ($p = .63$) or between the duration of utterances and the duration of "taking-in-food" actions ($p = .15$). It seems that the participants spent time for eating regardless of how deeply they were engaged in speech activity. This is counterintuitive if we assume that eating and speaking are controlled in a time-sharing manner, which would expect a negative correlation.

One possibility is that the participants did eating actions even when they were speaking. To see this possibility, we next focus on eating actions performed during speaking.

3.3. Eating actions while speaking

Figure 5 shows the total duration of each participant's eating actions, classified into the four categories, for the data limited to those performed while producing a substantial utterance. Interestingly, the three active speakers, i.e., A in the high-school-student group, B in the university-student group, and B in the married-woman group, devoted a considerable time for the "taking-in-food" action.

This is statistically evident; there was a strong and reliable correlation between the duration of utterances and the duration of "taking-in-food" actions while speaking ($r = .84$, $p < .005$). The more the participant was engaged in speech activity, the more time she spent for the "taking-in-food" action when she was speaking. No such correlation was

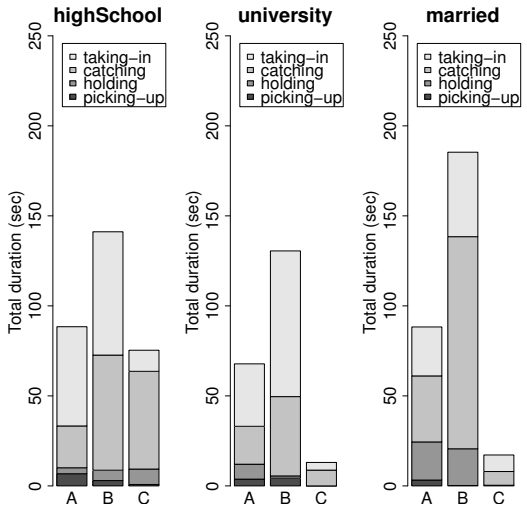

Figure 5: Total duration of eating actions (while speaking)

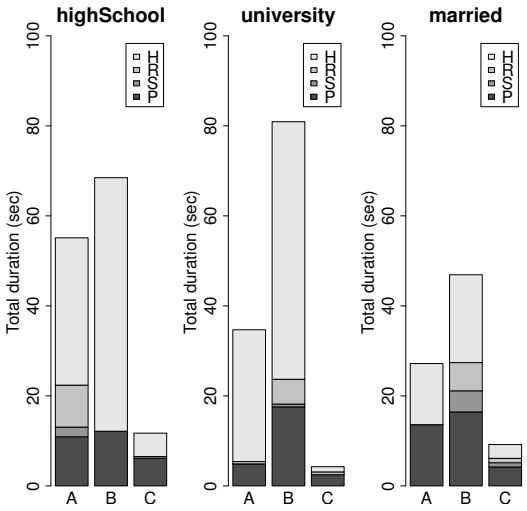

Figure 6: Total duration of phases in the "taking-in-food" action (while speaking)

found when the participant was not speaking ($p = .94$), meaning that the active speakers did not compensate for the lack of time for eating when they were not speaking.

3.4. Phases in "taking-in-food" actions while speaking

To understand more accurately the result shown in the previous section, we examined the duration of the time devoted to each of the four action phases, i.e., preparation, stroke, retraction, and hold, in the "taking-in-food" action while the participant is speaking. Figure 6 shows the result. Obviously, the hold phase occupied the majority of the time, and the time devoted to the hold phase was correlated with the total duration of utterances, although the correlation was moderate ($r = .67$, $p < .05$).

It is now clear that the active speakers coordinated their eating actions with their speech by an elaborate means; instead of compensating for the lack of time for eating when they were not speaking, they were involved in eating actions even when they were speaking, but sometimes halted the movement, and retained the location, of the hand with chopsticks and food before putting food in the mouth.

4. Concluding remarks

In this paper, we developed a scheme for annotating eating activity in multi-party table talk, and showed that active speakers were involved in eating actions even when they were speaking, but locally coordinated their eating actions with the speech by halting an ongoing eating action. A next interesting question would be when this halting of eating action occurs. To answer this question, we have to pay more attention to its sequential contexts in the data. Our detailed annotation would enable us to conduct such a fine-grained analysis easily, and that is our future direction.

5. Acknowledgement

The authors would like to thank Prof. Naoki Mukawa and the members of his laboratory for their giving us permission to use their corpus in our research.

6. References

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

Y. Den, H. Koiso, T. Maruyama, K. Maekawa, K. Takanashi, M. Enomoto, and N. Yoshida. 2010. Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme. In *Proceedings of LREC2010*, pages 2103–2110, Valletta, Malta.

Y. Den, N. Yoshida, K. Takanashi, and H. Koiso. 2011. Annotation of Japanese response tokens and preliminary analysis on their distribution in three-party conversations. In *Proceedings of Oriental COCOSDA 2011*, pages 168–173, Hsinchu, Taiwan.

J. W. Du Bois and R. Englebretson, 2005. *Santa Barbara corpus of spoken American English, Part 4*. Linguistic Data Consortium, Philadelphia.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICASSP-92*, pages 517–520, San Francisco, CA.

C. Goodwin. 1984. Notes on story structure and the organization of participation. In J. M. Atkinson and J. Heritage, editors, *Structures of social action: Studies in Conversation Analysis*, pages 225–246. Cambridge University Press, Cambridge.

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

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

N. Mukawa, H. Tokunaga, M. Yuasa, Y. Tsuda, K. Tateyama, and C. Kasamatsu. 2011. Analysis on utterance behaviors embedded in eating actions: How are conversations and hand-mouth-motions controlled in three-party table talk? *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences (Japanese edition)*, J94-A:500–508.

Classifying the feedback function of head movements and face expressions

Patrizia Paggio*±, Costanza Navarretta*
University of Copenhagen*, University of Malta±
patrizia.paggio@um.edu.mt±, costanza@hum.ku.dk*

Abstract

This paper deals with the automatic classification of the feedback function of head movements and facial expressions in the Danish NOMCO corpus, a collection of dyadic conversations in which speakers meet for the first time and speak freely. Two classification tasks are carried out with good results. In the first one, head gestures with a feedback function are learnt. In the second one, the direction of the feedback – whether given or elicited – is predicted. In both cases, we achieve good accuracy (an F-score of 0.764 in the first task and 0.922 in the second), and the best results are obtained when features concerning the shape of both gesture types as well as the words they co-occur with are taken into consideration.

Keywords: multimodal annotated corpora, feedback, classification

1. Introduction

We know from many earlier studies (Yngve1970, Duncan 1972, McClave 2000) that head movements play a crucial role in conversation, in that they are used to give and elicit feedback, to regulate turn taking, to express focus, and to mark syntactic and prosodic phrase boundaries. Especially their function in expressing different types of feedback has been emphasised not only for English, but also for languages as different as Japanese (Maynard1987), Chinese (Lu and Allwood2011) and Danish (Paggio and Navarretta 2010).

Recently, considerable attention has been given to the issue of how to apply machine learning algorithms to multimodal corpora, with the twofold purpose of making sense of data that are often very large and complex, and of training models for the generation of intelligent gestural behaviour in conversational agents. For example, Reidsma et al. (2009) show that there is dependence between focus of attention (a combination of head, gaze and body features) and the assignment of dialogue act labels. Feedback expressions (head nods and shakes) have also been successfully predicted from speech, prosody and eye gaze in interactions with embodied agents as well as human communication (Fujie et al. 2004, Morency et al. 2009). However, none of these studies focus particularly on how head movements *in combination* with facial expressions can be used to predict feedback behaviour. In previous studies (Paggio and Navarretta 2010, Navarretta and Paggio 2010) we trained classifiers for the recognition of dialogue acts in a map-task corpus of Danish based on multimodal behaviour (head, face and speech) and obtained promising results.

In a recent study (Paggio and Navarretta 2011) we began to experiment with the automatic classification of multimodal feedback behaviour in a conversational corpus, the NOMCO corpus. In this paper, we return to the same issue, but using data containing richer annotations which allow us to experiment with different combinations of features and ultimately to achieve better accuracy. Our goal is to i. predict which of the head gestures in our corpus are used to express feedback as opposed to other conversational functions, ii. classify feedback gestures in terms of their direction, i.e. whether they are being used to give or to elicit feedback.

2. The Danish NOMCO corpus

The corpus used in this study is a collection of dyadic conversations in Danish between subjects who meet for the first time. It was recorded and annotated within the NOMCO (Nordic Multimodal Communication) project, and is one of a number of parallel multimodal corpora showing different types of interaction in Swedish, Danish, Finnish and Estonian (Paggio et al. 2010).

2.1 The recordings

The Danish NOMCO first encounters corpus consists of 12 dialogues for a total of about an hour of interaction. The participants, six males and six females, all native speakers of Danish, did not know each other beforehand. Each subject participated in two interactions, one with a female and one with a male. Subjects were standing opposite each other, and were recorded by three cameras, one taking a long shot of their entire bodies from the side, and the other two taking mid shots of them from different angles. The two views are shown in Figure 1.

Figure 1: Recordings from the Danish NOMCO dialogues: total and split views

2.2 The annotation

Speech was orthographically transcribed and aligned at the word level in Praat. Gestures were annotated with ANVIL (Kipp 2004) using a subset of the features defined in the MUMIN coding scheme (Allwood et al. 2007).

Head movements and facial expressions are annotated in different tracks, so their mutual correspondence is given by temporal overlap. The gesture annotation features are shown in Table 1.

Attribute	Value
HeadMovement	Nod, Jerk, HeadForward, SideTurn
	HeadBackward,Tilt, Shake,Waggle,
	HeadOther
HeadRepetition	Single, Repeated
General face	Smile, Laugh, Scowl, FaceOther
Eyebrows	Frown, Raise, BrowsOther
FeedbackBasic	CPU, SelfFeedback, FeedbackOther
FeedbackDirection	FeedbackGive, FeedbackElicit, FeedbackGiveElicit, FeedbackUnderspecified

Table 1: Annotation features for gestural behaviour

Features related to the shape of the gestures are self-explanatory. The functional features only concern feedback and are inspired by the framework advocated by Allwood et al. (1992), where feedback is described as unobtrusive behaviour that has the purpose of either signalling or eliciting signals of *contact*, *perception* and *understanding* (CPU). FeedbackBasic is thus used to annotate if the feedback involved includes all three aspects (CPU), or only one of them (FeedbackOther). The feature SelfFeedback is used to annotate speakers' head movements expressing a comment to their own speech. FeedbackDirection indicates whether the gesture is a feedback signal or elicits a feedback signal.
Finally, each gesture is explicitly linked to a sequence of semantically related words: words uttered by the subject who is producing the gesture or produced by the other person. Words comprise here normal words and expressions such as *hm* and *ehm*.
An inter-coder agreement test was run in order to test to what extent three coders identified the same gestures and assigned the same categories to the recognised gestures. The results in terms of Cohen's kappa (1960) were in the range 0.6-0.7 for face attributes and 0.7-0.9 for head movements. The highest disagreement values for facial expressions were mainly due to disagreement on segmentation.

2.3 Corpus analysis

So far, 10 videos in the corpus have been analysed[1] and used for the machine learning experiments described below. The total duration of the videos is 3,280 seconds, they contain 12,032 words (including filled pauses) and a total of 3,511 head movements and facial expressions. The distribution of the gestures is shown in Table 2, together with frequency figures (gesture per word and gesture per second).

Gesture type	#	g/w	g/s
All gestures	3511	0.29	1.07
Head	2335	0.19	0.71
Face	1176	0.09	0.35

Table 2: Gestures in the Danish NOMCO corpus

The average number of head movements per person is 129.72, with a standard deviation of 34.68, whilst the average number of facial expressions per person is 61.89, with a standard deviation of 24.41. The data show a good variety of head movement types (Table 3).

Head movement	#
Tilt	403
SideTurn	333
Repeated Nod	324
HeadBackward	256
Simple Nod	224
HeadForward	207
Repeated Shake	182
HeadOther	161
Jerk (UpNod)	122
Simple Shake	62
Waggle	61
Total	2335

Table 3: Head movements in the Danish NOMCO corpus

Nods (either single or repeated), are the most common, followed by tilts and side turns. Face expressions are smiles, laughs or scowls, accompanied or not by eyebrow raises or frowns. Eyebrow raises and frowns may also occur on their own. 50% of all gestures (head and face) express feedback. Head is the preferred modality when it comes to feedback (60% of all feedback gestures), which is not surprising since the head is the preferred modality of gestural expression in general. The exact distribution among the 11 most frequent types, making up for 95% of all the feedback gestures, is shown in Table 4.

Gesture type	%
Smile	18
Repeated Nod	16
Raise	10
Single Nod	9
Tilt	8
HeadBackward	6
HeadForward	6
Jerk (UpNod)	6
Laughter	6
Shake	6
SideTurn	5
Total	95

Table 4: Most frequent feedback gestures in the Danish NOMCO corpus

Finally, feedback gestures fall into different types depending on the feedback direction. FeedbackGive is by far the most frequent type, followed by FeedbackElicit. There are also some cases of mixed direction, labelled FeedbackGiveElicit, and a few in which the annotators could not choose a specific value and assigned the label FeedbackUnderspecified (exact figures are given below in Section 3).

3. Classification experiments

3.1 The tasks

The following two tasks were defined for the study: a) classification of head movement function; b) classification of feedback type.
For the first task, our hypothesis was that, based on formal characteristics of the head movements and the related

[1] In one of the videos, only the gestures of one of the speakers had been annotated at the time this study was conducted.

facial expressions, as well as the word tokens that the movements are linked with, it would be possible to distinguish feedback gestures from non-feedback gestures. Differing from our previous work (Paggio and Navarretta 2011), here we have not only access to gesturer's but also to the other person's words. In addition, we are taking into consideration a greater number of gestures.

As for the second task, we wanted to investigate to what extent, assuming the same features used in task one, plus knowledge of whether the gestures are feedback gestures, it would be possible to distinguish between feedback giving and feedback eliciting gestures. However, it must be noted that the data on this point are strongly biased in that the attribute FeedbackGive is by far the most frequent.

3.2 The dataset

Head movement and facial expression features annotated in ANVIL were combined if they are performed by the same participant, and they overlap temporally. Temporal overlap was calculated by taking into account start and end points of each gesture given in milliseconds. No restriction was posed on the temporal overlap.

Since overlapping gestures can have different durations, one gesture type can overlap with more gestures of the other type and viceversa. Head movements are more frequent than facial expressions in our data, thus we have extracted the annotations of all head movements, and we have added to them the annotations of the overlapping facial expressions. If a head movement overlaps with two facial expressions one after the other, it is represented in the dataset as two instances with the same head movement features, but with different facial expression features. This may seem somewhat artificial. On the other hand, since this situation only happens a limited number of times, we decided to accept the duplication. In the opposite case, in which a head movement has no overlapping facial expression, a *None* value is assigned to all facial features. The resulting dataset consists of 2725 elements, that is 390 elements more than the observed head movements. Out of these, 1205 have an overlapping facial expression. Concerning the different feedback types, 80% of the CPU feedback movements are annotated as FeedbackGive, 18% as FeedbackElicit, and 2% as either FeedbackGiveElicit, or as FeedbackUnderspecified. The instances of SelfFeedback, which in the table is counted together with FeedbackNone, are 77. The counts for the various types are shown in Table 5.

Head movement function	**#**
FeedbackGive	995
FeedbackElicit	221
FeedbackGiveElicit	30
FeedackUnderspecified	2
Total Feedback	1248
Feedback None/SelfFeedback	1477
Total head movement	2725

Table 5: Head movement instances in the dataset

3.3 The results

Given the datasets, for each of the tasks we tested classifiers with different combinations of gesture and speech features. All features were extracted from the manually annotated data. The results for the first task are shown in Tables 6 to 8 in terms of Precision (P), Recall (R) and F-score (F). We experimented first only with head movement features (Table 6), then only with facial expression features (Table 7), and finally with head and face in combination (Table 8). In all tables, the first row shows the baseline obtained with the ZeroR algorithm, which always chooses the most frequent class. The remaining rows show results obtained with a support vector classifier, sequential minimal optimisation (SMO), on more feature combinations: i. only shape features (either Head, Face or Head and Face together), ii. shape features plus the co-occurring words by the gesturer, iii. the co-occurring words by the other person, and iv. the co-occurring words by both participants.

Classifier	**P**	**R**	**F**
ZeroR	0.263	0.513	0.348
Head	0.647	0.661	0.647
Head+GesturerWords	0.762	0.739	0.729
Head+OtherWords	0.669	0.675	0.657
Head+AllWords	0.772	0.745	**0.734**

Table 6: Classification of feedback function of head movements.

Classifier	**P**	**R**	**F**
ZeroR	0.245	0.495	0.328
Face	0.503	0.537	0.51
Face+GesturerWords	0.676	0.632	0.622
Face+OtherWords	0.503	0.537	0.51
Face+AllWords	0.688	0.646	**0.635**

Table 7: Classification of feedback function of facial expressions.

As expected, the best result (an F-score of 0.764) is obtained when both head movements and facial expressions are used with all the co-occurring words.

Classifier	**P**	**R**	**F**
ZeroR	0.264	0.514	0.349
Head+Face	0.637	0.647	0.63
Head+Face+GesturerWords	0.782	0.765	0.757
Head+Face+OtherWords	0.67	0.676	0.657
Head+Face+AllWords	0.792	0.772	**0.764**

Table 8: Classification of feedback function in multimodal head gestures (movements and expressions).

If only head movements are considered, the best result is obtained when both gesturer's and non-gesturer's words are used in combination with the gesture shape. If facial expressions are considered alone, the best results are also produced when all words are considered.

In Table 9 we show the results for the second task, i.e. the classification of feedback direction.

Classifier	**P**	**R**	**F**
ZeroR	0.294	0.542	0.381
Head+Face	0.833	0.906	0.865
Head+Face+GesturerWords	0.922	0.929	0.915
Head+Face+OtherWords	0.834	0.906	0.866
Head+Face+AllWords	0.924	0.932	**0.922**

Table 9: Classification of FBDirection

We took both head and face features directly, since this is the combination where feedback can be predicted with the highest accuracy, and classified feedback direction again in the same incremental way that was described above. The FeedbackBasic of the head movements was also used in the classification. Also in this task, the best result (an F-score of 0.922) is produced using both the gesturer and the interlocutor's words.

4. Discussion and conclusions

The results obtained in the classification tasks are generally positive. The classifiers perform better than the baseline in both tasks, and achieve in fact pretty high accuracy. This is especially interesting in relation with the first task, which is quite a complex one, as testified by the fact that the human annotators had an inter-annotator agreement not higher than about 0.6. However, several issues merit discussing.

Concerning the first task of distinguishing head movements related to feedback from others, our expectation that the formal features of the head movements would play a significant role, is confirmed. On the other hand, when the formal features of the corresponding facial expressions are added to the classifier, the accuracy does not improve unless all the co-occurring words are also added. The explanation is presumably the fact that facial expressions are in themselves rather ambiguous with respect to their conversational function, but that their function is clarified by the co-occurring words.

In our previous study (Paggio and Navarretta 2011), the effect of adding the words to the classifiers was not as clear as it is here. Having added both participants' words has increased the accuracy of the results, a fact that seems intuitively understandable.

In general, distinguishing gestures that have a feedback function from those that don't, is a rather simplistic task. Once the full set of functional attributes from the MUMIN scheme is annotated in the corpus, we will be able to experiment with learning several communicative classes, e.g. feedback, turn and sequencing.

As for the second task, the fact that the F-measure value is as high as 0.922, and so much better than the baseline, is largely due to the fact that the classifier trivially discards false positives for the most frequent category, *None*, every time the Feedback value of the gesture is CPU. However, the next most frequent value, FeedbackGive, is also classified correctly 97% of the time, while the percentage goes down to 30% for FeedbackElicit, which only has 221 instances in the corpus.

Although the classifiers described here perform quite well, in future we would like not only to extend our investigation to other communicative functions, but also to investigate how the directly preceding and following context, words as well as gestures, can be used to predict the occurrence of the various functional classes.

Acknowledgments

The NOMCO project is funded by NORDCORP under the Nordic Research Councils for the Humanities and the Social Sciences (NOS-HS).The annotators of the Danish NOMCO corpus have been Sara Andersen, Josephine B. Arrild, Anette Studsgård and Bjørn N. Wesseltolvig.

5. References

Allwood, J., Nivre, J. and Ahlsén, E. (1992) On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26.

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

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

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

Fujie, S., Ejiri,Y., Nakajima, K., Matsusaka, Y. and Kobayashi, T. (2004) A conversation robot using head gesture recognition as para-linguistic information. In *Proceedings of the 13th IEEE Workshop on Robot and Human Interactive Communication*, pp.159 – 164.

Kipp, M. (2004) *Gesture Generation by Imitation – From Human Behavior to Computer Character Animation*. Boca Raton, Florida: Dissertation.com.

Lu, J. and Allwood, J. (2011) A pilot study on feedback in chinese and swedish mono-cultural and in-tercultural interactions. In *Proceedings of the 3rd Nordic Symposium on Multimodal Communication*.

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

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

Morency, L.-P., de Kok, I., and Gratch, J. (2009). A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20:70–84, Springer.

Navarretta, C. and Paggio, P. (2010). Classification of Feedback Expressions in Multimodal Data. *Proceedings of ACL 2010*, Uppsala, Sweden, Juli 11-16, 2010, pp. 318-324.

Paggio, P. and Navarretta, C. (2010) Feedback in Head Gestures and Speech. In *LREC 2010 Workshop Proceedings on Multimodal Corpora* Malta, May 17, pp. 1-4.

Paggio, P. and Navarretta, C. (2011) Learning to classify the feedback function of head movements in a Danish Corpus of first encounters. In *Proceedings of ICMI2011 Workshop Multimodal Corpora for Machine Learning*, Alicante, Spain November, 8 pages.

Reidsma, D., Heylen,D., and op den Akker, R. (2009) On the Contextual Analysis of Agreement Scores. *Multimodal Corpora From Models of Natural Interaction to Systems and Applications*, number 5509 in LNAI, pages 122–137. Springer.

Yngve, V. (1970) On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago.*

Who am I speaking at?
Perceiving the head orientation of speakers from acoustic cues alone

Jens Edlund[1], Mattias Heldner[2], Joakim Gustafson[1]
[1]KTH Speech, Music and Hearing, Stockholm, Sweden
[2]Linguistics, Stockholm University, Stockholm, Sweden
E-mail: edlund@speech.kth.se, mattias.heldner@ling.su.se, jocke@speech.kth.se

Abstract

The ability of people, and of machines, to determine the position of a sound source in a room is well studied. The related ability to determine the orientation of a directed sound source, on the other hand, is not, but the few studies there are show people to be surprisingly skilled at it. This has bearing for studies of face-to-face interaction and of embodied spoken dialogue systems, as sound source orientation of a speaker is connected to the head pose of the speaker, which is meaningful in a number of ways. We describe in passing some preliminary findings that led us onto this line of investigation, and in detail a study in which we extend an experiment design intended to measure perception of gaze direction to test instead for perception of sound source orientation. The results corroborate those of previous studies, and further show that people are very good at performing this skill outside of studio conditions as well.

1. Introduction

Gaze and head pose shifts are central to studies of human face-to-face interaction. They are becoming equally important for spoken dialogue systems research as the interest for embodied systems keep increasing. At the same time, the way in which we collect and use our corpora is changing. More and more corpora are not only multimodal in the traditional sense of containing both audio and video, but also hold other information such as movement data. And increasingly, we study dialogue within the situation: we attempt to model not only the dialogue itself and its semantic context, but facts about the space in which it takes place, about the moods and motivations of its participants, or about the events taking place in its vicinity. Finally, a growing community of researchers focus on developing spoken dialogue systems that are first and foremost humanlike, either because they are convinced that humanlikeness will improve spoken dialogue as a human-machine interface, or because they are interested in testing their hypotheses about how human interaction works.

In light of this altogether more holistic view of dialogue research, we have investigated the extent to which a listener can perceive a speaker's facing angle under normal conversational circumstances. To the extent that human speakers' facing angles are important, the auditory perception of a speaker's facing angle is likely to be important as well. We start out with a background of the area and of related research that serves as motivation for this study, continue with a brief description of the preliminary mini-studies that led us to the present study, and conclude with a detailed description of the present study, its method and its results.

2. Background and related work

The spatial relation between speakers and listeners is an important part of the dynamically changing situation in which a conversation unfolds. This spatial relation can be modelled using corpora in which reliable data describing each participant's orientation and location in the room is available from for example motion capture, such as the Spontal database (Edlund et al., 2010). Modelling the acoustic effects of these spatial relations would require, minimally, the addition of binaural microphones in each participant's ears. No available and sizeable corpus to date holds both binaurally captured sound and positional data.

2.1 Perception of sound source orientation

Whereas studies of people's ability to judge the position of a sound source are plentiful, there are only a handful studies of our ability to judge the orientation of directional sound sources.

In the early 2000s, Neuhoff and colleagues showed that people can indeed distinguish between different orientations of a directional loudspeaker. Neuhoff (2001) shows subjects' ability to detect the facing angle of a loudspeaker playing recorded speech in an empty room, and find that factors influencing this ability include whether the sound source is stationary or rotating (the movement helps); the distance to the sound source (closer is better); and the facing angle itself (the task is easier when the loudspeaker faces the listener straight on). Neuhoff et al. (2001) determines a just noticeable difference (JND) for facing angles by having subjects judge the orientation of a loudspeaker producing broadband noise in an anechoic chamber. As predicted by the findings in Neuhoff (2001), the JND varies with the distance to the loudspeaker and with the facing angle itself. The work is brought together and discussed in Neuhoff (2003), where greater weight is given to the bearing of these results on spoken interaction research. Neuhoff and colleagues implicate the inter-aural level difference (ILD) as the most likely cue to sound source orientation.

Kato and colleagues later took the potential relevance for realistic human-to-human telecommunication as their main motivation to perform similar studies. Kato et al. (2010a) and Kato et al. (2010b) both report on a study where a male speaker poised on a pivot chair in an anechoic chamber speak utterances at different horizontal

and vertical angles. We focus on the horizontal angles here. 12 blindfolded listeners were asked to indicate the speaker's facing direction. The results, including an average horizontal error of 23.5 degrees, are comparable to or better than those achieved with loudspeakers, adding evidence to the idea that interlocutors may be able to hear the head pose of the speaker from acoustic cues alone. A clear effect of the facing angle was observed, with head-on utterance being much easier to judge correctly. Kato and colleagues also analyse the acoustic transfer function from a speaker's mouth to the ears of a listener using binaural microphones, and like Neuhoff and colleagues, they find ILD to be the prime cue for horizontal orientation.

Finally, Nakano et al. (2008) and Nakano et al. (2010) contributed a comparison between perception in what they term a *real environment* - a normal room stripped bare of all furniture - and an anechoic chamber. Their stimuli is a live human speaker. Their subjects do better in the anechoic chamber. They also compare performance before and after a training session, and get an improvement from training.

2.2 Sound source orientation and face-to-face interaction

It is well attested that gaze, and in particular mutual gaze is important for the interaction in face-to-face dialogue. A typical gaze pattern, at least in Europe and in Northern America, is that the listener looks fairly constantly at the speaker, while the speaker looks at the listener in the vicinity of speaker changes or backchannels (e.g. Bavelas & Gerwing, 2011; Kendon, 1967). Hence, auditory perception of speaker facing direction might provide a redundant correlate of gaze in visible conditions, and a correlate of gaze in non-visible face-to-face conditions, such as in the dark. Note also, as mentioned above, that several studies report that listeners are particularly sensitive when the sound source is directed straight at them, that is, the situation correlated to mutual gaze in visible conditions.

2.3 Sound source orientation and embodied spoken dialogue systems

Currently, there are no interactive systems that detect and make use of sound source orientation, and systems that use gaze and head pose as a part of their expressive repertoire routinely produce audio through fixed loudspeakers without concern for what the acoustic effects of the head movements they display would be. Nakano et al. (2010), however, show a machine trained on acoustic data from an array microphone that perform better than chance but poorer than human subjects on the task of detection the facing angle of a speaker.

Given the importance of gaze in face-to-face interaction, there is considerable scope for improving the interactional capabilities of interactive avatars and robots by endowing them with means to produce and perceive visible as well as audible facing direction.

3. Preliminary studies

The idea that speaker head orientation may be heard by listeners struck us for no good reason during a conversation about turntaking a number of years ago. The thought immediately fascinated us, and we immediately proceeded to run impromptu tests and to track down and read up on the work of Neuhoff and colleagues, but time constraints came in the way of proper replication and publication. The tests we did run had a few things in common. They tested five orientations only - head on towards the listener, and 45 as well as 90 degrees in either direction. We felt that those directions were sufficient to study the effects the acoustics of face orientation might have on spoken face-to-face interaction. We used a real human speaker reading a predefined sentence, sacrificing the control afforded by a recording replayed in a directional loudspeaker for the ecological validity of a real human speech production apparatus. Tests in a number of environments, including offices, snow-clad fields and noisy bars, and at distances ranging from 1 metre to 10 metres all showed that subjects were able to indicate the direction in which speaker was facing from listening only with an accuracy was much above random choice. As we have recently increased our studies of co-presence (Edlund et al., 2011) as well as our efforts to create situated and embodied conversational partners (Al Moubayed et al., in press), we decided to resume these studies and repeat these tests under more controlled circumstances. And while the studies published to date were all performed in studios or rooms designed to minimize or normalize echoes, we choose to focus on a real everyday environment, sacrificing control for ecological validity.

4. Method

4.1 The subject/target experimental paradigm

We employed an experimental paradigm first used in Beskow & Al Moubayed (2010), where it was developed to allow experimenters to quickly gather large amounts of data on human perception of gaze targets/direction. We have generalized the paradigm here, and adapted it to work for perception of directional audio. In its generalized form, the paradigm is used to gauge subjects' ability to perceive the intended target of a directional stimulus, and can be described as follows.

A group of N subjects are placed in a circle or semi-circle, so that there is one point at their centre which is equidistant to each subject, from which all stimuli are presented (the *centre*). Subjects positions are numbered P_1 to P_N, and the angle between each subject's position, that of the centre, and that of the subject's closes neighbouring subjects ($A(P_1P_{2)}$... $A(P_NP_{1)}$) is calculated. Subjects may or may not be equidistant from their closest neighbours.

All subjects double as targets for the directional stimuli (hence the *subject/target paradigm*). During an experiment, directional stimuli are aimed at each of the subjects. The order is varied systematically, and the number of stimuli is such that each subject is targeted as

Figure 1: The experiment environment

many times as the others in one set of stimuli. A set of stimuli, then, contains a multiple R of N for a total of R*N stimuli. Once one set is completed, the subjects rotate - they shift their positions by one step and the process of presenting a set of N*R stimuli is repeated. The rotation is repeated N times, until each subject has been in each position once, making the total number of stimuli presented in an experiment N*R*N.

Each time a stimulus has been presented, each subject is asked to point out the intended target in such a manner that the other subjects cannot see it. The result is N judgements for each stimulus, for a total of N*R*N*N data points in one experiment. If more than one experiment condition is to be tested, the entire process is repeated from the beginning.

We now turn to the specifics of the present experiment.

4.2 Subjects

Two conditions were tested in a between-group design, and groups with five participants (N=5) were used. The subjects were students and university employees. Four of the subjects were female and six were male. All reported having normal hearing on both ears.

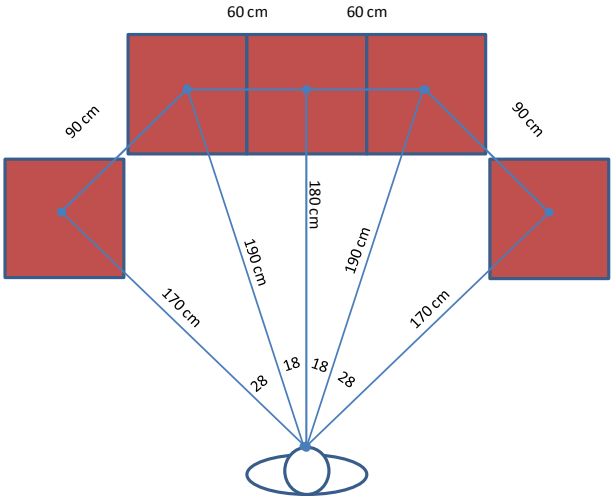

Figure 2: Schematic view of the experimental setup

4.3 Spatial layout and surroundings

The main motivation for the experiment was to test the subjects' ability to perceive acoustic (speech) directionality in normal, everyday conditions. For this reason, an existing recreational sofa group in busy office surroundings was chosen, and no attempt were made to stop other people from walking through the area or talking nearby. The sofa group was left standing as it is normally, and subjects were seated in five of the seats, as seen in Figure 1. A result of this was that the distance to the nominal "centre" from which stimuli were presented was not identical for all seats. The actual measures are shown in Figure 2, which also shows the distances and angles between adjacent subjects.

4.4 Stimuli

The experiment conductor spoke the sentence "Who am I speaking to now", while facing one of the subjects head-on from the nominal centre position. Each group contained two readings directed at each target (R=2) for a total of ten readings, after which the subjects were rotated.

4.5 Conditions

A between-group design was employed, in which the first group (NOFEEDBACK) were presented with stimuli exactly as described above, while the second group (FEEDBACK) received feedback after each utterance, once all five judgements had been recorded. Feedback consisted of the reader saying "I was talking to number N", where N was a number between 1 and 5 referring to the five seats from left to right. The subjects in this group had been informed about this procedure beforehand.

4.5 Responses

The subjects used hand signs to show which listener they thought the reader was facing: one, two, three or four fingers on the left hand to signify one, two, three and four steps to the left, respectively; one, two, three or four fingers on the right hand to signify one, two, three and four steps to the right; and a pointing gesture towards the chest to signify themselves (see figure 3).

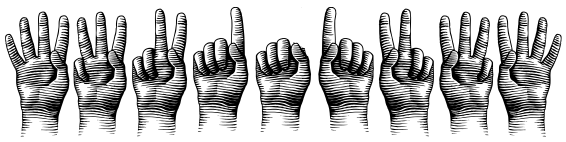

Figure 3: Signs used to indicate target position

All in all, the utterance was spoken 5*2*5=50 times for each condition. With five responses for each utterance, a total of 500 judgements were collected, 250 for each group and condition.

5. Results

Combined over the two conditions, the subject got the target exactly right in 259 out of 500 cases, or 52 % of the time. Random choice yields a 20 % baseline, and chi-square test shows that the result deviates significantly from a random choice ($\chi^2(1, N=24)=488.79, p=0.0001$).

The confusion matrix for all data is shown in Table 1.

Variance analysis of the errors (ANOVA), assuming equidistant positions, show significant main effects for condition, with the FEEDBACK condition resulting in a smaller error ($F(1,496)=4.23; p=.04$). No main effects

were found for gender *(F(1,496)=0.23; p=.63)*, nor were there any interactions between gender and condition.

Table 1. Confusion matrix for all subjects and conditions.

Target position \ Estimated target position	1	2	3	4	5	Total
1	**62**	23	7	8	0	**100**
2	13	**40**	38	8	1	**100**
3	9	15	**47**	24	5	**100**
4	1	8	37	**35**	19	**100**
5	0	2	6	17	**75**	**100**
Total	**85**	**88**	**135**	**92**	**100**	**500**

6. Discussion and future work

The results of the present study show that listeners are quite good at distinguishing between different facing angles in a speaker not only in anechoic chambers and emptied out, silent rooms, but also under conditions in which conversations normally occur - in furnished, asymmetric rooms with background noise and people passing by. This is consistent with an idea that the acoustic properties of speech and facing angle may be a redundant cue that interlocutors take into consideration in face-to-face spoken interaction. We further argue that modelling the acoustic properties of speakers' position and orientation is an important step in achieving a realistic model of situated interaction.

The data (see Table 1) also indicate that some directions in our fully furnished environment were easier to detect than others. This suggests that listeners use more than ILD to judge the facing angle of a speaker, but rather maintain an model of their acoustic environment into which they fit acoustic stimuli. As an example, when the speaker faced straight towards the large window set on his right side, subjects on all seats were more likely to judge the direction correctly, possibly due to the special acoustic character of the reflection against the window. This leads us to our next goal: to compare listeners' performance in everyday environments to anechoic chambers. If models of the acoustic environment are involved, one might expect poorer performance in an anechoic chamber; if it is all IDL, the anechoic chamber should instead help.

7. Acknowledgements

This work was funded by the Riksbankens Jubileumsfond (RJ) project *Prosody in conversation* (P09-0064:1-E).

8. References

Al Moubayed, S., Beskow, J., Skantze, G., & Granström, B. (in press). Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction. To be published in Esposito, A., Esposito, A., Vinciarelli, A., Hoffmann, R., & C. Müller, V. (Eds.), *Cognitive Behavioural Systems. Lecture Notes in Computer Science*. Springer.

Bavelas, J. B., & Gerwing, J. (2011). The listener as addressee in face-to-face dialogue. *International Journal of Listening, 255*(3), 178-198.

Beskow, J., & Al Moubayed, S. (2010). Perception of Gaze Direction in 2D and 3D Facial Projections. In *The ACM / SSPNET 2nd International Symposium on Facial Analysis and Animation*. Edinburgh, UK.

Edlund, J., Al Moubayed, S., & Beskow, J. (2011). The Mona Lisa Gaze Effect as an Objective Metric for Perceived Cospatiality. In Vilhjálmsson, H. H., Kopp, S., Marsella, S., & Thórisson, K. R. (Eds.), *Proc. of the Intelligent Virtual Agents 10th International Conference (IVA 2011)* (pp. 439-440). Reykjavík, Iceland: Springer.

Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., & House, D. (2010). Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., & Tapias, D. (Eds.), *Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (pp. 2992 - 2995). Valetta, Malta.

Kato, H., Takemoto, H., Nishimura, R., & Mokhtari, P. (2010a). Spatial acoustic cues for the auditory perception of speaker's facing direction. In *In Proc. of 20th International Congress on Acoustics, ICA 2010*. Sydney, Australia.

Kato, H., Takemoto, H., Nishimura, R., & Mokhtari, P. (2010b). On the human ability to auditorily perceive human speaker's facing angle. In *In Proc. of the 4th International Universal Communication Symposium (IUCS), 2010* (pp. 387 - 391). Beijing.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 26*, 22-63.

Nakano, A. Y., Nakagawa, S., & Yamamoto, K. (2010). Auditory perception versus automatic estimation of location and orientation of an acoustic source in a real environment. *Acoustical Science and Technology, 31*(5), 309-319.

Nakano, A. Y., Yamamoto, K., & Nakagawa, S. (2008). Auditory perception of speaker's position, distance and facing angle in a real enclosed environment. In *Proc. of Autumn Meeting of Acoustic Society of Japan* (pp. 525-526).

Neuhoff, J. G., Rodstrom, M-A., & Vaidya, T. (2001). The audible facing angle. *Acoustics Research Letters Online, 2*(4), 109-114.

Neuhoff, J. G. (2001). Perceiving acoustic source orientation in three-dimentional space. In *Proc. of the International Conference on Auditory Display*. Espoo, Finland.

Neuhoff, J. G. (2003). Twist and shout: audible facing angles and dynamic rotation. *Ecological Psychology, 15*(4), 335-351.

Incremental Collection of an Activity-based Multimodal Corpus and its Use in Activity-based studies

Jens Allwood & Elisabeth Ahlsén
SSKKII Interdisciplinary Center/Communication and Cognition
University of Gothenburg
jens@ling.gu.se, eliza@ling.gu.se

Abstract

Activity-based communication Analysis is a framework, which puts social activity in focus and analyzes communication in relation to the determining and determined factors of the activity. Given an activity-based approach, it is essential to collect multimodal corpora with a variation of social activities, in order to study similarities, as well as differences between activities and possible influencing factors. The Gothenburg Spoken Language Corpus was collected as a corpus representing communication in a wide range of social activities. The paper describes and briefly discusses the purpose and some of the features of the corpus. The usefulness of activity-based multimodal corpora is exemplified by the analysis of spoken feedback in a specific activity (the physical examination in doctor-patient interaction).

The framework of Activity-based Communication Analysis (ACA) for Studying Multimodal Communication

Activity-based communication Analysis is a framework developed by Allwood (1976, 2000, 2001), which puts social activity in focus and analyzes communication in relation to influencing and influenced factors of an activity. The framework is on inspired by work in philosophy, linguistics, anthropology, psychology and sociology and sees communication as action , involving degrees of coordination, collaboration and cooperation, in particular social activities.

Some influencing factors in an activity are global, i.e. influence the activity as a whole, while others are local, i.e. influence specific parts of an activity. Some of the influencing activity factors are collective, which means that they influence all participants in an activity, while others are individual, which means that they influence only individual participants.

Besides influencing factors, there are influenced parameters in the interaction, which can also be global or local and collective or individual. Among the collective Among the collective influenced parameters we find for example, interaction patterns which are produced collaboratively, while examples of individual influenced parameters are particular traits of communicative behavior or particular traits of perceiving/understanding speech and gestures for each of the participants.

A summary of the framework is found in table 1 below.

Influencing factors of an activity
- Collective: purposes and function of the activity, roles of the activity, the sub-activity structure of the activity, artifacts and other instruments used in the activity as well as social and physical environment of the activity - Individual: goals of the individual participants, individual role interpretations, individual artifacts as well as individual interpretations of the environment
Influenced factors in an activity
- Collective: interaction patterns, such as those to be found in interactive communication management (turn management, feedback patterns and sequences) - Individual: communicative behavior and perception of communication particular to individual participants (e.g. production and perception of vocabulary, grammar, pronunciation, gestures)

Table 1. Summary of Activity-based Communication Analysis

The ACA theory and framework rest on a strong belief that activity factors are important and lead to important differences between social activities, so that normally only some of the findings based on a study of communication in a particular social activity can be generalized to other social activities. Understanding how activity variation affects features of communication is therefore an important goal of using this framework. Since both the physical conditions (non-communicative, but nonetheless informative) actions as well as the use of gestures, tools etc. can vary

between social activities, an analysis of multimodal communication is always relevant.

The Need for Multimodal Corpora

Given an activity-based approach, it is essential to collect multimodal corpora with a variation of social activities. This makes possible a study of similarities, as well as of differences between activities and possible influencing factors. Depending on available resources, an activity-based corpus can be collected during a limited period of time as a project in itself or incrementally, by accumulating multimodal recordings from different projects, involving different activity types, as in the corpora described in this paper. The purpose of the corpus is an important initial consideration. The corpora presented in this paper have all been collected with the purpose of studying multimodal interaction in different, mostly naturalistic, settings, with a priority on ecological validity. This emphasis on field recordings means that the best possible quality, given this condition, has been achieved, but that naturalness has been more important than studio quality. Since face-to-face interaction has been prioritized, most of the field recordings have been made using one camera, with all participants visible in the same picture. In studio made recordings, three cameras have sometimes been used, together with separate microphones, in order to make the material useful for in-depth analysis of, for example, facial expressions and speech characteristics. But the field recordings made of naturalistic interactions, where the recording did not interfere too much, provide enough information to study what was said, what gestures were used, how the body posture varied etc. in

Incremental Data Collection and Structure of the Corpus

Since the early 1980's, the Gothenburg Spoken Language Corpus (GSLC) has been incrementally collected, i.e. new social activities have gradually been added from different projects and other sources. The corpus consists of mostly videorecorded interactions in Swedish from 25 general activity types, see table 2. The size of the corpus is around 1 400 000 transcribed words. The included activities are: Arranged discussions, Auction, Bus driver/passenger, Church, Consultation, Court, Dinner, Discussion, Factory conversation, Formal meeting, Games & play, Hearing, Committe on the constitution, Hotel, Informal conversation, Interview, Lecture, Market, Meeting, Phone, Political debate, Retelling of article, Role play, Shop, Task-oriented dialogue, Therapy, Trade fair, Travel agency, and TV

Since the corpus is dynamic and grows mainly by the inclusion of new activities from new projects, there is more material from some activities and less from others, something which has to be taken into account when activities are compared. One important feature of an activity-based corpus is to have metadata organized, so that different features can be extracted and compared. The GSLC videorecordings, transcriptions and codings have headers with some of the metadata easily available and retreivable. A corpus browser allows different search procedures based on the transcriptions and headers. Table 2 presents an example header with metadata. All names are pseudonyms.

@ Activity type, level 1: Consultation
@ Activity type, level 2: -
@ Activity type, level 3: P-D: Radiation
@ Recorded activity title: Patient-Doctor Conversation: Radiation Control
@ Recorded activity date: 890914
@ Recorded activity ID: A500302
@ Transcription name: A5003021
@ Transcription System: MSO6
@ Duration: 00:07:53
@ Short name: Radiation
@ Participant: D = (Dr. Bengtsson)
@ Participant: P = (Patient)
@ Anonymized: yes
@ For external use: no
@ Kernel: yes
@ Transcriber: Unknown
@ Transcription date: 950815
@ Checker: Elisabeth Kovacs
@ Checking date: 950828
@ Project: doctor-patient conversations
@ Comment:
@ Time coding: yes
@ Transcribed segments: all
@ Tape: a5003, ka5003
@ Section: 1: Start
@ Section: 2: Main reason
@ Section: 3: Physical
@ Section: 4: Diagnosis
@ Section: 5: History
@ Section: 6: Ordination
@ Section: 7: Diagnosis end
@ Section: 8: Frame
@ Section: 9: End

Table 2. The header (excluding basic statistics)

As we can see in table 3, the activity is further divided into sub-activities or sections, which often have specific characteristics.

A more advanced relational database could also be very useful, but requires more administrative effort and is not as easily available to users of the corpus.

The videorecorded and/or audiorecorded activities have all been transcribed, using a standardized format the Modified Standard Orthography (MSO6) (Nivre, 1999) and the Gothenburg Transcription Standard (GTS 6.4) (Nivre, 2004). The transcriptions have been checked by a second transcriber and by a transcription checking tool, in order to ensure that they can be merged and that a number of tools for calculating types of behavior, making concordances of words, counting and sorting various features can be used. The transcriptions can be

used in different formats: e.g. the transcribed spoken language variant and the written language equivalent variant. This enables a transcription close to speech for spoken language analysis and a written language version for comparisons between spoken and written language. Some of the activities have also been annotated for multimodal communication, either using the comment function of GTS or using multimodal transcription tools, such as Praat and ANVIL. Other annotations have also been made for specific purposes.
In addition to the GSLC, activity based multimodal corpora of face-to-face interaction, baed on the same principles as the GSLC have also been collected in a number of other countries, which makes interlinguistic and intercultural comparisons of sub-corpora possible

Situation-based Multimodal Analysis - An Example

We will now consider an example of the use of a multimodal activity-based corpus – a study of feedback in the physical examination sub-activity/phase of a typical doctor-patient interaction. This example of how an activity-based multimodal corpus can be used illustrates that even if, as in this case, spoken output was in focus, a multimodal corpus provides information on what goes on in the activity, which is important for determining what is actually said and why. In this case, linguistic feedback in three types of sub-activity in doctor-patient interactions was analyzed.
The results of counting utterances, words, feedback words and the relative share of feedback out of the total speech as well as a classification of the type of feedback, based on the transcription and coding of a specific doctor-patient interaction is shown in table 4. The numbers are given for each of the following three phases: case history (case hist), physical examination (phys ex) and ordination (ordin) totally and separately for the doctor (D) and patient (P) in each of the phases. The *share of feedback* is the share of feedback word tokens out of the total number of word tokens – it indicates how much feedback is used in relation to other words. The *share of utterances containing initial feedback units* (i.e. a feedback word like *yes, no* or *m* at the beginning of an utterance) and the *share of utterances containing only feedback words* indicate the role of feedback and the type of utterances dominating an activity. The *share of totally overlapped feedback units* can tell us if there is a great deal of back-channeling from one participating during long utterances or narratives produced by the other participant. The *share of interrupting feedback* shows if participants interrupt each other frequently, e.g. because the interaction is fast.
Why do we find the numbers related to spoken feedback that appear in the table for the different sub-activities, i.e. what do they reflect in terms of influencing factors and typical patterns of interaction in the three phases of the doctor-patient interaction?

Sub-activities	Number of utter-ances	Number of ords	Total Number of feed-back words	Feed-back share of speech
Case hist	711	5680	530	9.3
Case hist D	373	2315	249	10.8
Case hist P	338	3365	281	8.4
Phys ex	492	3317	358	10.8
Phys ex D	251	2166	176	8.1
Phys ex P	241	1151	182	15.8
Ordin	831	7473	667	8.9
Ordin D	410	5344	268	5.0
Ordin P	421	2129	399	18.7

Sub-activities	Initial FB	Only FB	Inter-rupting FB	Over-lap-ped FB
Case hist	23.6	27.4	2.8	8.0
Case hist D	18.2	32.7	2.4	12.6
Case hist P	29.5	21.6	3.3	3.0
Phys ex	21.1	24.4	3.3	4.9
Phys ex D	20.7	16.7	2.8	4.4
Phys ex P	21.6	32.4	4.4	5.4
Ordin	23.6	30.2	3.5	11.0
Ordin D	22.4	14.6	2.7	6.3
Ordin P	24.7	45,4	4.3	15.7

Table 3. Feedback measures, utterances and words for doctors and patients in three subactvities (phases) of patient-doctor consultation: case history, physical examination, and ordination sub-activities

If we take a look at the physical examination, it can be distinguished by the physical conditions of the examination being different from that of the case history and the ordination and by the focus of action rather than speech. Some feedback characteristics are that the physical examination contains fewer utterances, words and feedback expressions totally, but a higher share of feedback from the patient and fewer overlapped feedback utterances than the other two sub-activities. The lack of overlap reflects a slower and more structured turn management. There is more focus on instruments and body parts, which also leads to less eye contact between the participants.
What do the case history and physical examination have in common? They both have similar purposes, i.e. the doctor collects information, but this is done in different ways, in the case history by listening to the patient' s story with the goal of obtaining information through dialog, and by the doctor's own examination, using observation more than dialog
What do the physical examination and ordination have in common? Both these phases contain considerably

more totally overlapped utterances consisting only of feedback from the patients than from the doctor. This shows that the doctor speaks the most in both these sub-activities. This is so even more in the ordination phase than in the physical examination. (In the case history, on the other hand, the patient speaks the most.)
In order to see what characterizes typical exchanges in the physical examination and how this relates to the quantitative findings, we can look at example 1 below. (English translations of the Swedish utterances are given in italics, ( ) encloses quiet speech, < > encloses comments about what happens, [ ] encloses overlap, /// = long pause).

Example 1.
D: ja ska ta de stående också om du ställer dej där borta
I will take it standing too if you stand over there
P: mm (där nej) <patient gets up> de ä så stelt å resa sej <doctor measures blood pressure>
mm (there no) <patient gets up> if is so stiff to get up <doctor measures blood pressure>
D: men du blir inte yr när du reser dej
but you don't get dizzy when you get up
P. joo ibland
yes sometimes
D: just när du [reser dej ur sängen]
right when you [get out of bed]
P: [joo ja kan inte]
resa mej hastigt [utan]tar de
[yes I can't] can't get up fast [but] take it
D: [nähä] <doctor measures blood pressure>
[no] <doctor measures blood pressure>
D: /// hundrasjutti sjutti
/// a hundred seventy seventy
P: jo då
yes then
D: jaa
yes

In a typical sequence of the physical examination, the doctor has initiative. This means that he does not have to start his utterances with feedback, he can change topic, ask questions and often gives feedback to events. Feedback as reactions to events in the interaction is common in both participants. The sequence can evolve as follows: The doctor asks a yes/no question or another question requiring only a short answer in relation to a specific part of the examination. The patient answers the question and the doctor gives feedback to the answer. However, the examination continues after this feedback and during the silence, the patient quite often makes a short comment.
In example 1, the doctor's first three utterances contain no initial feedback. The doctor questions the patient, while he measures his blood pressure and the patient answers. The last two feedback utterances are reactions to the result of the measurement (which is in this case uttered by the doctor).

Conclusions – The creation and use of multimodal corpora

The collection and use of the GSLC and other related corpora that have been briefly described here have been on-going for more than 30 years. The idea of a variety of social activities in different and mainly naturalistic settings has all the time been in focus and this has made possible a number of observations over the years that have become increasingly relevant for applications related to human-computer interaction, including the design of Embodied Communicative Agents and avatars for use in different types of activities and cultures. The corpus has not originally been collected with these applications in mind, but some of the more recent additions to it have been directly related to this domain. Activities related to different types of service provision, such as information about merchandise, tourism information, travel agency etc. are areas of application, which are represented in the corpus and more of this type of material can be included in the future. The possibility to study how different activity related factors interact is relevant for questions of what can be kept fairly stable and what should be varied in the behavior of interfaces attempting human-human like communication.

References

Ahlsén, E. Allwood, J. & Nivre, J. (2003). Feedback in Different Social Activities. In Juel-Henrichsen, P. (ed.) Nordic Research on Relations between Utterances. Copenhagen Working Papers in LSP, 3, pp. 9-37.
Allwood, J. (1976) Linguistic Communication as Action and Cooperation. Gothenburg Monographs in Lingusitics 2. Univ. of Gothenburg, Dept of Linguistics
Allwood, J. (1999). "The Swedish Spoken Language Corpus at Göteborg University". In Proceedings of Fonetik 99, *Gothenburg Papers in Theoretical Linguistics 81*, Univ. of Göteborg, Dept of Linguistics.
Allwood, J. (2000). An Activity Based Approach to Pragmatics. In Bunt, H., & Black, B. (Eds.) *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics*. Amsterdam, John Benjamins, pp. 47-80.
Allwood, J. (2001). Capturing Differences between Social Activities in Spoken Language. In Kenesei, I., & Harnish, R.M. (eds.) *Perspectives on Semantics, Pragmatics and Discourse*. Amsterdam: John Benjamins, pp. 301-319.
Allwood, J. (2008). Multimodal Corpora. In Lüdeling, Anke & Kytö, Merja (eds.) *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin. 207-225.
Nivre J. (2004)- Gothenburg Transcription Standard (GTS), -V. 6.4. Unversity of Gothenburg, Department of Linguistics.
Nivre, J. (1999). Modified Standard Orthography, Version 6 (MSO6). University of Gothenburg, Department of Linguistics.

A Framework for the Acquisition of Multimodal Human-Robot Interaction Data Sets with a Whole-System Perspective

Johannes Wienke, David Klotz, Sebastian Wrede

Research Institute for Cognition and Robotics (CoR-Lab), Bielefeld University
Universitätsstraße 25, 33615 Bielefeld, Germany
{jwienke,dklotz,swrede}@cor-lab.uni-bielefeld.de

Abstract

In this work we present a conceptual framework for the creation of multimodal data sets which combine human-robot interaction with system-level data from the robot platform. The framework is based on the assumption that perception, interaction modeling and system integration need to be treated jointly in order to improve human-robot interaction capabilities of current robots. To demonstrate the feasibility of the framework, we describe how it has been realized for the recording of a data set with the humanoid robot NAO.

1. Introduction

Improving capabilities of robots for interacting with people is a challenging task which needs to be addressed from various perspectives. Besides models of human-human interaction which guide the development of computational models and realizing software, sufficient perception abilities for people and the scene are required. To cope with the challenges and use opportunities of a fully integrated robot system, these challenges cannot be solved separately. Data sets incorporating such a *whole-system perspective* are required to develop integrated solutions with repeatable and realistic conditions, e.g. for benchmarking (Lohse and others, 2009). Creating such data sets is a complex and time-consuming task. In this work we present a conceptual framework for the creation of these data sets. Moreover, we demonstrate the suitability of the framework by explaining how a real data set involving the humanoid robot NAO was created. This includes the description of chosen technical solutions and lessons learned by the realization. We begin with a description of the scenario for later references and examples.

2. Scenario and Required Data

The scenario is part of the activities in a collaborative research project[1], which tries to improve the abilities of a robot interacting with a group of people through audio-visual integration. Here, the setting of a small vernissage where which visitors are guided by the humanoid robot NAO was chosen. It is inspired by (Pitsch and others, 2011). More detailed, naive participants entered a recording room in pairs and were greeted. Afterwards, the robot presented several paintings in the room using speech and matching gestures. These explanations included pauses intended to elicit comments by the visitors and also gave them the chance to tell the robot if they wanted to hear further explanations at specific points. After the explanations, the robot proceeded with a quiz asking several questions about the paintings and more general topics. During the recordings, speech and movements of the robot were remotely controlled by a human operator. This fact was unknown to the participants.

[1]HUMAVIPS, cf. `http://www.humavips.eu`

In order to address the research questions like addressee detection and visual focus of attention (VFOA), several requirements existed. First, absolute positions and orientation of each participant (specifically the head), the robot and all paintings needed to be known for being able to analyze the VFOA. For this task and for being able to detect addressees, annotations based on the orientation of heads and facial reactions were required. Hence, close video recordings of the faces are required. The same is true for spoken words of all participants. Apart from such external cues, internal sensory and status information of the robot are necessary in order to develop algorithms for a robotic platform that are able to cope with real environment restrictions. For instance, these information include CPU load, kinematics or odometry. Having these information available retains the ability of integrating them into developed algorithms.

3. Challenges in Creating Data Sets

One of the primary issues when recording multimodal data sets is the *synchronization* of all modalities. For cameras or audio streams this could be done in a post-processing phase. However, this is much more complicated with modalities which are less intuitive to observe for humans like robot internal states. Also it induces additional effort required in the post-processing phase. Hence, one challenge is the *reduction of required post-processing* already through the recording setup. This influences the choice of devices.

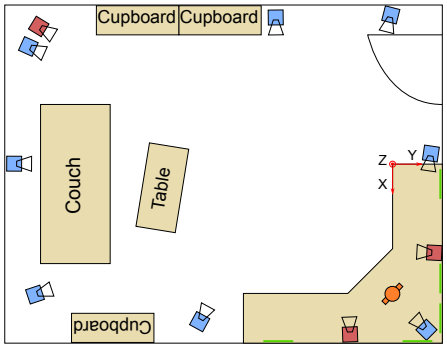

Figure 1: Qualitative overview of the recording room. Orange: NAO, cameras: blue – Vicon, red – HD, green lines: paintings, red: Vicon coordinate system

Moreover, it requires *automation* and *validation* possibilities during the recording time in order to prevent errors in the recorded data. This also concerns the *calibration* of recording devices. By e.g. calibrating all cameras with respect to a motion capturing system unplanned opportunities to use the data set are preserved.

From the whole-system perspective an important requirement is that the recorded data allows a smooth *application in the integrated system*. This means that developed system components can be used without changing their interfaces. While this restricts the recording tools and formats it is still important that the data set can be used without the system integration. Thus *export* facilities to common formats are necessary, e.g. for video.

Finally, for the annotation phase, *established tools* should be reusable and benefits for the annotation should be gained from the system modalities. To efficiently evaluate the system based on the data set the *availability of annotations for integrated components* is essential.

4. The Whole-System Framework

In order to address the aforementioned challenges we propose to directly use the communication system of the robot (i.e. the middleware) as the primary tool and format for data set recordings (data recorded in this way will be called *reference data*). By providing record and replay solutions tightly integrated to the middleware layer, system-internal data can be captured easily and in the native format of the system, as required for whole-system analysis. Moreover, we propose that additional recording devices are either captured directly with this system or their data is later integrated into it. This integration is also the proposed method for secondary data like annotations. Replaying this data enables the *application of the data set in the integrated system* while ensuring the *availability of annotations* without depending on a concrete annotation tool. This means unchanged system component can be used online on the data set with their usual inputs like audio and vision and also have the annotations available in the architecture. The *synchronized replay* in this case is a generic problem and needs to be solved only once, hence reducing parts of the *post-processing effort*. Figure 2 visualizes the proposed process of data management.

According to this process, external recording devices should be selected in a way that they can be recorded with the system infrastructure. This is for instance the case with network cameras, which can be recorded directly using the middleware layer of the system. In cases where this is not possible, the implementation of the framework in Section 5. demonstrates how to *automate* the *synchronization* and conversion of the external data to the reference data.

To address the requirement of *exporting* parts of the data set to common formats we propose a *view-based approach* on the reference data. Views are selective immutable exports of (parts of) the data sets. An exemplary use case of this approach can be the generation of a project for an annotation tool.

5. Applying the Whole-System Approach for a Multimodal HRI Data Set

We will now describe how the framework has been instantiated to record the scenario introduced in Section 2.. Experimental robotics applications in the project are realized on the humanoid robot NAO[2]. Integration is performed using an event-based middleware termed RSB (Wienke and Wrede, 2011) which allows full introspection of the internals of the robotics system. All information in the system is continuously sent over a logically unified bus, which can be composed of different transport layers (e.g., network or in-process communication) within events as the basic unit of communication. Each event contains data like sensor reading and a set of meta data including accurate timing information. Based on the introspection support, RSB includes a mechanism (RSBag) to record and replay the events stream with original timing and hence realizes the acquisition of the reference data. All internal data from NAO as well as the control commands for remote operation have been recorded using this mechanisms without needing modifications of the system. As RSB and the record and replay mechanism are generic, this architecture can be reused for other data set acquisition activities.

Besides this system-level data we utilized a Vicon motion capturing system[3] to acquire ground truth position data of participants and the robot, installed 3 HD cameras in the recording room to provide views for the annotation of addressees and VFOA, and equipped each participant with a close talk wireless microphone. Due to restricted interfaces, the Vicon system and the HD cameras could not be recorded based on the RSB system, so they form a test case for the later integration into the reference data during the post-processing. In contrast, the close talk microphones were attached to one of the recording computers and hence could be recorded inside the RSB architecture. As the internal clocks of all computers in the distributed system were synchronized using NTP, all data recorded using RSB is synchronized without manual work. An overview of the recording room and the placement of recording devices is depicted in Figure 1. Table 1 lists all recorded data streams.

Besides the actual recording of the scenario with different participants, *calibration* sequences have been recorded. From these runs a special Vicon marker has to be extracted at certain points in time to find out the locations of paintings and several other objects in the scene. Even though it would be possible to extract these positions manually, we increased the *automation* in the post-processing phase by presenting a clapperboard which was markered for the Vicon system each time the marker for measuring positions was finally placed at the desired position.

Besides this calibration aspect, we recorded a checkerboard pattern for all cameras (including NAO) so that distortions can be calibrated. Moreover, a special Vicon subject with 4 tracked markers has been presented to the HD cameras and the Vicon at once. Hence, the location of each HD camera in the Vicon coordinate system can be computed.

[2] http://aldebaran-robotics.com

[3] http://www.vicon.com

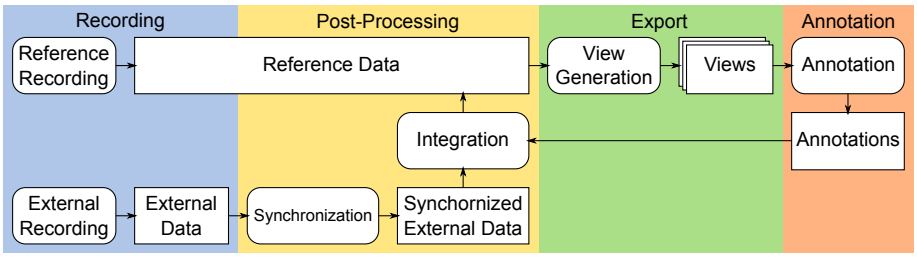

Figure 2: A schematic overview of the proposed framework. Rounded boxes indicate activities to be performed, other squares represent generated data. Different workflow phases while creating the data set are indicated through different background colors

5.1. Post-Processing

As previously mentioned, the Vicon system and the HD cameras could not be recorded directly using RSB. For Vicon, in principle, an online API exists so that special software can be written which receives the Vicon measurements and sends them using RSB. However, using this approach no tracking errors can be corrected afterwards using the Vicon Nexus software. Moreover, the online API up to our knowledge lacks accurate timing information. For these reasons we decided to use the internal recording capabilities of Vicon which allows manual error correction afterwards but also requires a manual processing, export and synchronization with the remaining recordings. The HD cameras were used as no cameras with Ethernet connection were available for the recording and the high resolution is a requirement for the annotation.

To synchronize the videos from the HD cameras we calculate the cross-correlation peak of the cameras' audio channels with a reference audio channel recorded in RSB where exact timing information are available. For this purpose we used the sound recorded by NAO's microphones as a good cross-correlation can be expected because NAO and the external cameras were always in the same room. This was not the case for the close talk microphones which were carried around by participants and could be muted.

Type	Specification
NAO video	Monocular uncompressed frames, VGA, variable frame rate (∼15 fps mean), YUV422 color mode.
NAO audio	4 channels, 48000 Hz, 16 bit signed.
NAO odometry	est. 2D location of robot body
NAO proprioception	Joint angles, stiffness, last command value, temperature
NAO system	CPU, memory, battery, modules
Demo system and control	Wizard commands, internal events for speech and gesture production
close talk microphones	4 channels, 44100 Hz, 24 bit signed
Vicon	6D pose for people and NAO, 100 Hz
External HD Cameras	3 perspectives, 1920 × 1080 pixels, 25 Hz. 5.1 channel sound, 48000 Hz.

Table 1: Detailed description of recorded data. Italic entries have been recorded using RSB.

Praat (Boersma and Weenink, 2011) was used to realize the cross-correlation calculation. Based on the correlation peak we deduced the offset of the external videos with respect to the audio from NAO, which in turn allowed us to compute the start time of the external videos in the RSB time frame.

To generate synchronized results from the Vicon system several steps were necessary. First, each recorded trial needs to be processed in Vicon which might involve manual labeling in situations where the Vicon system could not track its artificial markers sufficiently well. After this processing, an export of the tracking results was performed using the easy to parse CSV (comma-separated values) export format. Vicon uses a fixed frame rate, in our case 100 Hz. Hence, it is sufficient to know the RSBag time of one Vicon frame per trial. For this purpose we implemented a clapperboard detection which extracts the Vicon frame for the moment the clapperboard was shut. The detection is based on a sliding window approach on the Euclidean distance of the two clapperboard parts. Unfortunately, we did not find an automatic way to relate this Vicon frame to RSBag, as no easy detection of the clapperboard in the recorded modalities using this system was possible. Hence, we manually searched for the time of the clapperboard in an export of the audio recordings using the open source audio editor Audacity[4] which allows to display the exact sample count for the cursor. This provides a high precision and simplifies the calculation of the RSBag timestamp. The clapperboard detection was also used to automatically obtain the positions of paintings and other interesting locations from the calibration runs.

Both, Vicon exports to CSV and HD video can be integrated into the RSBag files by creating RSB events containing the data with the correct timestamps provided by the aforementioned synchronization procedures.

5.2. Annotation and View-based Access

We decided to use the well-established annotation tool ELAN (Wittenburg and others, 2006) for the data set. As ELAN is not capable of processing the RSBag format of the data set, we needed to provide an ELAN-view on it. To *automate* the creation of this task, we have developed a script which creates a view on the data set to enable the annotation in ELAN. It uses the synchronized data created during the post-processing, converts video and audio to file

[4] `http://audacity.sourceforge.net/`

formats compatible with ELAN using ffmpeg[5], and automatically creates a project file to load in ELAN. The annotations created in ELAN will be integrated into the RSBag files to ensure the *availability of annotations for integrated components*.

6. Related Data Set Acquisition Approaches

To the best of our knowledge, no work exists which provides a conceptual framework on how to capture data sets which contain HRI data in a whole-system manner, especially with the aspect of direct replay possibilities to the system architecture.

A closely related scenario with comparable modalities has been presented in (Green and others, 2006). However, it only covers a single sensor from the system and neglects the remaining communication in the system, especially the commands for remote control of the robot. Also, no generic approach for recording has been presented.

Regarding data sets without robotics involved, (Luz et al., 2006) describes the acquisition of computer-mediated meetings through a distributed system. Their system streams data over the network using RTP and also contains internal data of the collaborative editor. However, there is no unique recording format. (Roggen and others, 2010) describes the implementation of an acquisition system for corpora based on a largely distributed sensor system. Their recording system provides no intrinsic mechanism for synchronized recording. Instead manual inspection is proposed for this task. A recording architecture for meeting corpora is presented in (Banerjee and others, 2004) with comparable aims for extensibility as in our approach. The presented system uses NTP for time synchronization and has a comparable approach of timed events for stored data. Besides the recording aspects, an approach to collaboratively aggregate more data in data sets with a central server where the data is uploaded. The collaboration idea is further devised in (Chervenak and others, 2000). Our current approach does not cover such a level of dissemination and collaboration, even though highly required to facilitate research.

7. Conclusion and Outlook

The applicability of our devised generic framework for data set acquisition has been demonstrated to a large extend on a concrete implementation for the vernissage scenario. However, further validation of the concept with respect to the annotation as incremental addition of other data needs to be performed. Ultimately, the integration of our concept with the ideas presented in (Chervenak and others, 2000) is required to cover the whole workflow from recording to dissemination.

For the described implementation several optimizations are possible. The current integration of the Vicon system needs to be improved to provide further automation and reduce the post-processing effort. External cameras could be replaced with networked cameras. However, the current integration solutions are not tailored to the specific scenario and can be reused for other recordings. In the future we will continue to evaluate the application of further views, for the export of the data set. A prototype we have developed indicates that with a web-based interface potential users of the data set can easily browse the contents and request an appropriate synchronized view from the web-server without needing to install specialized tools.

Summing up, the framework provides a structured approach for the acquisition of data sets which include system level information. As the situational context is partially determined by the system, the frameworks helps in generating a better view on these context aspects in data sets.

[5] `http://ffmpeg.org`

8. Acknowledgments

This work was funded by the European FP7 project HUMAVIPS, theme ICT-2009.2.1, grant no. 247525. Thanks to Raphaela Gehle, Michael Götting, Dinesh Jayagopi, Stefan Krüger, Phillip Lücking, Jan Moringen, Karola Pitsch, Lars Schillingmann, Jens-Christian Seele, Samira Sheikhi, and all participants for their help with the data set.

9. References

Satanjeev Banerjee et al. 2004. Creating multi-modal, user-centric records of meetings with the carnegie mellon meeting recorder architecture. In *Proc. of ICASSP'04, the Int. Conf. on Acoustics, Speech, and Signal Processing, Meeting Recognition Workshop*, Montreal, Canada.

Paul Boersma and David Weenink. 2011. Praat: doing phonetics by computer (version 5.3.03). `http://www.praat.org`. Computer program.

Ann Chervenak et al. 2000. The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets. *Journal of Network and Computer Applications*, 23(3).

Anders Green et al. 2006. Developing a contextualized multimodal corpus for human-robot interaction. In *Proc. of the 5th Int. Conf. on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

Manja Lohse et al. 2009. Systemic Interaction Analysis (SInA) in HRI. In *Proc. Int. Conf. Human-Robot Interaction*.

Saturnino Luz, Matt-Mouley Bouamrane, and Masood Masoodian. 2006. Gathering a corpus of multimodal computer-mediated meetings with focus on text and audio interaction. In *Proc. 5th Int. Conf. on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

Karola Pitsch et al. 2011. Attitude of german museum visitors towards an interactive art guide robot. In *Proc. of the 6th int. conf. on Human-robot interaction*, HRI '11, pages 227–228, New York, NY, USA. ACM.

Daniel Roggen et al. 2010. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh Int. Conf. on Networked Sensing Systems (INSS)*, number 00. IEEE.

Johannes Wienke and Sebastian Wrede. 2011. A middleware for collaborative research in experimental robotics. In *2011 IEEE/SICE Int. Symposium on System Integration, SII2011*, Kyoto, Japan. IEEE, IEEE.

P. Wittenburg et al. 2006. Elan: a professional framework for multimodality research. In *Proc. of LREC 2006, Fifth Int. Conf. on Language Resources and Evaluation*.