Gefördert vom

bmb+f

Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie

Mensch-Technik-Interaktion

SmartKom

Interlabeller Agreement in SmartKom Multi-Modal Annotations

Irene Jacobi

Florian Schiel

Ludwig-Maximilians-Universität München

Technisches Dokument Nr. 26

Dez 2003

Dez 2003

Irene Jacobi
Florian Schiel

Ludwig-Maximilans-Universität München
Schellingstr. 3
80799 München

Tel.: (089) 2180-2758
FAX: (089) 2800362

E-Mail: schiel@phonetik.uni-muenchen.de

Dieses Technische Dokument gehört zu Teilprojekt 1:
Emprische Datensammlung

Das diesem Technischen Dokument zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01 IL 905 gefördert. Die Verantwortung für den Inhalt liegt beim Autor.

Contents

1 Introduction

Within the SmartKom project different groups of labellers categorized and segmented (henceforth summarized as 'labelling') utterances, facial expressions and gestures in recorded man-machine-dialogues using fixed sets of label categories. To work out the interlabeller agreement and therewith the reliability of a labelling, a certain amount of dialogues was labelled twice by different, non-overlapping groups of labellers. The accuracy and correctness values were calculated as well as the single label agreements. The results might give reason to either keep or neglect certain categories and subcategories when labelling interactive man-machine-dialogues.

1.1 Files and Labels

The dialogues for the interlabeller agreement probe were chosen by chance. Four different types of labellings were investigated: holistic (USH), gesture (GES), mimic (USM), and prosody (USP).
The **USH-files** contain facial expression labels that were assigned by the labellers whilst watching and listening to a video of the dialogue situation (cf. [1],[2]).
The labels of the **GES-files** were annotated within the same circumstances, marking any gesture activity in the facial area or in the cube i area[1] (cf. [3],[5]).
The **USM-files** contain mimic labels which were assigned by the labellers whilst merely watching the dialogue video, without listening to the audio (cf. [6]). Therefore, the USM-labeller got a filtered USH-file version to start with, in which all USH user-state labels were deleted, barring the 'Neutral' ones. The USM-labellers were instructed to ignore the neutral segments and to bring in their own labels and new ratings only within the segments previously marked as user states by the USH labellers. Consequently, USH-files and USM-files show a certain dependency. Still, USM-labellers had - and often used - the possibility to segment the neutral USH-segments anew.
The **USP-files** contain prosodic labels, annotated into a (already provided) transcription file while listening to the audio signal only (cf. [4],[6]).
Further information on categories und subcategories can be retrieved from [4],[5],[6], as well as information on which categories were based on former investigations or publications and which were self-designed.

1.2 Labelers

The first labelling was done in the years 2001 and 2002. It took place during the normal processing of the data of the Wizard-of-Oz recordings. The second labelling was done especially for the calculation of the interlabeller agreement. It took place in 2003.
Each session in the second labelling was done by a different labeller than in the first labelling. The time between the first and the second labelling comprised several months. The first labelling consistes of three steps: first labelling, correction and end-correction (USH, GES and USP) with the exception of USM that consisted only of two steps: first labelling and end-correction.
The second labelling consisted only of two steps: first labelling and correction. This was done

[1]The area above the display, where 2D hand gestures may be captured by the SmartKom system.

because for each group (USH, USM, USP, GES) there existed only one person for the end-correction to maximize the consistency. The labellers in each group were trained by one person (in most cases the same that did the end-correction and/or developed the labelling system). All had a good labelling experience, with exception of some of the labellers in the second labelling that had only medium labelling experience. The labellers were students with different study subjects (only some were Phoneticians). None had had labelling experience before being trained in SmartKom. All except two were native Germans (one was Spanish, one Russian, both spoke fluently German).
During the first labelling labellers sometimes discussed difficult cases. During the second labelling this was not allowed.

	first labelling	**second labelling**
Time	2001-2002	2003
number of steps (USH,USP,GES)	3	2
number of steps (USM)	2	2

2 Analysis

Within each category the labels of 20 files (labelling A) had to be compared to the labels of their 20 equivalent files (labelling B). Corresponding labels had to be identified and compared to each other. Therefore labels from the two files were matched in consideration of their time range, allowing insertions (i) and deletions (d), as well as two-fold labelling.

```
Schematical label occurences:


        file x        - - - - - - - - time - - - - - - - - - - - - ->

       labelling A:    [ label ][ label ]          [   longlabel    ]
       labelling B:    [ label ]          [ label ][ label ][ label ]
                         |         |         |             |
                         |         |         |             |
       matched?        match    delet.   insert.   two-fold-match
```

A perl programm was written to read in a set of label files and compare them one by one with the corresponding transcription files. The basic rule searches for majority overlaps between the segments of labelling A and B or decide if there is an insertion or deletion.

Compare fileA vs fileB - short version of matching rule

if $\lvert LabelstartA_i - LabelstartB_i\rvert < \lvert LabelstartA_i - LabelstartB_{i+1}\rvert$ and $\lvert LabelstartA_i - LabelstartB_i\rvert < \lvert LabelstartB_i - LabelstartA_{i+1}\rvert$ then match $LabelA_i$ to $LabelB_i$ else if $LabelstartA_i < LabelstartB_i$ then mark A_i as a deletion since there's no matching B_i, else if $LabelstartA_i > LabelstartB_i$ then mark an insertion before A_i since there's no match to B_i

Since there was no clear reference file during analysis, each of the two corresponding files was one time classified as the reference file and one time classified as the corresponding file, resulting in two correctness values and two accuracy values for each pair of files. Following this, overall accuracy (EndAcc) resp. overall correctness (EndCorr) for one pair of files was then the sum of both accuracy values resp. correctness values, devided by two.

```
matching example:
                              left tabular to right tabular -> change of reference file

lines:
matchA  startA  startB  labelA -- corresp labelB             matchB  startB  startA  labelB -- corresp labelA

s 0 0 Überlegen/Nachdenken -- Neutral                        s 0 0 Neutral -- Überlegen/Nachdenken
d 1920 0 Neutral -- s.o.                                     i 0 1920 s.o. -- Neutral
r 9600 10240 Restklasse -- Restklasse                        r 10240 9600 Restklasse -- Restklasse
r 16640 17920 Neutral -- Neutral                             r 17920 16640 Neutral -- Neutral
r 161760 163040 Restklasse -- Restklasse                     r 163040 161760 Restklasse -- Restklasse
r 197120 193920 Neutral -- Neutral                           r 193920 197120 Neutral -- Neutral
r 220800 203520 Überlegen/Nachdenken -- Überlegen/Nachdenken r 203520 220800 Überlegen/Nachdenken -- Überlegen/Nachdenken
s 241920 243840 Ratlosigkeit -- Neutral                      s 243840 241920 Neutral -- Ratlosigkeit
i 241920 252160 s.o. -- Restklasse                           d 252160 241920 Restklasse -- s.o.
i 241920 263040 s.o. -- Neutral                              d 263040 241920 Neutral -- s.o.
r 273280 278400 Freude/Erfolg -- Freude/Erfolg               r 278400 273280 Freude/Erfolg -- Freude/Erfolg
s 295040 291200 Neutral -- Überlegen/Nachdenken              s 291200 295040 Überlegen/Nachdenken -- Neutral
d 331520 291200 Überlegen/Nachdenken -- s.o.                 i 291200 331520 s.o. -- Überlegen/Nachdenken
d 362240 291200 Überlegen/Nachdenken -- s.o.                 i 291200 362240 s.o. -- Überlegen/Nachdenken
r 437760 450560 Neutral -- Neutral                           r 450560 437760 Neutral -- Neutral
i 437760 465280 s.o. -- Überlegen/Nachdenken                 d 465280 437760 Überlegen/Nachdenken -- s.o.
i 437760 711040 s.o. -- Neutral                              d 711040 437760 Neutral -- s.o.
i 437760 734080 s.o. -- Überlegen/Nachdenken                 d 734080 437760 Überlegen/Nachdenken -- s.o.
i 437760 770560 s.o. -- Neutral                              d 711040 437760 Neutral -- s.o.
r 773120 773760 Überlegen/Nachdenken -- Überlegen/Nachdenken r 773760 773120 Überlegen/Nachdenken -- Überlegen/Nachdenken
r 790400 787840 Neutral -- Neutral                           r 787840 790400 Neutral -- Neutral
```

Figure 1: Example for a match of two files, r = correct match, s = substitution, i = insertion, d = deletion (s.o. = still the same label as above)

Percentage of correctness is based on the number of correct labels devided by the total number of labels, multiplied by hundred:

$$corr = \frac{N_{ref} - d - s}{N_{ref}} 100\% \tag{1}$$

Percentage of accuracy is based on the total number of labels minus substitutions, minus deletions, minus insertions devided by the total number of labels and multiplied by hundred:

$$acc = \frac{N_{ref} - d - s - i}{N_{ref}} 100\% \tag{2}$$

3 Results

The matrices show the interlabeller analysis results for 2x20 USH-files, 2x20 USM-files, 2x20 USP-files and 2x20 GES-files. For each matrix, the labels of the first labelling are written above horizontally, the second labelling's labels are written on the left vertically.
'None' means either that no label was assigned at all, meaning that there's a pause which can only occur within GES-files. The meaning of 'none' can also be that no new label was assigned, with the precedent label expanding over the relating time section (thus causing a deletion or insertion in the final count).

```
examples for 'none' in the B labelling:

                    - - - - - - -time- - - - - ->

              A:   [ label ]         [label  ][ label ][    longlabel    ]
              B:            [ label ][   longlabel    ][ label ]
                       |        |        |        |         |        |
                       |        |        |        |         |        |
       labelling B    none     label    label    none      label    none
```

3.1 USH

An agreement between USH labelling is with an accuracy of 1.95% almost non existent, ascribed to twice as much insertions/deletions and substitutions as concordant labels. The lot of insertions/deletions is attributed to the fact that one label group set 763 labels in twenty dialogues whereas the other group set only 494 labels to categorize user states within the same dialogues. The first labelling (not rep.-files) with the smaller amount of assigned labels shows less misses than the second labelling and therewith a higher reliability.
When matching segmentally, more than a fourth of the segments are judged differently by the two labellings.
Even the categories 'Freude' (joy) and 'Überlegen' (thinking) with the largest amount of concordant user state labels (barring 'Neutral') are not very authentic when related to the amount of their misses.

Within the USH-files there is a small amount of labels attributed with 'strong', describing a certain perceived strength of the labeled mimic. Matched segment pairs containing at least one label with the attribute 'strong' show an accuracy of 36.60%. This can be led back on one hand to the more alike amount of labels on both sides, 37 vs 46 assigned labels (compared to 494 vs 763 overall-USH-labels) and on the other hand to the higher amount of 'right' label matches related to substitutions/insertions/deletions.
Again 'Freude' and 'Überlegen' show the highest amount of concordance but only little reliability as the labellers' judgements diverse in at least half of the labels' occurances.

3.1.1 USH matrix

```
-----------------------------------------------------------------------------------------
Ref.-files/A-files =  20 original USH files       B-files = 20 rep. USH 20 files
Labels in A-Files = 494                           Labels in B-Files = 763

                                Label combinations USH
         r = 273
         s = 123
         d =  98
         i = 367                                                Correct =  55.26 %
Ref-labels = 494                                               Accuracy = -19.03 %
```

B \ Ref	Freude	Überlegen	Ärger	Überrasch	Ratlos	Rest	Neutral	none
Freude	47	1	2	0	2	2	4	43
Überlegen	1	41	2	1	1	8	5	63
Ärger	0	4	4	1	7	1	3	8
Überrasch	0	0	1	1	3	2	0	5
Ratlos	0	1	0	2	2	0	1	2
Rest	4	5	5	0	3	23	6	71
Neutral	6	12	6	2	3	16	155	175
none	3	16	9	5	1	17	47	

```
                               CHANGE OF REFERENCE FILE


B-files = 20 original USH files         Ref.-files/A-files = 20 rep. USH files
Labels in B-Files = 494                 Labels in A-Files = 763

                                Label combinations USH
         r = 273
         s = 123
         d = 367
         i =  98                                                 Correct = 35.78 %
Ref-labels = 763                                                Accuracy = 22.94 %
-----------------------------------------------------------------------------------------

                                # # # # # # # # # # #
                                #        USH        #
                                #  EndCorr  45.52%  #
                                #   EndAcc   1.95%  #
                                # # # # # # # # # # #
```

3.1.2 USH 'strong' matrix

Ref.-files/A-files = 20 original USH files B-files = 20 rep. USH 20 files
Labels in A-Files = 494 Labels in B-Files = 763

Label combinations USH with
at least one 'strong' label in A or B

r = 21
s = 15
d = 1
i = 10
Ref-Labels = 37

Correct = 56.76 %
Accuracy = 29.73 %

B \ Ref	Freude	Überlegen	Ärger	Überrasch	Ratlos	Rest	Neutral	none
Freude	7	0	0	0	1	0	0	3
Überlegen	0	12	1	1	0	0	2	6
Ärger	0	0	2	0	3	0	0	0
Überrasch	0	0	0	0	3	0	0	0
Ratlos	0	1	0	0	0	0	1	1
Rest	1	1	0	0	0	0	0	0
Neutral	0	0	0	0	0	0	0	0
none	0	1	0	0	0	0	0	

CHANGE OF REFERENCE FILE

B-files = 20 original USH files Ref.-files/A-files = 20 rep. USH files
Label insg. in B-Files = 494 Labels in A-Files = 763

Labelcombinations USH with
at least one 'strong' label

r = 21
s = 15
d = 10
i = 1
Ref-Labels = 46

Correct = 45.65 %
Accuracy = 43.48 %

```
# # # # # # # # # # #
#   USH 'strong'    #
#  EndCorr  51.20%  #
#   EndAcc  36.60%  #
# # # # # # # # # # #
```

3.2 USM

The agreement between USM labellers leads to an accuracy of 45.42%. This calculation does not include the concordant labelling of 'Neutral' segments as these are in large parts predetermined by the USH-labellers. Still it should be mentioned that the USH-groups labelled 221 resp. 375 times 'Neutral' whereas the USM-groups labeled 221 resp. 233 times 'Neutral'. Barring 'Freude', the variance is overall fairly high.

The segment matches including at least one 'strong' label show an accuracy of 64.87% and fairly reliable hit results regarding the labels 'Freude' and 'Überlegen'. The labelling of 'Ratlos' shows the highest spreading and therewith the worst reliability.

3.2.1 USM matrix

```
-----------------------------------------------------------------------------------------
Ref.-files/A-files = 20 original USM files          B-files = 20 rep. USM files
Labels in A-Files = 474                             Labels in B-Files = 493

                                 Label combinations USM
         r = 155 (barring 'Neutral'-concordance)
         s =  85
         d =  22
         i =  41                                                Correct = 59.16 %
Ref-Labels = 262                                               Accuracy = 43.51 %
```

B \ Ref	Freude	Überlegen	Ärger	Überrasch	Ratlos	Rest	Neutral	none
Freude	47	1	2	0	1	1	1	4
Überlegen	3	61	3	0	11	9	1	14
Ärger	0	2	10	0	3	1	0	1
Überrasch	0	0	0	4	7	0	0	1
Ratlos	3	3	10	4	17	4	1	1
Rest	1	6	0	0	4	16	0	2
Neutral	0	1	0	0	0	2	(212)	18
none	1	5	0	1	4	3	8	

```
                               CHANGE OF REFERENCE FILE


Ref.-files/A-files = 20 rep. USM files            B-files = 20 original USM files
Labels in A-Files = 493                           Labels in B-Files = 474

                               Label combinations USM
         r = 155 (barring 'Neutral'-concordance)
         s =  85
         d =  41
         i =  22                                                Correct = 55.16 %
Ref-Labels = 281                                               Accuracy = 47.33 %
-----------------------------------------------------------------------------------------

                                # # # # # # # # # # #
                                #        USM        #
                                #  EndCorr  57.16%  #
                                #   EndAcc  45.42%  #
                                # # # # # # # # # # #
```

3.2.2 USM 'strong' matrix

```
------------------------------------------------------------------------------------------
Ref.-files/A-files = 20 original USM files          B-files = 20 rep. USM files
Labels in A-Files = 474                             Labels in B-Files = 493

                               Label combinations USM with
         r = 64                at least one 'strong' label in A or B
         s = 27
         d =  1
         i =  5                                                Correct = 69.57 %
Ref-Labels = 92                                               Accuracy = 64.13 %
```

B \ Ref	Freude	Überlegen	Ärger	Überrasch	Ratlos	Rest	Neutral	none
Freude	14	0	0	0	0	0	0	1
Überlegen	0	36	0	0	5	0	0	3
Ärger	0	1	7	0	3	0	0	1
Überrasch	0	0	0	1	3	0	0	0
Ratlos	2	2	8	0	6	1	1	0
Rest	0	1	0	0	0	0	0	0
Neutral	0	0	0	0	0	0	0	0
none	0	0	0	0	1	0	0	

```
                                 CHANGE OF REFERENCE FILE


Ref.-files/A-files = 20 rep. USM files          B-files = 20 original USM files
Labels in A-Files = 493                         Labels in B-Files = 474

                               Label combinations USM with
         r = 64                at least one 'strong' label in A or B
         s = 27
         d =  5                                                 Correct = 66.67 %
         i =  1                                                Accuracy = 65.62 %
Ref-Labels = 96
------------------------------------------------------------------------------------------

                                # # # # # # # # # # #
                                #    USM 'strong'   #
                                #  EndCorr  68.12%  #
                                #   EndAcc  64.87%  #
                                # # # # # # # # # # #
```

3.3 USP

The main reason for the negative accuracy within the USP-files is the huge overall amount of insertions/deletions. The label 'Length-Syllable' shows the best - still poor - reliability with 45 matched pairs and 15 resp. 52 other (mis-paired) single occurances.

3.3.1 USP matrix

```
------------------------------------------------------------------------------------------
Ref.-files/A-files = 20 original USP files          B-files = 20 rep. USP files
Labels in A-Files = 373                             Labels in B-Files = 299

                                Label combinations USP
        r = 129
        s =  17
        d = 206
        i = 133                                                         Correct = 36.65 %
Ref-Label = 352                                                        Accuracy = -1.14 %
```

B \ Ref	CLEAR_ART	EMPHASIS	LAUGHTER	LENGTH_SYLL	PAUSE_PHRAS	PAUSE_WORD	STRONG_EMPH	none
CLEAR_ART	16	0	0	0	0	0	0	17
EMPHASIS	13	52	0	2	0	0	4	56
LAUGHTER	0	0	3	0	0	0	0	3
LENGTH_SYLL	6	1	0	45	0	0	0	37
PAUSE_PHRAS	0	0	0	0	5	4	0	7
PAUSE_WORD	0	0	0	0	6	8	0	13
STRONG_EMPH	0	0	0	0	0	0	0	0
none	81	95	1	13	14	2	0	

```
                                 CHANGE OF REFERENCE FILE

Ref.-files/A-files = 20 rep. USP files          B-files = 20 original USP files
Labels in A-Files = 299                         Labels in B-Files = 373

                                 Labelcombinations USP
        r = 129
        s =  17
        d = 133
        i = 206                                                     Correct = 46.24 %
Ref-Label = 279                                                    Accuracy = -27.60 %

EndCorr damit:  41.44%
 EndAcc damit: -14.37%
------------------------------------------------------------------------------------------
                                  # # # # # # # # # # #
                                  #       USP         #
                                  #  EndCorr  41.44%  #
                                  #  EndAcc  -14.37%  #
                                  # # # # # # # # # # #
```

3.4 GES

Within an overall accuracy of 52.32%, the most reliable label is 'I-Geste'. The first labeller group (not-rep.-files) had less misses, setting less labels than the second labeller group.

GES4 labels are further specified with the labels of subcategory GES5. If label pairs are concordant in category GES4, category GES5 labelling will be concordant in 90% of the cases as well. Accuracy for category GES4 including concordance in GES5 is therewith only little worse with 45.14%.

3.4.1 GES 4 matrix

```
------------------------------------------------------------------------------------------
Ref.-files/A-files = 20 original GES files           B-files = 20 rep. GES files
Labels in A-Files = 171                              Labels in B-Files = 227

                              Label combinations gesture category 4
         r = 142
         s =  21
         d =   8
         i =  64                                                     Correct = 83.04 %
Ref-Labels = 171                                                    Accuracy = 45.61 %
```

B \ Ref	I-Geste	R-Geste	U-Geste	none
I-Geste	78	3	3	19
R-Geste	1	50	0	41
U-Geste	3	11	14	4
none	3	3	2	

```
                                   CHANGE OF REFERENCE FILE


Ref.-files/A-files = 20 rep. GES files        B-files = 20 original GES files
Labels in A-Files = 227                       Labels in B-Files = 171

                         Label combinations gesture category 4
         r = 142
         s =  21
         d =  64
         i =   8                                                     Correct = 62.56 %
Ref-Labels = 227                                                    Accuracy = 59.03 %
------------------------------------------------------------------------------------------

                                  # # # # # # # # # # #
                                  #       GES4        #
                                  #  EndCorr  72.80%  #
                                  #   EndAcc  52.32%  #
                                  # # # # # # # # # # #
```

3.4.2 GES5, GES 4+5 matrix

```
-------------------------------------------------------------------------------------------

Label gesture subcategory 5
---------------------------
r = 128
s =  14
d =   0
i =   0
agreeing labels from preceding (rougher) category = 142

              I-Geste agreement = 78
thereof agreeing in subcategory = 69

              R-Geste agreement = 50
thereof agreeing in subcategory = 47

              U-Geste agreement = 14
thereof agreeing in subcategory = 12


Rough labels with subcategoric label matches:
```

	I-deut	I-frei	I-kreis	I-tipp	R-emot	R-UFO	U-les-k	U-les-p	U-such-k	U-überl-k	U-überl-p	U-zähl-k	n erkbar
I-Geste	7	1	4	57	0	0	0	0	0	0	0	0	0
R-Geste	0	0	0	0	11	36	0	0	0	0	0	0	0
U-Geste	0	0	0	0	0	0	0	0	6	6	0	0	0

```
# # # # # # # # # # #
#       GES5        #
#  EndCorr  90.14%  #
# # # # # # # # # # #


# # # # # # # # # # #
#      GES4+5       #
#  EndCorr  65.61%  #
#   EndAcc  45.14%  #
# # # # # # # # # # #
```

References

[1] S. Steininger, F. Schiel, A. Glesner (2002) Labeling Procedures for the Multi-modal Data Collection of SmartKom. In: Proceedings of the 3rd Language Resources & Evaluation Conference (LREC) 2002, Las Palmas, Gran Canaria, Spain, pp. , editors: Manuel Gonzalez Rodriguez and Carmen Paz Suarez Araujo.

[2] S. Steininger, F. Schiel, O. Dioubina, S. Rabold (2002) Development of User-State Conventions for the Multimodal Corpus in SmartKom. In: Proceedings of the Workshop 'Multimodal Resources and Multimodal Systems Evaluation' 2002, Las Palmas, Gran Canaria, Spain, pp. 33-37.

[3] S. Steininger, B. Lindemann & Th. Paetzold (2001) Labeling of Gestures in SmartKom – The Coding System. In: I. Wachsmuth & T. Sowa (Eds.): Gesture and Sign Languages in Human-Computer Interaction, International Gesture Workshop 2001, London, UK. Berlin: Springer, pp. 215-227.

[4] N. Beringer, V. Penide-Lopez, S. Neubauer, D. Oppermann, S. Biersack (2000) Prosodie-Labeling. SmartKom TechDok-NR-11.

[5] S. Steininger, B. Lindemann, Th. Paetzold (2001) Labeling von Gesten im Mensch-Maschine-Dialog – Gesten-Kodierkonvention SmartKom. SmartKom TechDok-NR-14.

[6] S. Steininger. O. Dioubina, R. Siepmann, C. Beiras-Cunqueiro, A. Glesner (2001) Labeling von User-States im Mensch-Maschine Dialog – User-State Kodierkonventionen SmartKom. SmartKom TechDok-NR-17.