| | |
|---|---|
| **Project ref. no.** | IST-1999-10647 |
| **Project title** | ISLE Natural Interactivity and Multimodality Working Group |

| | |
|---|---|
| **Deliverable status** | Public |
| **Contractual date of delivery** | 30 September 2001 |
| **Actual date of delivery** | February 2002 |
| **Deliverable number** | D8.1 |
| **Deliverable title** | Survey of NIMM Data Resources, Current and Future User Profiles, Markets and User Needs for NIMM Resources |
| **Type** | Report |
| **Status & version** | Final |
| **Number of pages** | 366 |
| **WP contributing to the deliverable** | WP8 |
| **WP / Task responsible** | Niels Ole Bernsen, NISLab |
| **Editors** | Malene Wegener Knudsen, Jean-Claude Martin and Laila Dybkjær |
| **Authors** | Malene Wegener Knudsen, Jean-Claude Martin, Laila Dybkjær, Stephen Berman, Niels Ole Bernsen, Khalid Choukri, Ulrich Heid, Sotaro Kita, Valerie Mapelli, Catherine Pelachaud, Isabella Poggi, Gijs van Elswijk, Peter Wittenburg |
| **EC Project Officer** | Brian Macklin |
| **Keywords** | Natural Interactivity, Multimodality, Data resources, facial expression, gesture, speech |
| **Abstract (for dissemination)** | This ISLE Deliverable 8.1 from the ISLE Natural Interactivity and Multimodality (NIMM) Working Group presents a survey of NIMM data resources and a strategic description of current and future user profiles, markets and user needs for NIMM data resources.<br><br>The report reviews 36 facial data resources and 28 gesture data resources some of them including speech. In addition, 28 answers to a questionnaire distributed at the November 2001 Dagstuhl workshop are presented in an Appendix.<br><br>The descriptions of reviewed data resources use a common structure. The purpose of this structure is to facilitate comparison across data resources by providing similar information about each resource to the extent possible. |


# ISLE Natural Interactivity and Multimodality Working Group Deliverable D8.1

## Survey of NIMM Data Resources, Current and Future User Profiles, Markets and User Needs for NIMM Resources

**February 2002**

## Authors

Malene Wegener Knudsen[1], Jean-Claude Martin[5], Laila Dybkjær[1], Stephen Berman[4], Niels Ole Bernsen[1], Khalid Choukri[3], Ulrich Heid[4], Sotaro Kita[6], Valerie Mapelli[3], Catherine Pelachaud[2], Isabella Poggi[2], Gijs van Elswijk[6], Peter Wittenburg[6]

1: NISLab, University of Southern Denmark. 2: DIS, University of Rome, Italy. 3: ELRA, Paris, France.

4: IMS, Stuttgart University, Germany. 5: LIMSI-CNRS, Orsay, France. 6: MPI, Nijmegen, The Netherlands

# Contents


1 Introduction ..... 1
1.1 Definitions ..... 2
1.2 Approach ..... 2
1.3 Surveyed data resources ..... 5
1.4 NIMM data resources – purposes, needs and use(r)s ..... 7
2 Dynamic Facial Data Resources with Audio ..... 9
2.1 Advanced Multimedia Processing Lab ..... 9
2.2 ATR Database for bimodal speech recognition ..... 14
2.3 The BT DAVID Database ..... 18
2.4 Data resources from the SmartKom project ..... 23
2.5 FaceWorks ..... 33
2.6 M2VTS Multimodal Face Database ..... 39
2.7 M2VTS Extended Multimodal Face Database – (XM2VTSDB) ..... 46
2.8 Multi-talker database ..... 52
2.9 NITE (Natural Interactivity Tools Engineering) Floor Plan Corpus ..... 60
2.10 Scan MMC (Score Analysed MultiModal Communication) ..... 64
2.11 VIDAS (VIDeo ASsisted with audio coding and representation) ..... 68
2.12 /'VCV/ database ..... 74
3 Dynamic Facial Data Resources without Audio ..... 81
3.1 LIMSI Gaze Corpus (CAPRE) ..... 81
4 Static Facial Data Resources ..... 85
4.1 3D_RMA: 3D database ..... 85
4.2 AR Face Database ..... 89
4.3 AT&T Laboratories Database of Faces ..... 94
4.4 CMU Pose, Illumination, and Expression (PIE) database ..... 98
4.5 Cohn-Kanade AU-Coded Facial Expression Database ..... 102
4.6 FERET Database Demo ..... 106
4.7 Psychological Image Collection at Stirling (PICS) ..... 110
4.8 TULIPS 1.0 ..... 115
4.9 UMIST Face Database ..... 119
4.10 University of Oulu Physics-Based Face Database ..... 124
4.11 VASC – CMU Face Detection Databases ..... 129
4.12 Visible Human Project ..... 135
4.13 Yale Face Database ..... 141
4.14 Yale Face Database B ..... 145
5 Lesser Known/Used Facial Data Resources ..... 149
5.1 3D Surface Imaging in Medical Applications ..... 149
5.2 ATR Database for Talking Face ..... 150
5.3 Audio-Visual Speech Processing Project ..... 151

5.4 Facial Feature Recognition using Neural Networks ..... 152
5.5 Image Database of Facial Actions and Expressions ..... 154
5.6 JAFFE Facial Expression Image Database ..... 158
5.7 Multi-modal dialogue corpus ..... 159
5.8 Photobook ..... 160
5.9 Video Rewrite ..... 161
6 Gesture Data Resources ..... 163
6.1 ATR Multimodal human-human interaction database ..... 163
6.2 CHCC OGI Multimodal Real Estate Map ..... 168
6.3 GRC Multimodal Dialogue during Work Meeting ..... 172
6.4 LIMSI Multimodal Dialogues between Car Driver and Co-pilot Corpus ..... 175
6.5 LIMSI Pointing Gesture Corpus (PoG) ..... 179
6.6 McGill University, School of Communication Sciences & Disorders, Corpus of gesture production during stuttered speech ..... 183
6.7 MPI Experiments with Partial and Complete Callosotomy Patients Corpus ..... 186
6.8 MPI Historical Description of Local Environment Corpus ..... 189
6.9 MPI Living Space Description Corpus ..... 192
6.10 MPI Locally-situated Narratives Corpus ..... 195
6.11 MPI Narrative Elicited by an Animated Cartoon "Canary Row" Corpus 1 ..... 198
6.12 MPI Narrative Elicited by an Animated Cartoon "Canary Row" Corpus 2 ..... 202
6.13 MPI Narrative Elicited by an Animated Cartoon "Maus" and "Canary Row" Corpus ..... 205
6.14 MPI Natural Conversation Corpus ..... 208
6.15 MPI Naturalistic Route Description Corpus 1 ..... 212
6.16 MPI Naturalistic Route Description Corpus 2 ..... 215
6.17 MPI Traditional Mythical Stories Corpus ..... 218
6.18 MPI Traditional Mythical Stories with Sand Drawings Corpus ..... 221
6.19 National Autonomous University of Mexico, DIME multimodal corpus ..... 224
6.20 National Center for Sign Language and Gesture Resources ..... 229
6.21 RWC Multimodal database of gestures and speech ..... 239
6.22 University of Chicago Origami Multimodal corpus ..... 246
6.23 VISLab Cross-Modal Analysis of Signal and Sense Data and Computational Resources for Gesture, Speech and Gaze Research ..... 249
7 Lesser Known/Used Gesture Data Resources ..... 254
7.1 ATR sign language gesture corpora ..... 254
7.2 IRISA Georal Multimodal Corpus ..... 256
7.3 LORIA Multimodal Dialogues Corpus ..... 259
7.4 University of California Video Series on Nonverbal Communication ..... 260
7.5 University of Venice Multimodal Transcription of a Television Advertisement ..... 262
8 Market and User Needs ..... 265
8.1 Introduction ..... 265
8.2 Methodology ..... 265
8.3 Results ..... 266

8.4 Conclusion ..... 269
Acknowledgements ..... 270
9 Appendix 1. Questionnaires collected at Dagstuhl November 2001 ..... 271
9.1 Utrecht University - Denk project ..... 271
9.2 LORIA ..... 274
9.3 MIT Media Lab ..... 277
9.4 Universität Bielefeld - Situated Verbmobil Artificial Communicators ..... 280
9.5 MITRE Corporation - Multimodal referent resolution in map-based interaction ..... 283
9.6 AIST Tokyo - JEITA Multimodal corpora ..... 286
9.7 Microsoft Research ..... 289
9.8 University of Art and Design Media Lab ..... 292
9.9 Linköping University - Swedish Dialogue System ..... 295
9.10 DFKI ..... 298
9.11 Tilburg University ..... 301
9.12 University of Edinburgh HCRC ..... 302
9.13 European Media Laboratory GmbH ..... 305
9.14 SRI / LIMSI / LINC ..... 308
9.15 MITRE ..... 311
9.16 University of Ulster ..... 312
9.17 Universität Erlangen-Nürnberg ..... 313
9.18 Oregon Graduate Institute ..... 316
9.19 Università di Roma La Sapienza ..... 319
9.20 IRST ..... 322
9.21 DFKI ..... 325
9.22 DFKI ..... 328
9.23 LORIA ..... 331
9.24 TU Berlin ..... 334
9.25 Lotus ..... 337
9.26 Mitsubishi Electric Research Laboratories ..... 340
9.27 DFKI ..... 343
9.28 IBM T.J. Watson Research Center ..... 346
10 Appendix 2. Questionnaire ..... 349
11 Appendix 3. Statistics ..... 352

# 1 Introduction

This ISLE (International Standards for Language Engineering) Natural Interactivity and Multimodality (NIMM) Working Group report D8.1 provides a survey of NIMM data resources which include facial expression and/or gesture possibly combined with speech. The report forms part of a series of European ISLE NIMM WG reports on data resources, coding schemes and coding tools for natural interactivity and multimodality. Report D11.1 on NIMM coding tools and report D9.1 on NIMM coding schemes are available at http://isle.nis.sdu.dk. This series of ISLE reports continues the work on coding schemes (deliverable D1.1) and coding tools (deliverable D3.1) surveys done for spoken dialogue in the MATE project (Multilevel Annotation Tools Engineering, 1998-2000). MATE reports are available at http://mate.nis.sdu.dk. The present survey comprises 64 resources world-wide, 36 of which are facial resources and 28 are gesture resources. Several data resources combine speech with facial expression and/or gesture. The report also includes a survey of market and user needs produced by ELRA (the European Language Resources Agency) and 28 questionnaires collected by LIMSI-CNRS at the Dagstuhl workshop on Coordination and Fusion in Multimodal Interaction held in late 2001.

The present report is the result of work in two sub-groups, one looking at facial data resource with or without accompanying speech, and one looking at gesture data resources with or without accompanying speech. The work on facial data resources was led by NISLab in close collaboration with U-ROME and with contributions from ELRA, while the work on gesture resources was led by LIMSI-CNRS with contributions from IMS and MPI. The distribution of expertise in the ISLE NIMM WG made it natural to proceed in this way. Resources involving both facial expression and gesture have of course been included as well. These typically focus on gesture and were thus described by the gesture group. Resources including lip movement and eye/gaze behaviour have usually been included among the facial coding schemes unless the main focus of the resource is on gesture. The data resources include both human-human communication, human-system communication and recordings from artificial situations in which subjects were asked, e.g. to speak certain words in isolation.

The present report seems to be by far the largest existing survey of NIMM data resources. In 2000, COCOSDA (The International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques) has produced a survey of NIMM data resources (http://www.slt.atr.co.jp/cocosda/beijing/multi-modal.files/frame.htm) which includes a relatively small number of NIMM data resources. COCOSDA sent a questionnaire to researchers working in the field of audio-visual speech. The questionnaire included questions about objectives, applications, and contents of data resources, their role in evaluations of multi-modal systems, and how to measure performance. In addition, a few general questions were asked, including the name of the respondent, activity information about the respondent's organisation, suggestions for where COCOSDA might send the questionnaire, projects or events that should appear on COCOSDA's web pages, other relevant web sites, and suggestions for domain-promoting activities. Seven replies seem to have been received, three of which came from the same site. There seems to be partial but very limited overlap with the data resources described in the present report. However, this could not be fully verified since none of the links to the filled questionnaires worked when the website was visited. The page containing links to the filled questionnaires also has links to six other data resources. In these cases, the descriptions seem to have been taken from publications. Again, this could not be verified since the links did not work. Some of these linked-to resources are also described in the present ISLE report.

We hope that the present report will be of interest to colleagues from academia and industry who have a need for, or take an interest in, working with NIMM data resources.

In the following, we first briefly define some central concepts (Section 1.1). We then describe our approach in terms of how the surveyed data resources were selected and described (Section 1.2). Section 1.3 provides an overview of the reviewed resources and whether contact was established to the resource creator(s) of each resource. Finally, Section 1.4 draws some general conclusions from this report, including for which purposes data resources are created, in which areas they are used, and which types of resources users mainly request.

The following chapters describe (i) the facial data resources as divided into dynamic facial data resources with audio (Chapter 2), dynamic facial data resources without audio (Chapter 3), static facial data resources (Chapter 4), and lesser know facial data resources for which we have found little information (Chapter 5), and (ii) the gesture data resources as divided into gesture data resources (Chapter 6) and lesser known gesture data resources for which we have found little information (Chapter 7). Chapter 8 provides information from the ELRA market study together with Appendices 2 and 3 which contain the questionnaire used and statistics, respectively. Finally, Appendix 1 includes the 28 questionnaires collected at the Dagstuhl workshop mentioned above.

## 1.1 Definitions

Despite the fact that this report is termed a survey of NIMM (Natural Interactivity and Multimodality) data resources, it may be useful to point out that its focus is on natural interactivity rather than on multimodality. The following definitions are necessary for understanding this claim.

A *modality* is a particular way of presenting information in some medium. A *medium* is a physical substrate or vehicle for information presentation, such as light/graphics which is being perceived visually, or sound/acoustics which is being perceived auditively. Obviously, there are many different ways of representing information in a particular medium. This is why we need to distinguish between different modalities presented in the same medium. For instance, spoken language is a modality expressed in the acoustic medium, whereas written language and facial expression are modalities expressed in the medium of light/graphics. In the form of lip movements or textual transcription, spoken language may also be expressed in the medium of light/graphics. Thus, the same modality can be represented (more or less adequately) in different media. During human-human-system interaction (see below), a modality may be used as an *input* modality (from a human to the system or to other humans) or as an *output* modality (from the system to humans or, rarely today, to another system), or both. *Multimodal* representations are representations which can be decomposed into two or more *unimodal* modalities. For more details on these basic concepts in Modality Theory, see Bernsen, N. O.: Multimodality in language and speech systems - from theory to design support tool. To appear in Granström, B. (Ed.): *Multimodality in Language and Speech Systems*. Dordrecht: Kluwer Academic Publishers 2002. *Interaction* refers to communication or information exchange between humans, possibly mediated by a computer system, or between humans and computer systems. The term *natural* qualifies the interaction and refers to the ways in which humans normally exchange information with one another.

Given those definitions, it is clear that natural interactive communication is multimodal, using several media and a large number of different modalities of information representation and exchange. However, multimodal exchange of information is not necessarily a case of natural interaction. Multimodal exchange of information might, for instance, and in fact often does, include media which are not perceivable by humans, such as magnetic fields, radar, or ultrasound. It follows that the modalities used in those media are not ones used by humans in their natural interactive communication with each other. In other words, the term "multimodal interaction" is a generic term which subsumes natural interaction as well as information representation and exchange which cannot be described as natural in this sense. It is clear from the surveyed data resources below that the vast majority of resources directly address, or are at least relevant to, natural interaction and its understanding for scientific or application-oriented purposes.

## 1.2 Approach

The approach adopted for producing the present report was to (i) first identify a common set of criteria for selecting the data resources to be described and decide upon issues concerning quality of content as well as of presentation; then (ii) establish a common template for describing each data resource; (iii) identify relevant data resources world-wide based on the web, literature, networking contacts among researchers in the field, etc.; and, finally (iv), interact with the data resource creators to the extent

possible in order to gather information on their resources and ask them to verify the data resource descriptions produced.

### 1.2.1 Selection criteria

To keep the survey focused, the following criteria were adopted for selecting the data resources to be included below:

*Accessibility:* The data resource must be accessible for research and/or industrial purposes. The survey should include an indication of whether a resource is free or if there is a fee to be paid.

*Annotation:* If a data resource has been marked up this is considered an advantage. The coding scheme used should then be included in ISLE report D9.1 if it satisfies the requirements for inclusion in that report. If not, a short informal coding scheme description should be included in the present report alongside the resource description. If not marked up, the resource should be highly suitable for markup and the types of phenomena which could be marked up, should be indicated.

*Exceptions:* Exception to the above is only to be made if a data resource is so rare or innovative for its domain that its very existence might be of interest to researchers in the field.

### 1.2.2 Quality of content and presentation

NIMM data resources are not always easy to get access to. We have adopted the following guidelines for contents inspection, validation and presentation:

*Access:* It is highly desirable that the describer of a certain data resource has actually had access to that resource. If this has not been possible, it should be clearly indicated in the description.

*Validation:* All descriptions should be validated by someone other than the describer, if at all possible with the data resource creator in the loop, either as describer or as validator.

*Examples:* Whenever permissible, a short example of the data resource should be presented in this report. If, for whatever reason, it has not been possible to access and inspect a resource example first-hand, this should be stated clearly in the description.

### 1.2.3 Common description template

In order to help describers and providers of data resource information to document data resources, facilitate the reading of this report, and allow some measure of easy comparison among the data resources presented, resource descriptions have a common structure which, to the extent possible, provides the same information about all data resources. The common structure includes 8 main entries in addition to the resource name as header as shown in Figure 1.2.1. Each main entry subsumes a number of more specific information items.

The common description template went through several revisions as work on the survey proceeded, for instance in order to take into account types of information which would be useful to include but which had not been anticipated from the outset.

| |
|---|
| **Reference (specify resource by project name, main authors or laboratory)** |
| **Description header** |
| Main actor (name and email) |
| Verifying actor (name and email) |
| Date of last modification of the description |
| **References** |
| Web site(s) (Make a short description on what can be found on the site.) |
| Short description |

Illustrative sample picture or video file

References to additional information on the reviewed resource (journal or conference paper, white paper…)

**Recorded human behaviour**

How many different humans have been recorded in the whole resource (none, one, two, more than two)?

How many humans are recorded at the same time (visible in the same frame)?

What is their profile (age, gender, profession…)?

Which human body parts are visible in the resource (face, arms, hand, whole body…)?

Which modalities are annotated (speech, hand gesture, arm gesture, body posture, facial expression …)?

Which other modalities are available/visible in the resource but have not been annotated (speech, hand gesture, arm gesture, body posture, facial expression …)?

**Recorded computer behaviour**

Which interactive media are visible/audible in the resource and are used by the humans (none, graphical screen, computer pen, tactile screen, data glove, loudspeakers…)?

**Recording**

What are the file types included in the resource? Are they organised in a database structure?

How much data does the resource contain (measured in duration, number of dialogues or Mb)?

Who created the resource and when?

How was the resource created?

What is the application area (none, tourism, education, arts … )?

What was the original purpose of creating the resource?

**Accessibility**

How does one get access to the resource?

Is the resource available for free or how much does it cost?

Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?

Did the reviewer have access to the resource to write his/her contribution to 8.1?

**Usage**

Which purpose(s) can the resource be used for/has the resource been used for?

Who used the resource so far/who are the target users of the resource?

Is the resource language dependent (which language(s)) or language independent?

**Conclusion**

How interesting/important/high quality is the resource?

What do the authors regret (if anything) not to have done while building the resource?

**Figure 1.2.1.** Common data resource description template.

## 1.2.4 Interaction with data resources creators

Close interaction with the creators of data resources has been sought throughout the writing of this report, first, of course, to seek their permission to publicly describe their data resource, and secondly to invite their collaboration in producing as useful and accurate information about the data resource as possible. Creators of the data resources reviewed were invited to comment on the description of their data resource and to validate the final description, resulting in feedback on and validations of more than half of the descriptions (including the lesser known ones) in this report, cf. Figure 1.3.1. Many data resource creators pointed out the potential value of the present survey. In a few cases, data resource creators had already answered a questionnaire from COCOSDA (see above) and did not want to repeat a similar exercise.

## 1.3 Surveyed data resources

For each data resource described in the following chapters, an indication is included of which ISLE partner was the main actor in making the description, i.e. the partner that had the main responsibility for describing that particular data resource. In most cases, resource descriptions were verified by another ISLE partner or by the data resource creator, which is then indicated as well. Only for a few resources on which very little information was available, no verifying actor was involved. For each described resource we tried to establish contact to the creator(s) to invite them to verify our descriptions and possibly provide additional information. In a number of cases, we received valuable feedback while in other cases we never succeeded in getting a response. Figure 1.3.1 lists the surveyed data resources in the order in which they are described in this report and indicates for each of them whether or not we received feedback from their creator(s).

* after a data resource name indicates that the creator(s) of the data resource provided feedback on our description.

+ means that the data resource was created at the main actor's site and that feedback on our description thus was provided by a person located at the main actor's site.

- means that we did not succeed in establishing contact to the creator(s) of the data resource.

| | |
|---|---|
| **2** | **Dynamic Facial Data Resources with Audio** |
| 2.1 | Advanced Multimedia Processing Lab- |
| 2.2 | ATR Database for bimodal speech recognition- |
| 2.3 | The BT DAVID Database- |
| 2.4 | Data resources from the SmartKom project* |
| 2.5 | FaceWorks- |
| 2.6 | M2VTS Multimodal Face Database- |
| 2.7 | M2VTS Extended Multimodal Face Database – (XM2VTSDB)- |
| 2.8 | Multi-talker database- |
| 2.9 | NITE Floorplan Corpus (Natural Interactivity Tools Engineering)+ |
| 2.10 | Scan MMC (Score Analysed MultiModal Communication)+ |
| 2.11 | VIDAS (VIDeo ASsisted with audio coding and representation)- |
| 2.12 | /'VCV/ database* |
| **3** | **Dynamic Facial Data Resources without Audio** |
| 3.1 | LIMSI Gaze Corpus (CAPRE)+ |
| **4** | **Static Facial Data Resources** |
| 4.1 | 3D_RMA: 3D database* |
| 4.2 | AR Face Database- |
| 4.3 | AT&T Laboratories Database of Faces* |
| 4.4 | CMU Pose, Illumination, and Expression (PIE) database* |
| 4.5 | Cohn-Kanade AU-Coded Facial Expression Database* |
| 4.6 | FERET Database Demo- |
| 4.7 | Psychological Image Collection at Stirling (PICS)- |
| 4.8 | TULIPS 1.0* |
| 4.9 | UMIST Face Database* |
| 4.10 | University of Oulu Physics-Based Face Database* |
| 4.11 | VASC – CMU Face Detection Databases* |
| 4.12 | Visible Human Project* |
| 4.13 | Yale Face Database* |

| | |
|---|---|
| 4.14 | Yale Face Database B* |
| **5** | **Lesser Known/Used Facial Data Resources** |
| 5.1 | 3D Surface Imaging in Medical Applications- |
| 5.2 | ATR Database for Talking Face- |
| 5.3 | Audio-Visual Speech Processing Project- |
| 5.4 | Facial Feature Recognition using Neural Networks- |
| 5.5 | Image Database of Facial Actions and Expressions- |
| 5.6 | JAFFE Facial Expression Image Database- |
| 5.7 | Multi-modal dialogue corpus- |
| 5.8 | Photobook- |
| 5.9 | Video Rewrite- |
| **6** | **Gesture Data Resources** |
| 6.1 | ATR Multimodal human-human interaction database- |
| 6.2 | CHCC OGI Multimodal Real Estate Map- |
| 6.3 | GRC Multimodal Dialogue during Work Meeting- |
| 6.4 | LIMSI Multimodal Dialogues between Car Driver and Copilot Corpus+ |
| 6.5 | LIMSI Pointing Gesture Corpus (PoG)+ |
| 6.6 | McGill University, School of Communication Sciences & Disorders, Corpus of gesture production during stuttered speech- |
| 6.7 | MPI Experiments with Partial and Complete Callosotomy Patients Corpus+ |
| 6.8 | MPI Historical Description of Local Environment Corpus+ |
| 6.9 | MPI Living Space Description Corpus+ |
| 6.10 | MPI Locally-situated Narratives Corpus+ |
| 6.11 | MPI Narrative Elicited by an Animated Cartoon "Canary Row" Corpus 1+ |
| 6.12 | MPI Narrative Elicited by an Animated Cartoon "Canary Row" Corpus 2+ |
| 6.13 | MPI Narrative Elicited by an Animated Cartoon "Maus" and "Canary Row" Corpus+ |
| 6.14 | MPI Natural Conversation Corpus+ |
| 6.15 | MPI Naturalistic Route Description Corpus 1+ |
| 6.16 | MPI Naturalistic Route Description Corpus 2+ |
| 6.17 | MPI Traditional Mythical Stories Corpus+ |
| 6.18 | MPI Traditional Mythical Stories with Sand Drawings Corpus+ |
| 6.19 | National Autonomous University of Mexico, DIME multimodal corpus* |
| 6.20 | National Center for Sign Language and Gesture Resources* |
| 6.21 | RWC Multimodal database of gestures and speech- |
| 6.22 | University of Chicago Origami Multimodal corpus- |
| 6.23 | VISLab Cross-Modal Analysis of Signal and Sense Data and Computational Resources for Gesture, Speech and Gaze Research- |
| **7** | **Lesser Known/Used Gesture Data Resources** |
| 7.1 | ATR sign language gesture corpora* |
| 7.2 | IRISA Georal Multimodal Corpus- |
| 7.3 | LORIA Multimodal Dialogues Corpus- |
| 7.4 | University of California Video Series on Nonverbal Communication- |
| 7.5 | University of Venice Multimodal Transcription of a Television Advertisement- |

**Figure 1.3.1.** Verification of data resource by its creator(s).

## 1.4 NIMM data resources – purposes, needs and use(r)s

This section briefly summarises the purposes for which the surveyed data resources were created and which needs they address. Also, conclusions from the ELRA market study on data resources are briefly described.

### 1.4.1 Collected data resources

The collected data resources reflect a multitude of needs and purposes, including the following (in random order):

- automatic analysis and recognition of facial expressions, including lip movements;
- audio-visual speech recognition;
- study of emotions, communicative facial expressions, phonetics, multimodal behaviour, etc.;
- creation of synthetic characters, including, e.g., talking heads;
- automatic person identification;
- training of speech, gesture and emotion recognisers;
- multimodal and natural interactive systems specification and development.

In many cases, the people working with the data, in particular those working with image analysis, have created their own resource databases. Algorithms for image analysis are sometimes dependent on lighting conditions, picture size, subjects' face orientations, etc. Thus, computer vision research groups often have had to create their own image databases. Image analysis using computer vision techniques remains a difficult task, and this may be the reason why we have primarily found static image resources produced by workers in this field.

In other areas, video recordings - mostly including audio - are needed. For example, studies of lip movements during speech, co-articulation, audio-visual speech recognition, temporal correlations between speech and gesture, and relationships among gesture, facial expression, and speech, all require video recordings with audio.

Significantly, across all the collected data resources, re-use is a rare phenomenon. If a resource has been created for a specific application purpose, it has usually been tailored to satisfy the particular needs of its creators, highlighting, e.g., particular kinds of interaction or the use of particular modality combinations. However, the lack of re-use may also to some extent be due to the fact that existing resources may be difficult to locate. On the other hand, it should be mentioned that vendors of data resources exist (e.g. ELRA and LDC).

### 1.4.2 Market study

The ELRA market study described in Chapter 8 shows that the NIMM data resources in which people seem most interested include audio, video and image resources. Audio is most popular (mentioned by 84% of the respondents) followed by video (mentioned by 52%) and static images (mentioned by 28%). If a data resource has also been annotated, this is considered an advantage since value has been added. In many cases, the users of data resources produce the resources they need themselves. Sometimes these resources are offered to other users.

The questionnaire which was used to collect information on data resources and user needs mentioned six general task categories for which data resources may be used: authentication, recognition, analysis, synthesis, control, and other. For each category, a number of more specific possibilities were listed. Respondents were supposed to indicate the kinds of applications they were interested in. The primary applications of data resources according to the answers are: information retrieval, speech recognition, speech verification, face recognition, speech/lips correlation, and voice control.

To get an idea of the overall market areas for data resources, the questionnaire listed five possibilities (including “other”) among which respondents were asked to choose the ones they found appropriate to their work. The area mentioned most frequently was research followed by information systems development (e.g. banking, tourism, telecommunication), web applications development, education/training, and edutainment. Other areas proposed include security, control of consumer devices, and media archiving for content providers.

# 2 Dynamic Facial Data Resources with Audio

This introduction covers chapters 2, 3, 4 and 5, all of which have a primary focus on facial data resources. Chapter 2 presents dynamic facial data resources with audio, Chapter 3 describes dynamic facial data resources without audio, Chapter 4 describes static facial data resources, and in Chapter 5 we describe lesser know facial data resources for which we have found little information.

Most of the data resources involving faces were found on the web. Many of the databases can be downloaded from the web for free. In all cases, information is provided on how to contact the creators of a particular data resource and ask them how to get access to it.

Facial data resources have been created for a number of different purposes, including, but not limited to:

- automatic analysis and recognition of facial expressions, including lip movements;
- audio-visual speech recognition;
- study of emotions, communicative facial expressions, and phonetics;
- creation of synthetic characters, e.g. talking heads;
- automatic person identification;
- training of emotion recognisers.

Depending on the purpose, the data resource may include, e.g., dynamic images with certain lighting conditions, picture size, and subjects' face orientations, VCV productions (vowel-consonant-vowel) by several speakers, or (for studies of lip movements) (a) speaker(s) saying a few words from a given dictionary.

## 2.1 Advanced Multimedia Processing Lab

### 2.1.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

May 14th, 2001.

### 2.1.2 References

***Web site***

The Advanced Multimedia Processing Labs project web site:
http://amp.ece.cmu.edu/projects/AudioVisualSpeechProcessing

The form to use for accessing the resource:
http://amp.ece.cmu.edu/projects/AudioVisualSpeechProcessing/download.htm

***Short description***

A high quality digital data resource with facial expressions and speech.

***Illustrative sample picture or video file***

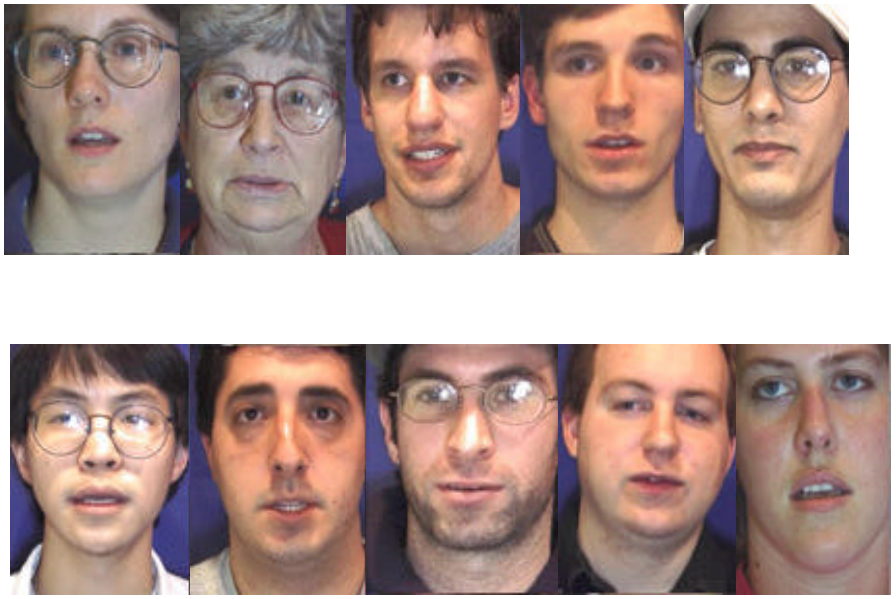

**Figure 2.1.1.** The ten subjects, whoose faces are recorded in the resource.

***References to additional information on the reviewed resource***

None.

## 2.1.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

10 subjects (7 males and 3 females) have been recorded. The audio consists of a vocabulary, which includes 78 isolated words commonly used for time, such as, "Monday", "February", "night", etc. Each word is repeated 10 times.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

No information has been provided on their profile.

***Which human body parts are visible in the resource?***

Only the face is visible in the resource.

***Which modalities are annotated?***

No information is available.

***Which other modalities are available/visible in the resource but have not been annotated?***

Since only the face in visible in the recordings the only modalities available besides facial expressions are lip movements, and perhaps gaze and eye behaviour.

### 2.1.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 2.1.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The format of the resource is QuickTime files.

***How much data does the resource contain?***

There are altogether 100 QuickTime files, each one containing one subject articulating all the words in the vocabulary. Each file is about 450MBytes. The audio signals have been sampled as PCM, 44.1KHz, 16 bit, mono.

***Who created the resource and when?***

The resource was created by Fu Jie Huang, Carnegie Mellon University.

***How was the resource created?***

The recording of the data resource was set up in a soundproof studio to collect noise-free audio data. Controlled light and blue-screen background were used to collect the image data. A SONY digital camcorder with tie-clip microphone to record the data on DV tapes was used. The data on DV tapes was transferred to a PC by the Radius MotoDV program and stored as QuickTime files. Since the data on the DV tapes were already digital, there was no quality loss when transferred to the PC.

***What is the application area?***

No information available.

***What was the original purpose of creating the resource?***

No information available.

### 2.1.6 Accessibility

***How does one get access to the resource?***

One has to fill out a form on the web:

http://amp.ece.cmu.edu/projects/AudioVisualSpeechProcessing/download.htm

When one has filled out the form and submitted it, one may either correct the information in the form, or "click to continue", which brings one to a new page where one can "click to download audio-visual speech processing dataset". Now one can download the texts files, sound files and transcription files in zip format.

The Quick Time files cannot be downloaded, due to the fact that it takes up to 60GB storage. Instead they are distributed upon request. Write Fu Jie Huang, jhuangfu@cmu.edu, if you want to request a copy of the Quick Time files.

***Is the resource available for free or how much does it cost?***

The resource is free and can be downloaded from the web. One is provided immediate access when filling in the form.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added to the original resource in the form of transcription files.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The reviewers had access to the resource through the web and to web information.

### 2.1.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource can be used to develop speech-reading techniques for higher speech recognition accuracy. The audio file and the extracted texts are primarily for lip-reading research purposes.

***Who used the resource so far/who are the target users of the resource?***

Since the resources can be freely downloaded from the web, it is assumed that there are no statistics available on this issue. It is furthermore assumed that researchers at CMU have used the resource.

***Is the resource language dependent or language independent?***

The resource is language dependent due to the audio being English.

### 2.1.8 Conclusion

***How interesting/important/high quality is the resource?***

The data resource is digital and of high quality. Furthermore, it has the advantage of being free.

***What do the authors regret ( if anything) not to have done while building the resource?***

No information available.

## 2.2 ATR Database for bimodal speech recognition

### 2.2.1 Description header

***Main actor***

IMS : Ulrich Heid (uli@IMS.Uni-Stuttgart.DE)

***Verifying actor***

IMS: Steve Berman (steve@ims.uni-stuttgart.de)
LIMSI: Jean-Claude MARTIN (martin@limsi.fr)

***Date of last modification of the description***

August, 4th (nakamura@slt.atr.co.jp has been contacted by email on August 8th)

### 2.2.2 References

***Web site(s)***

Dr. Satoshi Nakamura's own web site: http://isw3.aist-nara.ac.jp/IS/Shikano-lab/staff/teacher/nakamura/nakamura1_E.html
Same as above in Japanese: http://www.rwcp.or.jp/wswg/rwcdb/

***Short description***

The resource contains speech and talking face image data for bimodal speech recognition.

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Nakamura, S. et al: Multimodal Corpora for Human-Machine Interaction Research, Proc. of ICSLP, Volume IV, pp. 25-28, 2000

Kakihara, K., Nakamura, S., and Shikano, K.: Speech-to-Face Movement Synthesis Based on HMMs, Proc. ICME2000, 2000.

### 2.2.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

One human has been recorded in the whole resource.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

The subject is a female speaker. No other information available.

***Which human body parts are visible in the resource?***

The face of the subject is visible in the resource.

***Which modalities are annotated?***

No information available.

***Which other modalities are available/visible in the resource but have not been annotated?***

None.

### 2.2.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

No information available.

### 2.2.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

SGI movie (audio sampling at 16 bit, 48 kHz; video sampling at 30 frames/sec, size 160 * 120, colour: 8 bit RGB).

***How much data does the resource contain ?***

The resource contains a video capturing one female speaker on the ATR 5240 Japanese words task.

***Who created the resource and when?***

The resource has been created by a consortium of ATR, Sharp and Tsukuba Electrotechnical Laboratories. An exact date is not indicated. The level of development described here corresponds to the status of the project at the point of publication of.

***How was the resource created?***

The speaker is in a sound-proof room, with lighting set from the front. Two consecutive day of recordings; the speaker is asked to sit with her back against the back of the seat.

***What is the application area?***

No information available.

***What was the original purpose of creating the resource?***

The purpose of the creation of the resource is to get access to bimodal speech and talking face data, for speech recognition and for speech-to-lip generation (animated agents, talking face). As the resource is meant as a support tool for research, only the two modalities are observed, however by keeping track of different variation parameters (e.g. lighting position, head position etc.).

## 2.2.6 Accessibility

***How does one get access to the resource?***

The paper cited does not mention the possibility to access the data directly.

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

No information available.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

No. The review is based on the references listed above.

## 2.2.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for observations on the difference of lighting conditions, size of lips, and inclination of a face.

***Who used the resource so far/who are the target users of the resource?***

The resource is used by the consortium that has created it.

***Is the resource language dependent or language independent?***

Language dependent: The data are in Japanese.

### 2.2.8 Conclusion

***How interesting/important/high quality is the resource?***

Although the recording of the data has been done under controlled and carefully worked out conditions, the resource is only of secondary importance to ISLE. Reasons for this are the following:

- The resource only contains one speaker, and two modalities;
- The language is Japanese;
- It is not clear which kind of annotations there are;
- It is unclear, at least from the publications, whether the resource can be used for research outside the laboratory where it was produced.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 2.3 The BT DAVID Database

## 2.3.1 Description header

***Main actor***

ELDA: Niklas Paulsson (paulsson@elda.fr)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

June 27th, 2001.

## 2.3.2 References

***Web site***

Short description of the BT (British Telecommunications) DAVID (Digital Audio-Visual Integrated Database) database: http://galilee.swan.ac.uk/homepages/Home/data/data1.htm

***Short description***

The BT DAVID database contains full-motion video, showing a full face and a profile view of talking subjects and associated synchronous speech. The main purpose is to allow researchers to conduct work on audio-visual technologies in: speech and person recognition, synthesis and communication of audio-visual signals.

***Illustrative sample picture or video file***

**Figure 2.3.1.** Still pictures grabbed from the videos of the database, showing full face and profile view of talking subjects, plus a close-up view of the lips of one subject marked with blue in order to stand out more clearly.

***References to additional information on the reviewed resource***

C. C. Chibelushi, S Gandon, J. S. Mason, F. Deravi, and D Johnston: Design Issues for a Digital Integrated Audio-Visual Database. IEE Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication. London, Digest Reference Number 1996/213, pages 7/1-7/7, November 1996.

### 2.3.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

More than 100 subjects (males and females) have been recorded including 31 clients in 5 sessions and 92 impostors in 1 session.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

50/50 male/female. No other information is available.

***Which human body parts are visible in the resource?***

The face of each subject is visible in the resource.

***Which modalities are annotated?***

No information available.

***Which other modalities are available/visible in the resource but have not been annotated ?***

Facial expression is available as a modality in the resource as well as highlighted lip movements and synchronous speech.

### 2.3.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 2.3.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The file types are organised in a database structure and consist of video recordings as MPEG movies, and audio recordings in .wav and .au format.

***How much data does the resource contain?***

The DAVID database contains full-motion video, showing a full-face and a profile view of talking subjects, together with the associated synchronous sound. DAVID includes audio-visual material from more than 100 subjects including 30 clients recorded on 5 sessions spaced over several months. The utterances include the English digit set, English alphabet E-set {'B', 'C', 'D', 'E', 'G', 'P', 'T', 'V'}, vowel-consonant-vowel syllables, and phrases for the control of a video-conferencing session. The scenes include variable scene background complexity and illumination. Portions of the database include lip highlighting.

The database is divided into 4 corpora each addressing a particular research theme.

Corpus 1 is concerned with the theme of facial image segmentation. One subset of this recording, with one subject only, includes variable illumination and facial distractors such as glasses and hats. The second subset (6 subjects) includes backgrounds of variable complexity. The recordings are of the subjects uttering the digit set.

Corpus 2 is designed for research in automatic speech and person recognition. There are 6 subsets, the make-up of which includes multiple session recordings from 31 clients and single sessions from 92 impostors. Some subsets have complex scene backgrounds and some incorporate a profile view as well as a frontal view. The utterances are the E-set and digits. Two subsets include nine subjects with blue lip highlighting.

Corpus 3 is designed for use in speech-assisted video compression and synthesis of talking heads. Here the recordings are vowel-consonant-vowel-consonant-vowel sequences from 5 subjects recorded in one session.

Corpus 4 is also aimed at automatic speech and person recognition with particular application in the voice control of video-conferencing resources. Here sentences from a business control set are recorded from 31 clients and 92 impostors. One subset has plain scene backgrounds and profile views while the other has no face profile but includes complex scene backgrounds. The clients are recorded at 5 different sessions and the impostors are recorded only once.

***Who created the resource and when?***

The BT DAVID database was compiled by the Speech and Image Processing Group at the University of Wales, Swansea under a contact from BT Laboratories.

***The recording period lasted from 28 November to 12 August 1996.How was the resource created?***

The subjects were videotaped on analogue SVHS tapes.

The recorded data have been checked for quality; the overall quality scores are: 98.7 % of the recordings are ‘good’, 1.0 % are ‘usable’, 0.3 % are ‘bad’.

***What is the application area?***

Expected application areas are automatic speech/person recognition, synthesis and communication of audio-visual signals.

***What was the original purpose of creating the resource?***

The main purpose of the BT DAVID database is to allow researchers to conduct work on audio-visual technologies in: speech or person recognition, synthesis and communication of audio-visual signals. The content of the database, and hence its usefulness, is determined by the experimental themes supported by the database. Consequently, a set of experimental designs spanning the target application areas, were the starting point for the specification of the BT DAVID database.

### 2.3.6 Accessibility

***How does one get access to the resource?***

The complete BT DAVID database is distributed on 18 SVHS video tapes by ELDA.

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

No

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

No, the material was taken from descriptions given on the DAVID database web page as well as a report from UROME.

### 2.3.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The aim of the BT DAVID database was to design and collect an audio-visual database to support research in the following application areas:

- automatic speech/person recognition for:
    - terminal interfaces
    - automated transaction machines
    - telephone kiosks providing the BT charge card service
- speech assisted video (de)coding or synthesis of talking heads
- voice control of video conferencing resources

***Who used the resource so far/who are the target users of the resource?***

Researchers.

***Is the resource language dependent or language independent?***

The resource is language dependent due to the speech being in English.

### 2.3.8 Conclusion

***How interesting/important/high quality is the resource?***

The BT DAVID database is a large data resource with a variety of settings.

***What do the authors regret ( if anything) not to have done while building the resource?***

No information available.

# 2.4 Data resources from the SmartKom project

## 2.4.1 Description header

***Main actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Verifying actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Date of last modification of the description***

June 12th, 2001.

## 2.4.2 References

***Web site***

Information on or the data resource itself is not available from a web site.

The SmartKom web site: www.smartkom.com

The SmartKom website of the Institute of Phonetics and Speech Communication (where the data is recorded): http://www.phonetik.uni-muenchen.de/phonetik_frames.html. SmartKom can be found under the link "Projects".

The corpus will be publicly available from the BAS (Bavarian Archive for Speech Signals):

http://www.bas.uni-muenchen.de/Bas/

***Short description***

The data resource will consist of recordings of Wizard of Oz experiments, where the subject solves certain tasks (like planning a trip to the cinema, programming a VCR or navigating in a foreign town) with a multimodal dialogue system. No data is available at the time of writing this survey.

When the project is finished four resources will have been produced:

1. SmartKom Multimodal Corpus – Cinema. The collection of this corpus has ended at the time of writing.
2. SmartKom Multimodal Corpus - Tourist Information. The collection of this corpus is progressing at the time of writing.
3. SmartKom Multimodal Corpus – TV. The collection of this corpus is planned to take place in the future.
4. SmartKom Multimodal Corpus – Office. The collection of this corpus is planned to take place in the future.

***Illustrative sample picture or video file***

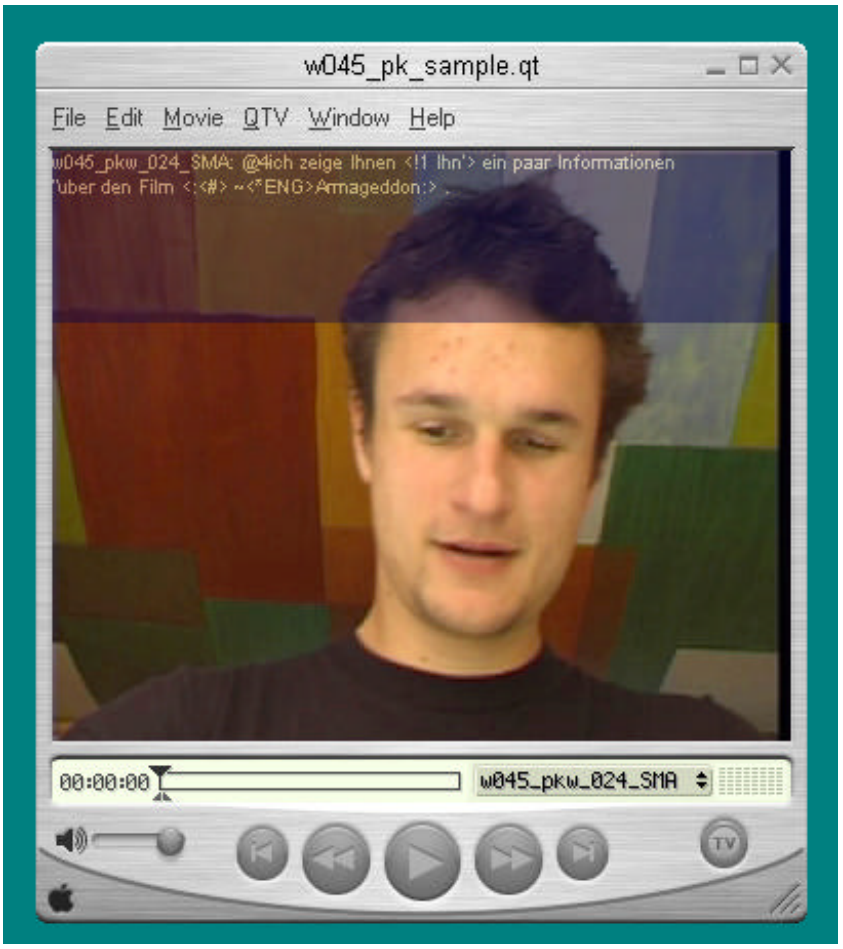


**Figure 2.4.1.** Default view on the data sample "w045" using the QuickTime Player. It shows the picture of the front camera, the audio transliteration and a popup menu for navigating to turn starts (on the right side of the timeline).

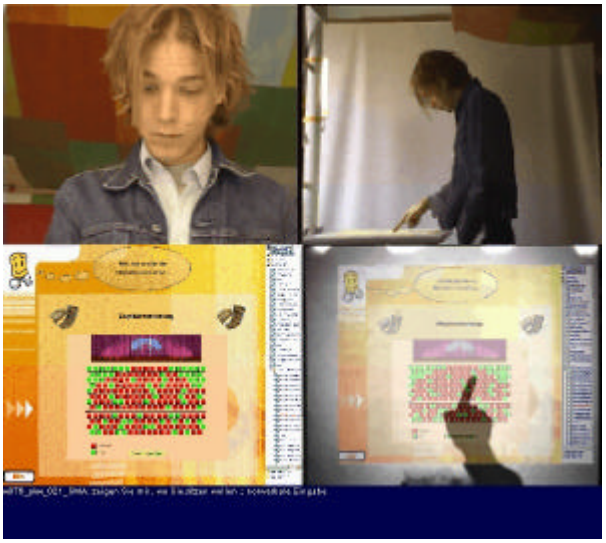

**Figure 2.4.2.** Still picture from a multi-picture video track used for gesture labelling. This will substitute the side view video track in the future.

Further examples can be found on http://isle.nis.sdu.dk under “Reports” -> Deliverable 8.1.

***References to additional information on the reviewed resource***

Beringer, N. et. al.: Transliteration spontansprachlicher Daten – Lexikon der Transliterationkonventionen – SmartKom (version 1). SmartKom Technishes Dokument Nr. 2, February 2000. These can be downloaded from: http://www.phonetik.uni-muenchen.de/phonetik_frames.html - Click “Projects”, then “SmartKom” and then “Transliterationskonventionen”.

Beringer, Nicole: Evoking Gestures in SmartKom - Design of the Graphical User Interface. Submitted to Gesture Workshop, London. Can be downloaded from: http://www.phonetik.uni-muenchen.de/SmartKom_LMU.html

Steininger, Silke: Transliteration of language and labelling of emotion and gestures in SmartKom, 2000. Proceedings from workshop on "Meta-Description and Annotation Schemes for Multimodal Language Resources" - Second International Conference on Language Resources and Evaluation, Athens 2000, pp. 49-51.

Steininger, Silke: Labeling of Gestures in SmartKom - Concept of the Coding System. Submitted to Gesture Workshop, London. Can be downloaded from: http://www.phonetik.uni-muenchen.de/SmartKom_LMU.html

Türk, Ulrich: The Technical Processing in SmartKom Data Collection - A Case Study, submitted for the Eurospeech 2001, Aalborg, Denmark, 3-7 September 2001. Can be downloaded from: http://www.phonetik.uni-muenchen.de/SmartKom_LMU.html

### 2.4.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

The SmartKom-Cinema-Recordings were done with 45 different subjects for Wizard-of-Oz recordings and additional 20 different subjects for test recordings (non naive subjects).

At the moment the SmartKom Multimodal Corpus - Tourist Information is being recorded. Since the recordings are not done, no information is available about the profile of the humans recorded.

Two additional sets of corpora are planned to be recorded later:

- SmartKom Multimodal Corpus - TV
- SmartKom Multimodal Corpus - Office

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

The user profiles are documented in detail in a database, which will be distributed together with the final data. The range covers native German user of all age (see below), non-native German users with good German language capabilities, experienced computer users as well as complete laymen.

- Wizard-of-Oz recordings: 25 female, 20 male; 30 students, 12 employed persons, 1 pupil, 1 retired person, 1 no information; age between 17 and 61 (mean: 29).
- Test recordings: 7 female, 13 male; 14 students, 8 employed persons; age between 22 and 47 (mean: 29,7).

***Which human body parts are visible in the resource?***

The face, arms, hands and upper body are visible in the resource.

***Which modalities are annotated?***

The available input modalities in the resource are spontaneous speech, gestures and facial expressions and upper body movements. All these modalities are being annotated for the SmartKom project.

***Which other modalities are available/visible in the resource but have not been annotated ?***

Body posture of the upper body is visible, but not labelled.

### 2.4.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

In the cinema recordings a pen is not used (and not visible), but it will be used in later recordings.

The microphone array, the directional microphone and the clip on microphone (if used) are visible. The subjects do not use them.

During the recordings different kind of background noise with different loudness levels is played from loudspeakers, but the loudspeakers are not visible.

### 2.4.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The format of the resource is for the moment video and audio recordings. These are in the process of being transformed into QuickTime format.

***How much data does the resource contain?***

The resource is distributed on DVDs with about 3700 MB to 4470 MB on each. The DVDs contain one to three dialogues and each dialogue is 4-5 minutes long.

Content of one DVD:

- Digital video recording of the face
- Digital video recording of a side view of the subject
- 2-dimensional infrared recording of the gestures from above
- Beamer output (the display)
- Coordinates of pointing recorded with the gesture recognizer
- Several audio files (microphone array, a directional microphone and (alternating) a headset or a clip-on-microphone)

For each dialogue there is a whole set of files: one QuickTime File. (".qt") and several ".wav", ".mov" and ".avi" files.

The files can be used separately, e.g. the videos (.mov, .avi) can be viewed with a QuickTime player. The QuickTime file is used for a combined viewing of all or a subset of the files (video and audio and time aligned transcription).

Gesture and emotion label files will be added in the future (at the moment they are only distributed separately via FTP when they are finished).

The first part of the corpus comprises Wizard-of-Oz human-machine dialogues with the task »looking for information on movies/cinemas/restaurants«. There are:

- 90 WOZ-dialogues, which were made between 28 July 2000 and 31 January 2001.
- 35 dialogues with non-naive subjects. These are additional to the 90 WOZ-dialogues and are test recordings with the same set up but non-naive subjects. These were made between 16 March 2000 and 26 July 2000.

At the moment and during the following years of the project dialogues on different tasks will be collected (e.g. tourist information, telephone, email).

***Who created the resource and when?***

The data collection for the SmartKom project began in January 2000 and will continue until 2003. The data collection is done by the Institute of Phonetics and Speech Communication in Munich for the SmartKom consortium.

***How was the resource created?***

For the audio recordings a microphone array of 4 microphones a directional microphone or alternating a headset or a clip-on-microphone have been used.

For the video recordings a digital camera has been used to capture the face of the subjects in order to capture the facial expressions. A second digital camera has been used to capture the gestures of the subject from a side view of the full height of the subjects. Furthermore an infrared camera (from a gesture recogniser: SIVIT/Siemens) has been used to capture the 2-dimensional hand gestures. The coordinates of the pointing gestures on the workspace have been recorded with SIVIT as has the input from the pen.

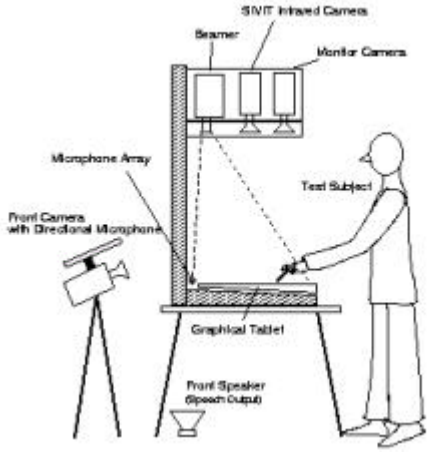


**Figure 2.4.3.** Schematic illustration of the SmartKom setup used for data collection.

***What is the application area?***

The application area of the resource is tourism, movie information, cinema information, restaurant information, telephone, e-mail, TV, fax and calendar. Hotel as 2nd task.

***What was the original purpose of creating the resource?***

The SmartKom project aims at developing an intelligent human-computer interface that allows computer novices to communicate naturally with an adaptive and self-explanatory machine. The original purpose of creating the resource was to collect data for the training of speech, gesture and emotion recognisers, to develop dialogue and context models and to investigate how users interact with a machine that has far greater communication skills than what is the usual situation at the time of writing. In the WOZ experiments the subjects are made to believe that the system they interact with is already fully functional, but many functions are only simulated by two "wizards" that control the system from another room.

## 2.4.6 Accessibility

***How does one get access to the resource?***

The data will be available one year after publication for the SmartKom partners, cf. table 1.3.1. Contact Dr. Florian Schiel, E-Mail: schiel@phonetik.uni-muenchen.de for information on the availability of the data.

Contact Dr. Silke Steininger, E-Mail: kstein@phonetik.uni-muenchen.de for more information about the corpus.

The corpus will be publicly available from the BAS (Bavarian Archive for Speech Signals):

http://www.bas.uni-muenchen.de/Bas/via DVD.

The transcription label files will be distributed together with the data on DVD, updates will be available via FTP.

The transcription and label files are distributed separately via FTP since they are sometimes produced after the completion of the DVDs.

List of dates for publication of the DVDs:

| DVD-Nr | Sessions | Date of public availability | DVD-Nr | Sessions | Date of public availability |
|---|---|---|---|---|---|
| 1.1 | d005_pk<br>d006_pk<br>d007_pk | 10.11.2001 | 21.0<br>22.0<br><br>23.0<br>24.0<br>25.0<br>26.0<br>27.0 | w062_pk<br>p022_pk<br>p024_pk<br>w048_pk<br>w049_pk<br>w059_pk<br>w060_pk<br>w045_pk | 30.02.2002 |
| 2.2<br><br>3.2<br><br>4.0 | w001_pk<br>w003_pk<br>w004_pk<br>w006_pk<br>w009_pk<br>w011_pk | 15.12.2001 | 28.1<br>29.1<br>30.0<br>31.0<br>32.0<br><br>33.0<br>34.0<br><br>35.0<br>36.0 | w040_pk<br>w054_pk<br>w075_pk<br>w039_pk<br>w010_pk<br>w035_pk<br>w041_pk<br>w036_pk<br>w043_pk<br>w042_pk<br>w022_pk<br>w026_pk | 30.03.2002 |

| DVD-Nr | Sessions | Date of public availability | DVD-Nr | Sessions | Date of public availability |
|---|---|---|---|---|---|
| | | | 37.0 | w023_pk | |
| | | | 38.0 | w027_pk | |
| | | | | w033_pk | |
| | | | 39.0 | w029_pk | |
| | | | | w031_pk | |
| 5.0 | w015_pk | 30.12.2001 | 40.0 | w090_pk | 15.04.2002 |
| | w037_pk | | 41.0 | w061_pk | |
| 6.1 | w002_pk | | 42.0 | w089_pk | |
| | w038_pk | | 43.0 | w012_pk | |
| 7.1 | w007_pk | | | w028_pk | |
| | w014_pk | | 44.0 | w088_pk | |
| 8.0 | p002_pk | | 45.0 | w077_pk | |
| | p006_pk | | 46.0 | w024_pr | |
| | | | 47.0 | w087_pk | |
| | | | 48.0 | w065_pk | |
| | | | 49.0 | w086_pk | |
| 9.0 | p007_pr | 15.01.2002 | 50.0 | w030_pk | 30.04.2002 |
| | p018_pr | | | w052_pk | |
| 10.0 | p027_pk | | 51.0 | w064_pk | |
| | w016_pk | | 52.0 | w068_pk | |
| 11.0 | w018_pk | | 53.0 | w046_pk | |
| | w013_pk | | | w055_pk | |
| | w020_pk | | | | |
| 12.0 | w057_pk | 15.02.2002 | 54.0 | w047_pk | 15.05.2002 |
| 13.0 | w017_pk | | 55.0 | w084_pk | |
| | w025_pk | | | | |
| 14.0 | w058_pk | | | | |
| 15.0 | w066_pk | | | | |
| 16.0 | w063_pk | | | | |
| 17.0 | w019_pk | | | | |
| | w050_pk | | | | |
| 18.0 | p040_pk | | | | |
| | p041_pk | | | | |
| 19.0 | p032_pk | | | | |
| 20.0 | w044_pk | | | | |

**Figure 2.4.4.** List of dates for publication of the DVDs with data. Only the highest version of a given DVD is available because the lower version contains mistakes that were corrected in the next version.

### *Is the resource available for free or how much does it cost?*

The price of the resource has not been decided upon yet, but at the moment it is 70 DM for each DVD for the partners. The price of the public available data will be determined by the costs that the BAS has to cover for the final distribution. The final prices for scientific and commercial usage will be published on the BAS catalogue as well as on the ELRA catalogue.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added to the original resource in the form of a labelling of the recorded spontaneous speech at the word level using a broad orthographic transcription system based on the system used in the VerbMobil project. http://www.phonetik.uni-muenchen.de/phonetik_frames.html. Click "Projects", then "SmartKom" and then "Transliterationskonventionen additional information" http://www.phonetik.uni-muenchen.de/Bas/BasKorporaeng.html; http://verbmobil.dfki.de/overview-us.html

Complementary to the orthography a list of conventions is used to code such things as corrections, repetitions, reductions, hesitations as well as technical artefacts and superimposition of the speech of the user and the machine.

The transcription categories that have been used for annotating the data are:

- Lexical units (e.g. words, classified words, compounds)
- Syntactical-semantical structure (e.g. sentences, repetitions, false starts)
- Non-verbal articulatory productions (e.g. hesitations, breathing, laughing)
- Noises, technical artefacts
- Acoustic superimpositions
- Comments (peculiarities of the grammar, pronunciation)
- Prosody (pauses, phrase boundaries, accentuation)

The transcription system has certain limitations due to being a broad annotation system. E.g. it does not support phonological or phonetical annotation.

The alignment of the audio signal is made via turn markers. The audio signal is segmented manually into turns. The names of the turn markers can be found again in the transcription. An additional automatic segmentation is planned for further annotations.

Moreover, there had to be made some additions, changes and deletions in the annotation rules of the Verbmobil system to make it suitable for the requirements of the SmartKom system and to fasten annotation and correction passes. For this aim the differentiation of noise categories in the annotation was reduced to a minimum of three noise categories. Changes were made in the conventions for names and compound words. However, some new features had to be added, which cannot be found in Verbmobil, like the marking of prosodic boundaries, intonation and accents, keywords, system-related pauses and Off-Talk.

The labelling system for gestures has been defined Instead of coding the precise morphological shape the researchers in the project will try to use a simplified, practice-oriented system, where two broad categories are labelled - head gestures and hand gestures.

The hand gestures are defined functionally/intentionally (not morphologically). Each gesture is first sorted in one of three categories: interactional gestures (requests to the systems or answers to questions from the system), supporting gestures (gestures that support non-interactional activities like searching or reading - mostly these are preparations for interactions with the system) and residual gestures (emotional gestures and non identifiable gestures).

The interactional gestures are: Pointing long and short, with and without touching of the gestures; circling with and without touching of the display, complex gestures.

The supporting gestures are divided in continual ones (hand is moving over the display) and punctual ones (focused to one point). There are: continual-reading, continual-pondering, continual-searching, continual-counting, punctual-pondering, punctual-reading.

The following modifiers are added to further describe the gestures:

- Reference zone (left upper corner, right upper corner, left lower corner, right lower corner, centre)

- Broad morphological form (e.g. "one hand finger pointing, "one hand circle", "two hands crossing")
- Stroke (beginning and end)
- Object (which object on the display were they pointing to)
- Reference word in the audio channel (e.g. "this", "here", "no")
- Comments

The head gestures are coded with regard to three broad morphological categories

- Head rotation
- Head incline forward/backward
- Head incline sideward

However, the researchers in the project say that they face some major problems with this coding system:

- How should they define a unit? - Beginning and end of a gesture cannot be perceived easily
- Meaning: morphological similar gestures can have different meanings
- Description: complex gestures can only be described roughly
- Reference: the reference word of the reference location cannot always be determined easily
- Categories: at the beginning of the project it is not known which categories will emerge as useful

The labelling system for facial expressions is also under development. It is planned to label emotional facial expressions in six categories:

- Anger/irritation
- Boredom/lack of interest
- Joy/gratification (being successful)
- Surprise/amazement
- Neutral/anything else
- Face partly not visible

The impressions will be rated as weak or strong. In future, the prosodic annotation of emotional speech described in Fischer (1998, 1999) will be adopted. The two main problems with respect to emotions are that first of all the judgement of emotions vary subjectively and secondly that for practical reasons it has not been possible to use a system like the "Facial Action Coding System" (FACS) by Paul Ekman, which allows objective categories. More information on FACS can be found in ISLE report 9.1, "Survey of Annotation Schemes and Identification of Best Practice", and in the following papers:

Fischer, Kerstin: Szenariodesign: Elizitieren von emotionalen Äußerungen. VERBMOBIL-MEMO 140. 1999

Fischer, Kerstin: Annotating Emotional Language Data. VERBMOBIL-REPORT 236. 1998.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The reviewers had access to examples from the resource and had contact to Silke Steininger, who is involved in the project. Silke Steininger kindly provided extensive help and information about the project and validated the final description.

Dr. Silke Steininger
E-Mail: kstein@phonetik.uni-muenchen.de
Institut für Phonetik und Sprachliche Kommunikation
Schellingstr. 3
80799 München

Tel.: +49 (0)89-2180 5751
Fax: +49 (0)89-2180 995751
Web site: http://www.phonetik.uni-muenchen.de/Mitarbeiter/steininger/Steininger.htm

## 2.4.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The purpose of the data resource has been to collect data for the training of speech, gesture and emotion recognisers, to develop dialogue and context models and to investigate how users interact with a computer that has far greater communication skills than humans are used to, in order to develop an intelligent human-computer interface that allows a computer novice to communicate naturally with a adaptive and self-explanatory computer system.

***Who used the resource so far/who are the target users of the resource?***

So far the data resource has been used only within the SmartKom project, but the SmartKom consortium takes into mind that the data is useful for other researchers too. The data is released to the public after one year because of competitive reasons.

***Is the resource language dependent or language independent?***

The dialogues are in German.

A translation into English of the gesture label files is being considered, but has not been decided upon yet.

A translation of the transcription files (without most of the markers for noise, prosody etc., i.e. only the spoken words) into English for the 90 cinema dialogues has started in May 2000, but it has not been decided whether this future translation should be made available to the public yet.

## 2.4.8 Conclusion

***How interesting/important/high quality is the resource?***

The resource was created to support the development of a NIMM system. A study of the included modalities may also provide useful information to other NIMM system developers. The resource includes speech, gestures, facial expressions, and upper body movements.

***What do the authors regret (if anything) not to have done while building the resource?***

A third camera with another view to the gestures used would have been useful but was not added because of economic reasons. Furthermore the SmartKom consortium regret not having asked for more money.

# 2.5 FaceWorks

## 2.5.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

May 16th, 2001

## 2.5.2 References

***Web site***

An overview and short description of FaceWorks:
http://www.interface.digital.com/overview/default.htm

***Short description***

FaceWorks is real-time 3D talking faces with synchronized speech.

DIGITAL FaceWorks™ is a piece of software that enables multi-media developers to create digital personalities.

***Illustrative sample picture or video file***

**Figure 2.5.1.** Examples of the different available facial models.

***References to additional information on the reviewed resource***

Fred I. Parke and Keith Waters: Computer Facial Animation. A. Peters Ltd., Boston Massachusetts, pp. 450, 1996. See: http://www.crl.research.digital.com/publications/books/waters/waters_book

Waters, K.: A Muscle Model for Animating Three-Dimensional Faces. From SIGGRAPH'87. Computer Graphics, Vol.21, No.4, pp. 17-24, July 1987.

K. Waters and T. Levergood: DECface: An automatic lip-synchronization algorithm for synthetic Faces. Digital Equipment Corp, Cambridge Research Laboratory, Technical Report Series 93/4, September 1993.

### 2.5.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

The files contain 3D synthetic faces with audio in the form of synchronized speech.

***How many humans are recorded at the same time?***

None.

***What is their profile?***

None.

***Which human body parts are visible in the resource?***

FaceWorks 3D talking faces represent face and shoulders.

***Which modalities are annotated?***

Both the visual representations and the audio have been annotated.

***Which other modalities are available/visible in the resource but have not been annotated ?***

Basic facial expressions like anger, fear, surprise, disgust, happiness and head nodding.

### 2.5.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 2.5.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The 3D facial models are defined by Facial Description (*.fd) files and the Facial Animation (*.fa) files created from *.wav files and jpeg images of faces.

***How much data does the resource contain?***

The resource contains representations of 5 different characters. However, any face can be constructed from a 2D jpeg image of the face.

***Who created the resource and when?***

The resource has been developed as part of Human-Computer Interaction research conducted at Compaq's Cambridge Research Lab under the supervision of Keith Waters.

***How was the resource created?***

The resource was created through the use of a synthetic agent with texture maps. The lip movements are computed by associating a lip shape to each phoneme and then by interpolating between successive shapes. Facial expressions are obtained by moving some key-points of the facial mask. It uses a pseudo-muscular approach that has been described in:

Waters, K.: A Muscle Model for Animating Three-Dimensional Faces. From SIGGRAPH'87. Computer Graphics, Vol.21, No.4, pp. 17-24, July 1987.

K. Waters and T. Levergood: DECface: An automatic lip-synchronization algorithm for synthetic Faces. Digital Equipment Corp, Cambridge Research Laboratory, Technical Report Series 93/4, September 1993.

***What is the application area?***

Multimedia development using ActiveX technology.

***What was the original purpose of creating the resource?***

The original purpose is to enable multimedia developers to create digital personalities.

## 2.5.6 Accessibility

***How does one get access to the resource?***

The resource can be downloaded from the web:

http://www.interface.digital.com/download/default.htm

One should make sure that one meets or exceeds the following system requirements before downloading FaceWorks Studio:

OS: Microsoft Windows 95, 98, or NT 4.0 with audio drivers.

Hardware: Intel Pentium PC (at least 133 MHz, with at least 16MB RAM) with audio card & speakers.

Once the installation file of FaceWork Studios has been downloaded, run it and follow the instructions. In case of problems please send an e-mail to the Compaq Corporate Research Downloads Team, downloads@crl.dec.com.

***Is the resource available for free or how much does it cost?***

The resource is available for free.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

When one runs FaceWorks a window appears which consists of several frames, cf. figure 1.5.2. The software offers several tools to create new characters with "personalities" and animation. The five faces from figure 1.5.1. are included as examples.

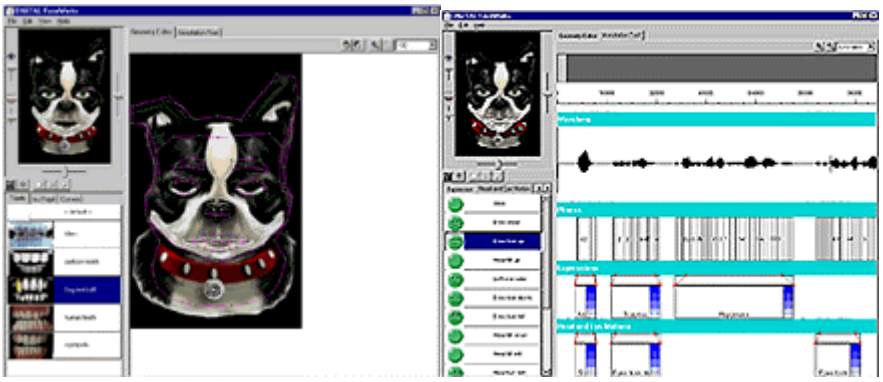

**Figure 2.5.2.** FaceWorks

One can easily transform any JPEG image into a 3D talking head. FaceWorks Studio can match voice audio tracks to the face automatically. Through the expression tools, one can add emotions and expressions. FaceWorks is easy to use and no intensive training is required. The program includes three key components:

*The Geometry Editor*

The geometry editor, cf. figure 1.5.3., takes any 2D image (JPEG) and maps it to a 3D face model. Key features, such as the eyes and the mouth, can be quickly located and changed. The geometry editor allows the user to choose from a pre-defined library with texture of the teeth, iris pupil, and cornea.

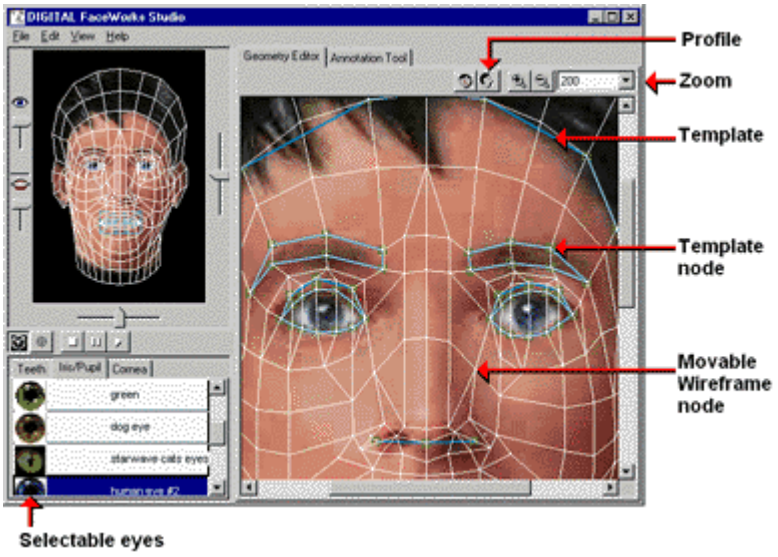


**Figure 2.5.3.** The FaceWorks Geometry Editor

*The Annotation Editor*

The annotation editor, cf. figure 1.5.4., allows users to record voices and analyse them. It automatically generates the linguistic information such as the list of phonemes and their duration, which is used to lip-sync a character's mouth to the voice. Users can then add head and eye motions, as well as basic facial expressions, to bring the character to life. The linguistic information as well as the expressions can be exported (or imported) in a file so that the user can save (respectively run) a defined animation. The possible facial expressions the user can choose from are: nod, anger, fear, shake, disgust, sadness, surprise and happiness. The head and eye motion library contains the following movements: stoic, eyes close, eyes look up/down/left/right, head tilt up/down/left/right,

right/left eyewink. The user needs to select one expression and place it along with the speech. At any time, the user can change the type of expression and its temporal value.

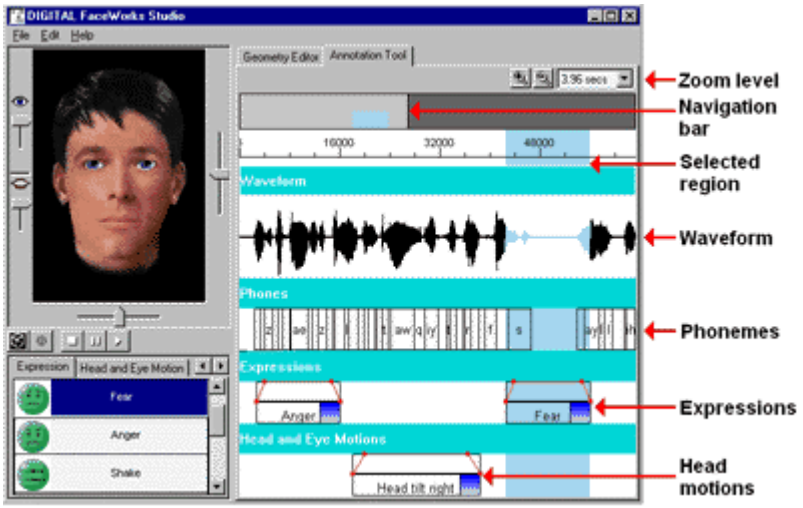


**Figure 2.5.4.** The FaceWorks Studio Annotation Editor.

*The Real-Time Face Display*

FaceWorks Studio has a real-time face display, cf. figure 1.5.5., allowing the user to observe editions as they happen. As nodes in the geometry editor are manipulated, the real-time face display displays the result. Likewise in the annotation editor, when a selection is played the face speaks in real-time.

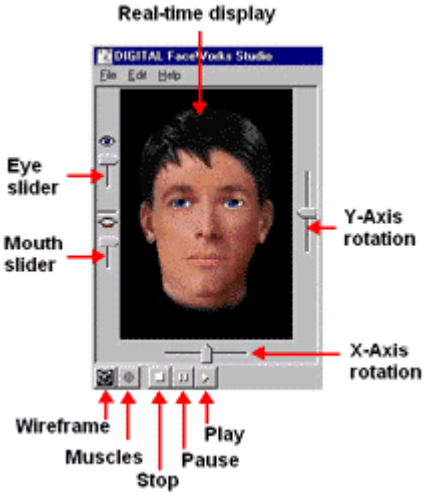


**Figure 2.5.5.** The Real-Time Face Display.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on hands on experience with the resource by the main actor and on web information. It is easy to use this software and create any animation. The lip movements were considered quite jerky. Animations are limited by the choice of expressions defined in the expression library. Moreover, one cannot superpose expressions on top of each other. Furthermore, the verifying actor had contact to the supervisor of the creation of the resource, Keith Waters (kwaters@mediaone.net), who kindly answered questions and validated the final description of the resource.

### 2.5.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource can be used for/has been used for the creation of synthetic characters on the web. The following are possible applications:

- Product and service information
- News updates and online columns
- Entertainment and comedy
- Education and training
- Electronic commerce

***Who used the resource so far/who are the target users of the resource?***

The target users are people who want to create and animate synthetic agents in real time.

***Is the resource language dependent or language independent?***

For the moment the faces can speak English, French, Spanish, Malay.

### 2.5.8 Conclusion

***How interesting/important/high quality is the resource?***

The resource is an animation system that is interactive, easy to use and public domain.

***What do the authors regret ( if anything) not to have done while building the resource?***

According to Keith Waters the creators of the resource regret nothing.

## 2.6 M2VTS Multimodal Face Database

### 2.6.1 Description header

***Main actor***

ELDA: Valerie Mapelli mapelli@elda.fr and Niklas Paulsson (paulsson@elda.fr)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

June 27th, 2001.

### 2.6.2 References

***Web site***

A short description of the project is on the UCL Laboratoire de télécommunication et télédétection web site: http://www.tele.ucl.ac.be/PROJECTS/M2VTS/

Download site for a report on the M2VTV face database in post script format: http://www.tele.ucl.ac.be/PROJECTS/M2VTS/m2vts_db.ps.Z

***Short description***

The database is intended for use within access control situation by using multimodal identification of human faces. The recognition efficiency is improved by combining single modalities. The modalities treated in this database are face and voice features with 185 pictures of 37 different faces.

***Illustrative sample picture or video file***

Both video and audio files can be downloaded from http://www.tele.ucl.ac.be/PROJECTS/M2VTS/test.html

***References to additional information on the reviewed resource***

Vassilios Chatzis Adrian: Multimodal Decision Level Fusion for Person Authentication. Can be downloaded from: http://citeseer.nj.nec.com/233665.html

C. Beumier, M. Acheroy: Final Report on the structured light modality. Deliverable 3.3.1 ``algorithm refinement''. Project M2VTS, Programme ACTS, July 98. Can be downloaded from: ftp://ftp.elec.rma.ac.be/user/beumier/PAPERS/deliv33.ps.gz

C. Beumier, M. Acheroy: Person Authentication with structured light. Deliverable 3.2.1 ``algorithm complexity evaluation and selection''. Project M2VTS, Programme ACTS, January 97. Can be downloaded from: ftp://ftp.elec.rma.ac.be/user/beumier/PAPERS/deliv32.ps.gz

C. Beumier, M. Acheroy. Person Authentication with structured light. Deliverable 3.1.1 ``multi-modal basic algorithm components", Project M2VTS, Programme ACTS, October 96. Can be downloaded from: ftp://ftp.elec.rma.ac.be/user/beumier/PAPERS/deliv31.ps.gz

Benoit Duc, Elizabeth Saers Bigun, Josef Bigun, Gilbert Maitre, and Stefan Fischer: Fusion of audio and video information for multi modal person authentication. In Pattern Recognition Letters, 18(9), pp. 835-843, 1997.

P. Jourlin, J. Luettin, D. Genoud, and H. Wassner: Acoustic-labial speaker verification. In Pattern Recognition Letters, 18(9), pp. 853-858, 1997.

Constantine Kotropoulos: Face Verification Based On Morphological Dynamic Link Architecture. Can be downloaded from: http://citeseer.nj.nec.com/104675.html

Constantine Kotropoulos: Linear Projection Algorithms And Morphological Dynamic Link Architecture For Frontal Face Verification. Can be downloaded from: http://citeseer.nj.nec.com/75876.html

Constantine Kotropoulos: Rule-Based Face Detection In Frontal Views. Can be downloaded from: http://citeseer.nj.nec.com/58561.html

C. Kotropoulos, and I. Pitas: Face authentication based on morphological grid matching. In Proceedings of the IEEE International Conference on Image Processing (ICIP 97), Vol. 1, pp. 105-108, Santa Barbara, California, U.S.A., 1997.

C. Kotropoulos, A. Tefas and Ioannis Pitas: Frontal Face Authentication Using Discriminating Grids with Morphological Feature Vectors. In IEEE Transactions on Multimedia, Vol. 2, no. 1, pp. 14-26, 2000. Can be downloaded from: http://citeseer.nj.nec.com/article/kotropoulos99frontal.html

C. Kotropoulos, A. Tefas and I. Pitas: Frontal face authentication using morphological elastic graph matching. In IEEE Transactions on Image Processing, 1999 Can be downloaded from: http://citeseer.nj.nec.com/kotropoulos99frontal.html

C. Kotropoulos, A. Tefas, and I. Pitas: Frontal face authentication using variants of Dynamic Link Matching based on mathematical morphology. In Proceedings of the IEEE International Conference on Image Processing, Vol. I, pp. 122-126, Chicago, U.S.A., October 1998.

C. Kotropoulos, A. Tefas and I. Pitas: Face Authentication Using Variants Of Elastic Graph Matching Based On Mathematical Morphology That Incorporate Local Discriminant Coefficients. Can be downloaded from: http://citeseer.nj.nec.com/20757.html

C. Kotropoulos, A. Tefas, I. Pitas, C. Fernandez, and F. Fernandez: Performance assessment of morphological dynamic link architecture under optimal and real operating conditions. In Proceedings from International Workshop on Nonlinear Signal and Image Processing, Antalya, Turkey, 1999.

Håkan Melin: Databases For Speaker Recognition: Activities. In Cost250 Working Group 2. Can be downloaded from: http://citeseer.nj.nec.com/345507.html

Athanasios Nikolaidis and Ioannis Pita: Robust Watermarking of Facial Images Based on Salient Geometric Pattern Matching. In IEEE Transactions on Multimedia, Vol. 2, no. 3, pp. 172-184, 2000. Can be downloaded from: http://citeseer.nj.nec.com/397808.html

S. Pigeon and L. Vandendorpe: The M2VTS multimodal face database. In J. Bigun, C. Chollet and G. Borgefors, Eds.: Lecture Notes by in Computer Science: Audio- and Video- based Biometric Person Authentication. 1206, pp. 403-409, 1997.

G. Richard et.al.: Multi Modal Verification for Teleservices and Security Applications (M2VTS). In Proceedings from IEEE International Conference on Multimedia Computing and Systems 1999, Vol. 1, Florence, Italy, pp. 1061-1064, 7-11 June 1999.

A. Tefas, C. Kotropoulos, and I. Pitas: Variants of dynamic link architecture based on mathematical morphology for frontal face authentication. In CVPR, 1998.

Anastasios Tefas, Constantine Kotropoulos and Ioannis Pitas: Enhancing The Performance Of Elastic Graph Matching For Face Authentication By Using Support Vector Machines. Can be downloaded from: http://citeseer.nj.nec.com/286314.html

A. Tefas, Y. Menguy, C. Kotropoulos, G. Richard, I. Pitas, P. Lockwood: Compensating For Variable Recording Conditions. In Frontal Face Authentication Algorithms A. Can be downloaded from: http://citeseer.nj.nec.com/137107.html

### 2.6.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

37 humans have been recorded in the whole resource.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

25 are male and 12 are female speakers. In some shots the subjects wear glasses.

***Which human body parts are visible in the resource?***

The face of each subject is visible in the resource.

***Which modalities are annotated?***

Some of the head movements of the subjects have been annotated.

***Which other modalities are available/visible in the resource but have not been annotated ?***

Speech, some head movements and facial expression are available as non-annotated modalities in the resource.

### 2.6.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None

### 2.6.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

Images are stored in CIF format with a resolution of 286x350 and 4:2:2 color components. The sound has been stored as raw data without header, sampled at 16 bits and a frequency of 48 kHz.

Sequences of images are stored in `.lcc` files, with 25 Hz frame frequency, using the following structure :

begin of file

Image 1 - luminance (286x350 bytes)

Image 1 - Cb (286x175 bytes)

Image 1 - Cr (286x175 bytes)
Image 2 - luminance (286x350 bytes)
Image 2 - Cb (286x175 bytes)
Image 2 - Cr (286x175 bytes)
Image 3 - ...
...
end of file

***How much data does the resource contain?***

The resource contains 10 Gb of images and sound.

***Who created the resource and when?***

The resource was created by the M2VTS Partners in December 1996. The M2VTS partners are:
Matra Communication, France
Banco Bilbao Vizcaya, Spain
Cerberus AG, Switzerland
Ecole Polytechnique Fédérale de Lausanne, Switzerland
Ibermatica SA, Spain
IMT Neuchâtel, Neuchatel
Institut Dalle molle d'Intelligence Artificielle Perceptive, Switzerland
Renaissance, Belgium
Université Catholique de Louvain, Belgium
Unidad Tecnica Auxiliar de la Policia, Spain
University of Surrey, United Kingdom
University of Thessaloniki - Aristotle, Greece
University of Carlos III, Spain

***How was the resource created?***

A Hi8 video camera (576x720, 50Hz-interlaced, 4:2:2) was chosen for the shooting and a D1 digital recorder for the recording and editing. These shots were taken at one week intervals or when drastic face changes occurred in the meantime. During each shot, people were asked to count from '0' to '9' in their native language (most of the people are French speaking), rotate the head from 0 to -90 degrees, again to 0, then to +90 and back to 0 degrees. Also, they were asked to rotate the head once again without glasses if they wore any. From the whole sequence, 3 parts have been extracted: the *voice* sequence, the *motion* sequence and the *glasses off* motion sequence (if any).

***What is the application area?***

The first sequence can be used for speech verification, 2D dynamic face verification (choosing the most appropriate picture out of the sequence) and for speech/lips correlation analysis. The other two sequences are meant for face recognition purposes only and provide information about the 3D face features thanks to the motion. They may be used to implement and compare other techniques like identification from 2D facial pictures, profile view or multiple views.

***What was the original purpose of creating the resource?***

The primary goal of the M2VTS project was to address the issue of secured access to local and centralised services in a multi-media environment. The main objective was to extend the scope of application of network-based services by adding novel and intelligent functionalities, enabled by automatic verification systems combining multimodal strategies (secured access based on speech, image and other information). The objectives were also to show that limitations of individual technologies (speech recognition, speaker verification...) can be overcome by relying on multi-modal decisions (combination or fusion of these technologies) and can find practical and important applications in the new emerging fields of advanced interfaces for tele-services. The main goals of the project were therefore :

- to implement and validate secured access schemes welded in existing voice-based services.
- to develop new security services exploiting emerging speech and image-based recognition technologies.
- to provide secured services on non secured networks (such as PSTN, ISDN, LAN).
- to develop new services for security applications (for alarm verification and access control).

## 2.6.6 Accessibility

***How does one get access to the resource?***

A copy of the M2VTS database can be requested through :

ELRA - Distribution Agency (ELDA)

55-57, rue Brillat-Savarin

75013 Paris

France

Tel: 0033-1-43.13.33.33 / Fax: 0033-1-43.13.33.30

***Is the resource available for free or how much does it cost?***

Distribution costs are fixed to 250 ECU for ELRA members, 500 ECU for non-members. An explanation of how to become an ELRA member and an application form can be found at: http://www.icp.grenet.fr/ELRA/org/appform.html

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Info files are provided for each shot and subject. They give useful information such as the date of the shot, sequences length, comments about the shot, etc. as given below.

```
*****************************************
* M2VTS / Multimodal Face Database ve 1.01 *
*****************************************

Shot Info

  Subject ID :       CD      (m)
  Shot Number :      03
  Date :             12/07/95
  Comments :         Takes his glasses off at the end of
                     the motion sequence
```

```
D1 tape references

  Tape no :     1
                        Start TC        Stop TC
  ------------------------------------------------
  voice sequence        00.24.40-13     00.24.47-20
  motion sequence       00.24.47-20     00.24.51-03
  motion sequence (#)   00.24.52-12     00.24.56-13
  ------------------------------------------------

Sequences Info
                        frames
  ----------------------------
  voice sequence        183
  motion sequence       84
  motion sequence (#)   102
  ----------------------------

Note :

     (#) Glasses off if any

     File name :     sp_03_m.lcc
                      |  | |  |
                      |  | |  |
      subject ID   <--/  | |  \--> Format
        shot ID      <--/  \--> file type

             type :  v(oice)
                     m(otion)
                     n(o glasses if any)
                     i(nfo)
             format : lcc (images)
                      raw (sound)
                      txt (text)
```

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The description is based on material from the web page of the M2VTS Database.

## 2.6.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for secured access to local and centralised services in a multimedia environment through user authentication.

***Who used the resource so far/who are the target users of the resource?***

The resource has been used by K. Jonsson (1), J. Matas (1,2) and J. Kittler (1) for the M2VTS Prototype System for Face Verification.

1.CVSSP, University of Surrey, Guildford GU2 7XH, Surrey, UK, web site: http://www.ee.surrey.ac.uk/CVSSP/
2.CMP, Czech Technical University, Prague, Czech Republic, web site: http://cmp.felk.cvut.cz/

The resource has also been used for user authentication, lip tracking and face recognition.

***Is the resource language dependent or language independent?***

The subjects are speaking in their native language, which is mainly French.

### 2.6.8 Conclusion

***How interesting/important/high quality is the resource?***

Good quality images. The database can be used for speech verification, 2D dynamic face verification (choosing the most appropriate picture out of the sequence), speech/lips correlation analysis and for face recognition purposes.

***What do the authors regret ( if anything) not to have done while building the resource?***

Some impairments can be noticed with respect to the theoretical case:

- some subjects did not rotate their head properly (horizontal translation of the head in the direction of the rotation, vertical tilt depending on the rotation angle, no full covering of the 180 frontal degrees...)
- some subjects might have had their mouth open during one rotation of the head, closed during the other, ending up on different shapes in the profile view
- some subjects closed their eyes while moving the head
- the direction of starting the rotation of the head was not fixed over the different shots
- some subjects were speaking very low, resulting in a poor sound SNR
- some subjects could not keep from smiling during the shot
- rotation speed could be highly variable between different shots, but also within the same shot
- reflections from eyes and glasses
- blurry images during fast head rotation, due to limited shutter speed

# 2.7 M2VTS Extended Multimodal Face Database – (XM2VTSDB)

## 2.7.1 Description header

***Main actor***

ELDA: Valerie Mapelli (mapelli@elda.fr) and Niklas Paulsson (paulsson@elda.fr)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

June 8th, 2001.

## 2.7.2 Reference

***Web site***

Main page of the data resource: http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/

***Short description***

The Extended M2VTS Multimodal Face Database is an extension of the M2VTS Multimodal Face Database and consists of more than 1,000 GBytes of digital video sequences, which makes it the biggest multimodal face database ever produced.

***Illustrative sample picture or video file***

Sample pictures are available at:
http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/order/datasets.html

**Figure 2.7.1.** Examples of the sample pictures that can be found at: http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/order/datasets.html

***References to additional information on the reviewed resource***

Results of the ICPR 2000 face authentication contest. Lists all identity verification results on XM2VTSDB frontal face data according to the Lausanne protocol published before June 2000. Can be downloaded from: http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/

Description of the content of the database. The protocol for evaluating verification algorithms on the database is presented:

http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/docs/messer-biosig98.ps.Z

Paper about the acquisition of the xm2fdb published at the BioSig98 conference in June 1998:

http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/docs/messer-biosig98.ps.Z

Instructions for CDS001 (same as for CDS002, 3, 5, 6, 8 and DVD001):

http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/docs/cds001.ps

User license agreements can be found at: http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/

## 2.7.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

295 volunteers from the University of Surrey were shot four times at approximately one month intervals. On each visit (session) two recordings (shots) were made. The first shot consisted of speech whilst the second consisted of rotating head movements.

***How many humans are recorded at the same time?***

One subject is visible in the same frame.

***What is their profile?***

The profile of the subjects is that they are adults of both sexes and of different ages.

***Which human body parts are visible in the resource?***

The face of each subject is visible in the resource.

***Which modalities are annotated?***

Speech and head rotation are the modalities that have been annotated.

***Which other modalities are available/visible in the resource but have not been annotated?***

None

### 2.7.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

A clip-on microphone is visible in the resource.

### 2.7.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The images in the resource are 720x576 pixels. They have been individually compressed using zip. They can be uncompressed using pkzip on a DOS/WIN95/WINNT machine and by using 'unzip' on a UNIX based machine. The images are stored in PPM (portable pixmap format).
The audio is stored in mono, 16BIT, 32 KHz, PCM wave files.
Video sequences of the database are being distributed in digital video encoded AVI file format. This format has the fixed compression ratio of 5:1 and stored at colour sampling resolution 4:2:0. The pixel resolution is 720x576 and it is full motion video (i.e. 25 frames per second). The audio track is 16-bit resolution and 32 KHz frequency stereo, but the database was recorded in mono.
A set of 293 3D VRML models and texture images are also being distributed. The VRML models correspond to each subject head and were acquired using a high precision stereo-based 3D camera.

***How much data does the resource contain?***

A set of images consists of one left-profile and one right profile image per person (295 * 2 * 4), per session, a total of 2,360 images. I.e. 295 sets * 2 images (left, right) * 4 sessions = 2,360 images.

Also available is a set of front profile images, 1 image per subject per session (295 * 4), 1,180 images in total.

One set of audio files consists of 3 sentences read twice per person per session (295 * 3 * 2 * 4), a total of 7080 files.

One set of 293 3D VRML models is also available.

***Who created the resource and when?***

The resource was created by the M2VTS partners in December 1996. The M2VTS partners are:

Matra Communication, France

Banco Bilbao Vizcaya, Spain

Cerberus AG, Switzerland

Ecole Polytechnique Fédérale de Lausanne, Switzerland

Ibermatica SA, Spain

IMT Neuchâtel, Neuchatel

Institut Dalle molle d'Intelligence Artificielle Perceptive, Switzerland

Renaissance, Belgium

Université Catholique de Louvain, Belgium

Unidad Tecnica Auxiliar de la Policia, Spain

University of Surrey, United Kingdom

University of Thessaloniki - Aristotle, Greece

University of Carlos III, Spain

***How was the resource created?***

The entire database was acquired using a Sony VX1000E digital cam-corder and DHR1000UX digital VCR. This captures video at a colour sampling resolution of 4:2:0 and audio at frequency 32kHz and sampling rate of 16 bits.

The subject, to whom a clip-on microphone had been attached, was asked to sit in a chair. He/she was then asked to read three sentences which were written on a board positioned just below the camera. The subjects were asked to read at their normal pace, to pause briefly at the end of each sentence and to read through the three sentences twice. The three sentences remained the same throughout all four recording sessions.

***What is the application area?***

Research.

***What was the original purpose of creating the resource?***

Reliable person identification methods already exist, e.g. finger print analysis and eye scans. However, for less complex application areas than high-security scenarios, the users finds it unacceptable to use such methods, thus requiring something less intrusive but still reliable. Personal identification systems based on analysis of speech, frontal or profile images of face are non-intrusive and therefore user-friendly. Multi-modal personal verification is one of the most promising approaches to user-friendly (hence acceptable) highly secure personal verification systems. The XM2VTS is a large multi-modal database, which will enable the research community to test their multi-modal face verification algorithms on a high-quality large dataset.

### 2.7.6 Accessibility

***How does one get access to the resource?***

One has to fill in a form of a user license agreement and either fax or e-mail it to:

Kieron Messer, Dept of Electronic Engineering, University of Surrey, Guildford, Surrey. GU2 5XH.

Fax number: +44 (0)1483 876031

Two signed copies of the user license agreement must be signed and sent by surface mail.

Payment can be done either by credit card or cheque.

***Is the resource available for free or how much does it cost?***

The cost varies according to which dataset one wishes to require. 14 CDs are available. Their prices vary according to the amount of information they contain. Academic institutions pay half price.

| *Dataset* | *Title* | *Media* | *Euros* |
|---|---|---|---|
| CDS001 | Frontal Image Set | 2 x CDROM | 171 (342 if non-academic inst) |
| CDS002 | Side Profile Image Set | 3 x CDROM | 256.5 (513 if non-academic inst) |
| CDS003 | Audio Data | 4 x CDROM | 342 (684 if non-academic inst) |
| CDS004 | Sets CDS000-003 | 9 x CDROM | 684 (1368 if non-academic inst) |
| CDS005 | 3D VRML Models | 1 x CDROM | 171 (342 if non-academic inst) |
| CDS006 | Frontal Image Set II | 2 x CDROM | 171 (342 if non-academic inst) |
| CDS007 | Sets CDS001 and CDS006 | 4 x CDROM | 342 (684 if non-academic inst) |
| CDS008 | Darkened Frontal View Images | 2 x CDROM | 171 (342 if non-academic inst) |
| DVD000 | Frontal Images on DVD-RAM | 1 x DVD-RAM | 171 (342 if non-academic inst) |
| DVD001 | Sentence 3 on DVD-RAM | 4 x 5.2 GByte DVD-RAM | 1026 (2052 if non-academic inst) |
| DVD002 | Head Rotation Shots | 12 x 5.2 GByte DVD-RAM | 1881 (3762 if non-academic inst) |
| DVD003a | Sentences 1,2, 4 and 5 - Client Set | 20 x 5.2 GByte DVD-RAM | 2565 (5130 if non-academic inst) |
| DVD003b | Sentences 1,2, 4 and 5 - Imposter and Test Set | 8 x 5.2 GByte DVD-RAM | 1539 (3078 if non-academic inst) |

**Table 1.6.2.** Overview of datasets, titles, media and price.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

No

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The description is based on material from the web page of XM2VTS.

## 2.7.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for personal identification systems based on analysis of speech, frontal or profile images of face. But the resource has also been used for lip tracking, eye coordinate determination, face and speech authentication.

***Who used the resource so far/who are the target users of the resource?***

The resource has so far been used by institutes working on personal identification systems.

***Is the resource language dependent or language independent?***

The speech dataset contains sentences read in English. All other datasets are language independent.

### 2.7.8 Conclusion

***How interesting/important/high quality is the resource?***

The resource contains high quality sets of data. It is an improvement of the M2TVS data resource as it contains speech sequences.

***What do the authors regret not to have done while building the resource?***

No information available.

# 2.8 Multi-talker database

## 2.8.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

June 27th, 2001.

## 2.8.2 References

***Web site***

A short description of the Development of a Facility for Simultaneous Recordings of Acoustic, Optical (3-D Motion and Video), and Physiological Speech Data:

http://www.hei.org/research/projects/comneur/speechdata.htm

A project summary of the KDI - Segmental and Prosodic Optical Phonetics for Human and Machine Speech Processing project:

http://nsf-workshop.engr.ucf.edu/papers/BERNSTEIN.asp

***Short description***

The Multi-talker database is a multidisciplinary, multilaboratory project, whose focus is optical and acoustic phonetic signals and their relationships to each other in speech production and perception.

The goals of this multidisciplinary project are to quantitatively characterize optical speech signals, examine how optical speech characteristics relate to acoustic and to physiologic speech characteristics, study several fundamental issues in human visual speech perception, and apply obtained knowledge to optical speech synthesis.

Across the entire project, the main questions are:

1. What speech information can perceivers get from seeing talkers?
2. How are optical and acoustic signals related to underlying speech articulations?
3. What are the perceptual and neuro-physiological bases for visual speech perception?
4. Can usefulness of this knowledge be demonstrated for developing synthesis of artificial talking faces?

***Illustrative sample picture or video file***

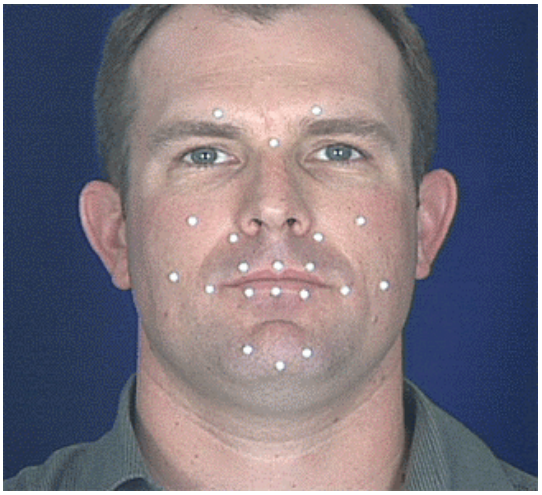

**Figure 2.8.1.** A face example, which shows the Qualisys™ Recording Positions. Copyright 2000, House Ear Institute.

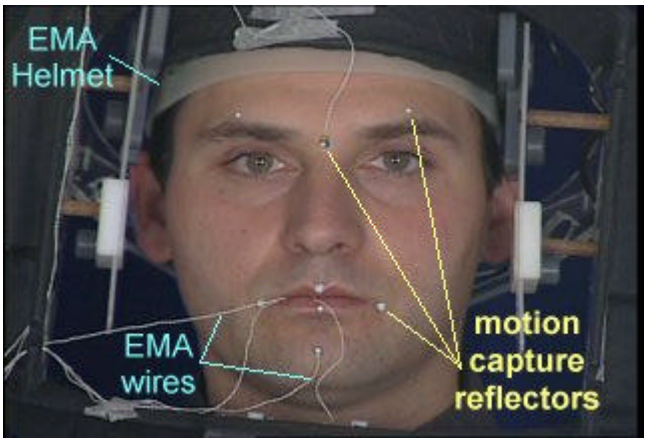


**Figure 2.8.2.** The subject's motions are captured with reflectors placed several places on the subject's face and an EMA helmet and EMA wires which make a computer image of the EMA recording positions, cf. figure 1.8.3. Copyright 2000, House Ear Institute.

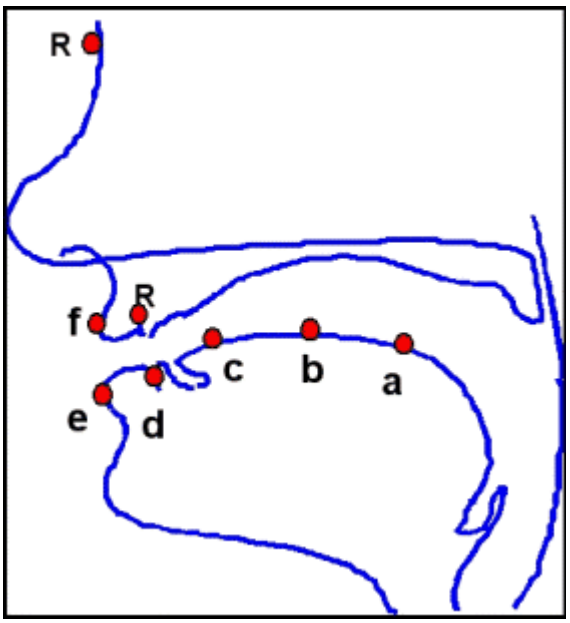


**Figure 2.8.3.** EMA recording positions.

***References to additional information on the reviewed resource***

Bernstein, L. E.: Segmental optical phonetics for human and machine speech processing. ICSLP2000, International Congress on Spoken Language Processing. Beijing, China, 16-20 October. 2000.

Bernstein, L. E., Auer, E. T., Chaney, B., Alwan, A., & Keating, P.: Development of a facility for simultaneous recordings of acoustic, optical (3-D motion and video), and physiological speech data. Journal of the Acoustical Society of America, 107, 2887. 2000.

Bernstein, L. E., Jiang, J., Alwan, A., & Auer, E. T., Jr.: Visual phonetic perception and optical phonetics. AVSP2001, Aalborg, Denmark, 7-9 September 2001.

Bernstein, L. E., Ponton, C., & Auer, E. T., Jr.: Electrophysiology of unimodal and audiovisual speech perception. AVSP2001, Aalborg, Denmark, 7-9 September 2001.

Bernstein, L. E., Ponton, C., Auer, E. T.: Audiovisual speech integration. XVIIth Biennial Symposium of the International Evoked Response Audiometry Study Group. Vancouver, British Columbia. 2001.

Bernstein, L. E., Ponton, C., Auer, E. T.: Is audiovisual speech integration an early perceptual effect? An event-related potential study of the McGurk effect. Cognitive Neuroscience Society, New York City, March 25-27, 2001.

Jiang, J., Alwan, A., Auer, E. T., & Bernstein, L. E.: Predicting visual consonant perception from physical measures. Eurospeech 2001. Aalborg, Denmark, September 3-6. 2001.

Jiang, J., Alwan, A., Bernstein, L. E., & Keating, P.: On the correlation between orofacial movements, tongue movements and speech acoustics. Journal of the Acoustical Society of America, 107, 2904. 2001.

Jiang, J., Alwan, A., Bernstein, L. E., Keating, P., & Auer, E. T.: On the correlation between facial movements, tongue movements, and speech acoustics. ICSLP2000, International Congress on Spoken Language Processing. Beijing, China, 16-20 October. 2000.

Jiang, J., Alwan, A., Keating, P., Bernstein, L. E., & Auer, E. T.: On the correlation between articulatory and acoustic data. Journal of the Acoustical Society of America, 108, 2508. 2001.

Keating, P. A., Cho, T., Mattys, S., Bernstein, L. E., Chaney, B., Baroni, M., Alwan, A.: Articulation of word and sentence stress. Journal of the Acoustical Society of America, 108, 2466. 2000.

### 2.8.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

Eight humans have been recorded in the resource.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

All of the talkers are young adults. Half are male. All were obtained through advertisements through the University of Southern California.

***Which human body parts are visible in the resource?***

The face and the midsagital tongue motion of the humans are available in the resource.

***Which modalities are annotated?***

The annotated modalities are the EMA (electromagnetic midsagital articulography) recording position and the 3D position of each retroreflector on speaker's face.

***Which other modalities are available/visible in the resource but have not been annotated ?***

None.

### 2.8.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 2.8.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The format of the resource is

- Raw video (BETACAM sp)
- 16-BIT audio (.raw audio file)
- EMA ASCII position files
- Video edit take lists (SMPTE time code with associated transcribed utterances)

The procedure of obtaining the data has been the following:
Post-Capture session data processing

- Raw video is copied to dubs.

- Video in and out point time code is identified for each utterance and a take-list is compiled.
- The 3D markers are tracked across frames and labelled.
- The 3D data can be stabilised for head motion.
- The EMA data are filtered and rotated into the occlusal plane.
- A master listing is generated to provide the synchronization information across files from different types of recordings, cf. figure 1.8.4.

| TAKE NUMBER | SMPTE TIME CODE | AUDIO FILENAME | QUALISYS FILENAME | QUALISYS OFFSET | EMA FILENAME | EMA OFFSET |
|---|---|---|---|---|---|---|
| 1 | 00:00:001:30 | XXX1.raw | XXX1.tsv | 3 | artmesXX1.txt | 15 |
| 2 | ... | XXX2.raw | XXX2.tsv | 2 | armesXX2.txtt | |
| ... | ... | ... | ... | ... | ... | ... |
| n | 00:02:59:59 | XXXn.raw | XXXn.tsv | 5 | artmesXXn.txt | 16 |

**Figure 2.8.4.** Example of master listing.

Part of the resource is available as a database.

***How much data does the resource contain?***

- Eight talkers screened to vary in visual intelligibility.
- 320 sentences recorded by each speaker.
    - IEEE/Harvard Sentences Selected for:
        - Similar mean word frequency of content words.
        - Current vocabulary/sense.
    - Audio, video, and 3-D motion recordings.
- CV nonsense syllables recorded by four speakers.
    - Audio, video, EMA, and 3D motion.
- Prosody Corpus recorded by three speakers.
    - Prosodic phenomena: lexical stress, weak syllables, word boundary, phrase break, and phrase stress.
    - Audio, video, EMA, and 3D motion recordings.

***Who created the resource and when?***

The resource was created by the House Ear Institute:
L.E. Bernstein, E.T. Auer, B. Chaney, House Ear Institute Los Angeles
A. Alwan and P.A. Keating, University of California, Los Angeles
The database was copyrighted in 1998.

***How was the resource created?***

The capture system and database are described below:

| DAT Recorder | VTR | Qualisys™ Motion Capture | Carstens Articulograph (Elecromagnetic Midsagital Articulography- EMA) |
|---|---|---|---|
| • HHB Portadat PDR1000<br>• Sennheiser ™ microphone MKH416P48U | • Sony™ UVW-1800 Recorder<br>• Sony™ DXC-D30 Digital Video Camera | • Cameras: MCU120/240Hz CCD Imager<br>• Passive retroreflectors<br>• Infrared flash<br>• www.qualisys.com | • Medizinelektronik AG100<br>• 10sensor recording<br>• Midsaggital plane<br>• www.articulograph.de |

**Figure 2.8.5.** Description of the equipment used for capturing the data.

***What is the application area?***

The application is primarily for research at this time. However, the goal is to apply the knowledge to communication problems of individuals with hearing impairment.

***What was the original purpose of creating the resource?***

The goals of the creation of this data resource are to quantitatively characterize optical speech signals, examine how optical phonetic characteristics relate to acoustic and to physiologic speech production characteristics, study what affects the intelligibility of optical speech signals, and apply obtained knowledge to optical speech synthesis and automatic speech recognition.

## 2.8.6 Accessibility

***How does one get access to the resource?***

One should contact:

L.E. Bernstein

House Ear Institute

2100 W. Third Street

Los Angeles, CA 90057

lbernstein@mailhouse.hei.org

regarding copies of the resource either as videodiscs or database.

Other data have been collected and archived in the following database:

Bernstein, L. E., & Eberhardt, S. P.: Johns Hopkins Lipreading Corpus I-II: Disc 1. Baltimore, MD: Johns Hopkins University, 1986.

Bernstein, L. E., & Eberhardt, S. P.: Johns Hopkins Lipreading Corpus III-IV: Disc 2. Baltimore MD: Johns Hopkins University, 1986.

Bernstein, L. E.: Lipreading Corpus V-VI: Disc 3. Gallaudet University, Washington, D.C., 1991.

Bernstein, L. E.: Lipreading Corpus VII-VIII: Disc 4. Gallaudet University, Washington, D.C., 1991.

Bernstein, L. E., Seitz, P. F., & Auer, E. T., Jr.: Lipreading Corpus IX-X: Disc 5. Initial Consonant Stimuli - Two Talkers. Gallaudet University, Washington, D.C., 1995.

Bernstein, L. E., Auer, E. T., Jr., & Seitz, P. F.: Lipreading Corpus XI-XII: Disc 6. Vowel Stimuli - Two Talkers. Los Angeles, CA: House Ear Institute, 1996.

Bernstein, L. E., Auer, E. T., Jr., & Seitz, P. F.: Lipreading Corpus XIII-XIV: Disc 7. Medial Consonant Stimuli - Two Talkers. Los Angeles, CA: House Ear Institute, 1996.

Bernstein, L. E., Auer, E. T., Jr., & Seitz, P. F.: Lipreading Corpus XV-XVI: Disc 8. Final Consonant Stimuli - Two Talkers. Los Angeles, CA: House Ear Institute, 1996.

Database:

Seitz, P. F., Bernstein, L. E., Auer, E. T., Jr., & MacEachern, M.: PhLex (Phonologically Transformable Lexicon): A 35,000-word computer readable pronouncing American English lexicon on structural principles, with accompanying phonological transformations, and word frequencies, 1998.

***Is the resource available for free or how much does it cost?***

Arrangements can be made to obtain database materials. For more information please contact L.E. Bernstein using the above address.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added to the original resource in the form of annotation of EMA recording position and 3D position of each retroreflector on speaker's face. Also, extensive perceptual data are being obtained on these recordings.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

Contact to Lynne E. Bernstein, lbernstein@mailhouse.hei.org was made, who kindly answered questions and validated the final description.

## 2.8.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource may be used to quantitatively characterize optical speech signals, examine how optical phonetic characteristics relate to acoustic and to physiologic speech production characteristics, study what affects the intelligibility of optical speech signals, and apply obtained knowledge to optical speech synthesis and automatic speech recognition.

One can also use the data to drive 3D talking faces.

***Who used the resource so far/who are the target users of the resource?***

The House Ear Institute and researchers at UCLA (Keating and Alwan) have used the resource for internal research projects.

***Is the resource language dependent or language independent?***

The acoustic signals are in English.

### 2.8.8 Conclusion

***How interesting/important/high quality is the resource?***

The resource is created especially for research of labial movement and is of very high quality.

***What do the authors regret ( if anything) not to have done while building the resource?***

No information available.

# 2.9 NITE (Natural Interactivity Tools Engineering) Floor Plan Corpus

## 2.9.1 Description header

***Main actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Verifying actor)***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Date of last modification of the description***

July 10th, 2001

## 2.9.2 References

***Web site(s)***

The NITE Project's own web site: http://nite.nis.sdu.dk/

***Short description***

The resource is a recording of a natural interaction between two subjects discussing a floor plan. The recording is divided into a three-screen view where everything from the waist and upwards is visible in the top frame, and a close up of the faces of the subjects is visible in the two lower frames, cf. figure 1.9.1. Therefore speech, gestures, arm gestures, body movements, facial expressions, lip movements and eye/gaze behaviour are available as modalities in the resource.

***Illustrative sample picture or video file***

**Figure 2.9.1.** Still picture taken from the video, showing the division into three views; one of both subjects and the two close-ups of their faces.

***References to additional information on the reviewed resource***

No additional information on the resource is available yet.

## 2.9.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

Two different humans have been recorded in the whole resource.

***How many humans are recorded at the same time?***

Two humans are visible in the top frame at the same time, and only one human is visible at the same time in the lower frames.

***What is their profile?***

The subjects are university professors, one male and one female.

***Which human body parts are visible in the resource?***

In the top frame everything from the waist and upwards is visible. In the two lower frames the faces of the subjects are visible.

***Which modalities are annotated?***

None, at the moment. In the future a crossmodality annotation of speech at multiple levels, gesture and facial expression will be made. The annotation of these modalities should be available in October 2001.

***Which other modalities are available/visible in the resource but have not been annotated?***

Speech, hand gestures, arm gestures, body movements, facial expressions, lip movements and eye/gaze behaviour are available as modalities in the resource.

### 2.9.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 2.9.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The resource includes one QuickTime file, one Mpeg-1 file and one .wav sound file. The files are not organised in a database structure.

***How much data does the resource contain?***

The QuickTime file contains 361 MB, the .mpeg file 185 MB and the .wav sound file 53 MB.

***Who created the resource and when?***

The resource was created and copyrighted by Natural Interactivity Systems Laboratory (NISLab), University of Southern Denmark, Main Campus: Odense University, Forskerparken 10, 5230 Odense M, Denmark on the 4th of April 2001.

***How was the resource created?***

The resource was created using three digital cameras; 1 for the view of both subjects and 1 for each of the close-up views of the faces of the subjects. The camera microphones plus an extra directional microphone placed on one of the cameras were used for the sound recordings. After the recording the videos were edited into one video using an analogue procedure to make the three-screen division, cf. figure 1.9.1. After this procedure the video was converted into the three formats described above.

***What is the application area?***

Research on natural interactivity.

***What was the original purpose of creating the resource?***

To create a test resource for cross level cross modality analysis of natural interactivity communication.

### 2.9.6 Accessibility

***How does one get access to the resource?***

The resource can be downloaded from the NITE project's web site: http://nite.nis.sdu.dk/. Or it can be obtained by requesting of copy of it on cd-rom from Niels Ole Bernsen (nob@nis.sdu.dk).

***Is the resource available for free or how much does it cost?***

The resource is available for free. If one requests a cd-rom one will have to pay for the price of the cd and the shipment of it.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Transcriptions and annotations of the speech and modalities of the resource will be available in 2002. For the audio transcriptions the tool Transcriber will be used and for the annotation work probably the tools ANVIL and Noldus Observer will be used.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review has been written by the creators of the resource.

## 2.9.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been created for use in the NITE project. It has been created for research on natural interactivity.

***Who used the resource so far/who are the target users of the resource?***

So far the partners in the NITE project has used the resource. For more information on NITE partners see: http://nite.nis.sdu.dk.

The target users are researchers in natural interactivity.

***Is the resource language dependent or language independent?***

The speech in the resource is in English.

## 2.9.8 Conclusion

***How interesting/important/high quality is the resource?***

The resource enables e.g. the study of the interrelationships among several different modalities in that it includes speech, gestures, arm gestures, body movements, facial expressions, lip movements and eye/gaze behaviour. The topic discussed in the resource is a floor plan for a new university building.

***What do the authors regret (if anything) not to have done while building the resource?***

The creators of the resource regret that they did not use a fourth camera, which filmed the interaction from above. This would have made a view of the sheets that the subjects are pointing to available, giving the future user a clear indication of what the pointing gestures refer to. The creators make evaluations of the resource as the work with it progresses in order to find ways to make even better resources as the NITE project continues.

# 2.10 Scan MMC (Score Analysed MultiModal Communication)

## 2.10.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

May 17th, 2001.

## 2.10.2 References

***Web site***

None.

***Short description***

The resource consists of 10 videotapes of 60 minutes each, where some fragments have been annotated. According to which part of the resource one is interesting in, everything from the whole body to just parts of the body, e.g., the face, are visible. The target users are scholars in Multimodal Communication.

***Illustrative sample picture or video file***

**Figure 2.10.1.** Example from the natural face-to-face interaction part of the resource.

***References to additional information on the reviewed resource***

I. Poggi and E. Magno Caldognetto: Mani che parlano. Gesti e psicologia della comunicazione. Padova: Unipress, 1997.

I. Poggi and C. Pelachaud: The meanings of gaze in animated faces. In P.McKevitt, S.Nuàllain and C.Mulvihill (Eds.) Language, vision and music. Amsterdam: John Benjamins, in press.

E. Magno Caldognetto and I. Poggi: The score of multimodal communication and the goals of political discourse. Quaderni dell'Istituto di Fonetica e Dialettologia del CNR, Padova, 1999.

## 2.10.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

No information available.

***How many humans are recorded at the same time?***

No information available.

***What is their profile?***

No information available.

***Which human body parts are visible in the resource?***

According to which part of the resource one is interesting in, everything from the whole body to just parts of the body, e.g., the face are visible.

***Which modalities are annotated?***

The annotated modalities are facial expression, head movement, gaze, gestures and body movement.

***Which other modalities are available/visible in the resource but have not been annotated ?***

The available modalities are audio and video. The video recordings consist of films, TV talk shows and natural face-to-face interactions, and the audio of the relating sound.

## 2.10.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

## 2.10.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The format of the resource is VHS videotapes. The VHS videotapes are in the process of being transferred to digitalized tapes.

***How much data does the resource contain?***

The resource consists of 10 videotapes of 60 minutes each, where some fragments have been annotated.

***Who created the resource and when?***

The resource was created by the main creator, I. Poggi and her students at courses in "General Psychology" and "Psychology of Communication" from 1994 to 2000 at the University of Padova.

***How was the resource created?***

The resource was created by recording films, TV talk shows and natural face-to-face interactions.

***What is the application area?***

Research

***What was the original purpose of creating the resource?***

The purpose was to do research on facial expressions and gestures.

## 2.10.6 Accessibility

***How does one get access to the resource?***

The procedure is to e-mail Isabella Poggi, poggi@uniroma3.it, and ask her for access to the resource.

***Is the resource available for free or how much does it cost?***

The resource will be for free. But the resource will not be available before all the VHS tapes have been transferred to digital tapes. The main creator cannot give a date of when resource will be available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added to the original resource in the form that some aspects of facial and gesture behaviour have been analysed according to coding scheme named The Musical Score (see D.9.1 for a description of the coding scheme).

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review was mainly written by the main creator of the resource.

## 2.10.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has and can be used for research on facial expressions and gesture behaviour.

***Who used the resource so far/who are the target users of the resource?***

So far only the creators of it have used the resource. The target users are scholars in Multimodal Communication.

***Is the resource language dependent or language independent?***

The speech in the resource is in Italian.

## 2.10.8 Conclusion

***How interesting/important/high quality is the resource?***

The resource is the only resource found, where the language is Italian.

***What do the authors regret ( if anything) not to have done while building the resource?***

No information available.

## 2.11 VIDAS (VIDeo ASsisted with audio coding and representation)

An ACTS project, 1995-1999.

### 2.11.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

May 14th, 2001.

### 2.11.2 References

***Web site(s)***

The Cordis record of the project: http://dbs.cordis.lu/cordis-cgi/srchidadb?ACTION=D&SESSION=45252001-5-14&DOC=1&TBL=EN_PROJ&RCN=EP_RCN:30501&CALLER=EISIMPLE_EN_PROJ

A description of the project on the Common European Newsletter, Multimedia Content Manipulation and Management web site: http://www.esat.kuleuven.ac.be/~konijn/vidas.html

A description of the project on the Digital Signal Processing Laboratory at University of Genova web site: http://www-dsp.com.dist.unige.it/projects/vidas.html

A description of the project on the MIRALab / University of Geneva web site: http://www.miralab.unige.ch/Vidas.html

The projects own web site: http://www.infowin.org/ACTS/RUS/PROJECTS/ac057.htm

***Short description***

"VIDAS" worked toward the objective of devising suitable methodologies and algorithms for time-correlated representation, coding and manipulation of digital A/V bit streams. In particular, the specific A/V content which VIDAS considered was the typical videophone scene where audio and video components consist of the talker's speech and face.

***Illustrative sample picture or video file***

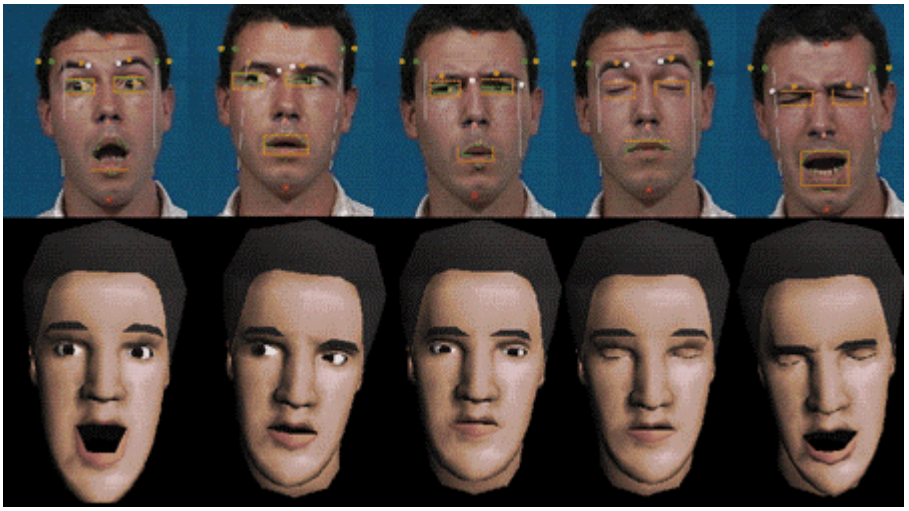

**Figure 2.11.1.** The top row shows faces taken from the corpus, and the below row the synthetic representation of the same.

***References to additional information on the reviewed resource***

None.

## 2.11.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

The data resource contains a synchronised audio-video corpus of 10 speakers, composed of recordings of single utterances of 700 English words. The resource does not contain the corresponding synthetic faces to all of the speakers, but just the ones found in the illustrative example above, cf. figure 1.11.1.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

The data resource contains recordings of both women and men. No other information is given on their profile.

***Which human body parts are visible in the resource?***

Only the face of the subjects is visible in the resource.

***Which modalities are annotated?***

The audio, the facial expressions and the lip shape have been annotated. In hybrid coding, the image of the speaker's face is analysed, features are tracked and parameterised, and the parameters transferred over the network to the facial model that is used to synthesize the face with the appropriate expressions, cf. figure 1.11.1. The rest of the image (background) is coded using more classical

region-based techniques. The project has worked towards an improvement of the MPEG-4 coding scheme. Cf. report 9.1 in the ISLE project, "Survey of Annotation Schemes and Identification of Best Practice"

***Which other modalities are available/visible in the resource but have not been annotated ?***

None.

## 2.11.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

## 2.11.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The format of the resource is video recordings, but the exact format of these has not been specified in the available information.

***How much data does the resource contain?***

No information available.

***Who created the resource and when?***

The participants of the ACTS project VIDAS created the resource from 1995-1999. The participants are:

DIST - University of Genoa

Aristotle University of Thessalonica

Ecole Politechnique de Lausanne

IRISA

Matra Communication

Modis S.p.A.

NAD - National Association for the Deaf

University of Geneva

University of Linköping

UPC - Universitat Politecnica Catalunya

***How was the resource created?***

No information available.

***What is the application area?***

No information available.

***What was the original purpose of creating the resource?***

"VIDAS" is working towards the objective of devising suitable methodologies and algorithms for time-correlated representation, coding and manipulation of digital A/V bit streams. In particular, the specific A/V content which VIDAS is considering is the typical videophone scene where audio and video components basically consist of the talker's speech and face, respectively.

The bimodal acoustic/visual nature of speech, both in production and in perception, suggests a strong correlation between the acoustic content of speech and the corresponding coherent movements of the talker's lips: the exploitation of this correlation both in the analysis (encoder) and synthesis (decoder) represents the eminent objective of VIDAS.

Two different applications are addressed: a short-term integration of VIDAS algorithms in H.324 videophone and a medium-term development of a hybrid coding scheme based on 3D A/V modelling techniques.

The improvements carried by VIDAS speech/image tools to H.324 will concern the quality of the decoded video at the receiving terminal and will be achieved through post-processing procedures in charge of smoothing out the movements of the talker's lips in synchronization with speech. This operation will lead to the generation of synthetic frames which will be interleaved with the real ones at presentation, thus simulating a frame rate up conversion. The quality of the synthesized video will be accessed by means of suitable evaluation tests carried out either with hearing impaired subjects, trained in speech reading, and with normal hearing subjects. The improved H.324 scheme will be first implemented in software and then integrated in real-time hardware.

The second objective of VIDAS consists of the implementation of a 3D virtual environment for the efficient representation of videophone scenes. Specific applications are in the field of advanced interpersonal communication services. This environment will allow efficient handling of acoustic-visual objects, either natural or synthetic, and their suitable composition, rendering and presentation. In accordance to the on-going discussion in the MPEG4 "ad hoc" group SNHC (Synthetic/Natural Hybrid Coding), VIDAS' virtual environment is also targeted to the implementation of functionalities like object geometry and texture scalability, object interactive manipulation and multimodal integration.

VIDAS' virtual environment includes two main objects being the talking actor and the background, each of them optionally natural, synthetic or hybrid. Multiple scenarios can be figured out depending on the particular object representation which is adopted and which, in principle, can be dynamically changed by interactive modification of the communication profile.

VIDAS' virtual environment will employ a flexible syntax so that the audio-video representation will make it possible to manipulate the multimedia content either on-line or off-line. This will allow not only the virtual representation of a natural audio-video scene but also the synthetic generation of virtual scenes with no relation at all to reality. In this case it would be possible to fuse multiple audio-video representations associated to natural acoustic-visual sources into new artificial representations.

It has not been possible to determine how many of these objectives the project has finished at the time of writing.

## 2.11.6 Accessibility

***How does one get access to the resource?***

No information is available but one can contact the prime participant of the project:

Fabio Lavagetto
DIST University of Genova
Via Opera Pia 13
16145 Genova
Italy

Tel: +39 010 353 2208
Fax: +39 010 353 2948

E-mail: Fabio@dist.dist.unige.it

An attempt of creating contact to Fabio Lavagetto has been made, but no response was returned. Therefore several questions remain unanswered.

***Is the resource available for free or how much does it cost?***

This has not been possible to find out.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added to the original resource in the form of an elaboration of:

- MPEG-4 facial model parameterisation
- Lip extraction and tracking
- Speech analysis and articulation estimation
- Audio assisted frame interpolation for increasing the frame frequency
- Segmentation of the scene into a component that can be modulated (the speaker's face) and another that cannot be modulated (the background)
- The region of the speaker's face is encoded through model-based algorithms assisted by speech analysis
- The background is encoded through region-based algorithms

Segmenting the speaker's face region from input images, extracting and tracking the speaker's facial parameters to be suitable and to produce realistic synthesis, either based on simple 2D meshes, or on complex deformable 3D models.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

It has not been possible to establish contact to the creators of the resource, nor to get access to it. Therefore, this description is made on the basis of web information.

## 2.11.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The main objective is to devise suitable methodologies and algorithms for time-correlated representation, coding and manipulation of digital A/V bit streams.

***Who used the resource so far/who are the target users of the resource?***

It is assumed that the participants in the project have used the resource, but since no contact has been established no further comments can be made at this point.

The number of users who could benefit from the project's results ranges from the normal consumer to the pathological hearing impaired. The goals of the project are oriented to the general improvement of the visual subjective quality of the images in a narrow-band videophone. Everyone will benefit from this improvement since the images will look more natural. Moreover, in case of hearing impairments, this benefit will be dramatic. In this case the videophone will not be a "useless" advanced telephone, but will become the privileged communication means. In some cases, rehabilitation to lip-reading

could even be done through remote teaching via videophone. In-between these two extremes, being the normal hearing user and the deaf, a large variety of intermediate possible consumers can be mentioned and, first of all, elderly people who could benefit much from the improvements on video-phone achieved by the project activity.

***Is the resource language dependent or language independent?***

The utterances of the recordings are in English.

## 2.11.8 Conclusion

***How interesting/important/high quality is the resource?***

This database allows for bimodal multi-speaker speech processing. The database was developed within the EU project VIDAS that has an active part in the MPEG-4 profiles definition for Face and Body Animation.

***What do the authors regret ( if anything) not to have done while building the resource?***

No information available.

# 2.12 /'VCV/ database

## 2.12.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

June 11th, 2001.

## 2.12.2 References

***Web site***

Information on the Elite Motion Analyser Technology can be found at:

http://www.bts.it/bts/products/eliteplu.htm

***Short description***

This database provides articulatory information for the production of /'VCV/ in Italian, (C being one of the 21 Italian consonants and V = /a, i, u/).

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Information on the Elite Motion Analyser Technology can be found in:

G. Ferrigno, A. Pedotti: ELITE: a digital dedicated hardware system for movement analysis via real-time TV signal processing. IEEE Trans. Biomed. Eng., BME 32, pp. 943-950, 1985.

G. Ferrigno, N.A. Borghese, A. Pedotti: Pattern recognition in 3D automatic human motion analysis. ISPRS Journal of Photogram, Remote Sensing, 45, pp. 227-246, 1990.

Information on the elaboration of the data resource can be found in the following papers:

Cosi, P., Dugatto M., Ferrero F., Magno Caldognetto E., Vagges K.: Bimodal recognition of Italian plosives. In Proc. of 13th International Congress of Phonetic Sciences, (Stockholm August 1995), Vol. 4, pp. 260-263, 1995.

Cosi P., Magno Caldognetto E.: Lips and jaw movements for vowels and consonants: Spatio-temporal characteristics and bimodal recognition application, in Stork D.G. and Hennecke M.E. (Eds.):

Speechreading by Humans and Machines: Models, Systems and Applications. NATO ASI Series, pp. 291-313, 1996.

Magno Caldognetto E., Vagges K., Zmarich C.: Visible articulatory characteristics of the Italian stressed and unstressed vowels. Proceedings of XVIII International Congress of Phonetic Sciences, (Stockholm, 13-19 August 1995), Vol. 1, pp. 366-369, 1995.

Magno Caldognetto E., Zmarich C., Cosi P., Ferrrero F.E.: Italian consonantal visemes: Relationships between spatial/temporal articulatory characteristics and co-produced acoustic signal. In Proceedings of the Workshop on Audio-Visual Speech Processing: Cognitive and Computational Approaches, Rhodes (Greece), pp. 5-8, 1997.

Magno Caldognetto E., Zmarich C., Cosi P.: Statistical definition of visual information for Italian vowels and consonants. In Proceedings of AVSP'98, International Conference on Audio-Visual Speech Processing, Terrigal –Sydney, pp. 135-140, 1998.

## 2.12.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

Four speakers have been recorded in the resource.

***How many humans are recorded at the same time?***

Only one human is recorded in the same frame.

***What is their profile?***

They are two female and two male university students, aged between 18 and 22, talkers of northern Italian.

***Which human body parts are visible in the resource?***

Only the markers attached to the face of the subjects are visible in the resource.

***Which modalities are annotated?***

The lip movements have been annotated. The markers have been applied on the subject's face on the: mid point of the upper lip, mid point of the lower lip, both lip corners, central point of chin. Other markers on the tip of the nose and on the ear lobes are used as reference points for the analysis program. The displacement of the markers in 3D as the subject pronounces sequences of the type /'VCV/ is recorded using the Elite system. The displacement of markers corresponds to kinematical curves. Adequate programs have been developed to compute parameters that have relevance in phonetics and phonology. Such parameters are lip height, lip width, protrusion of upper and lower lip.

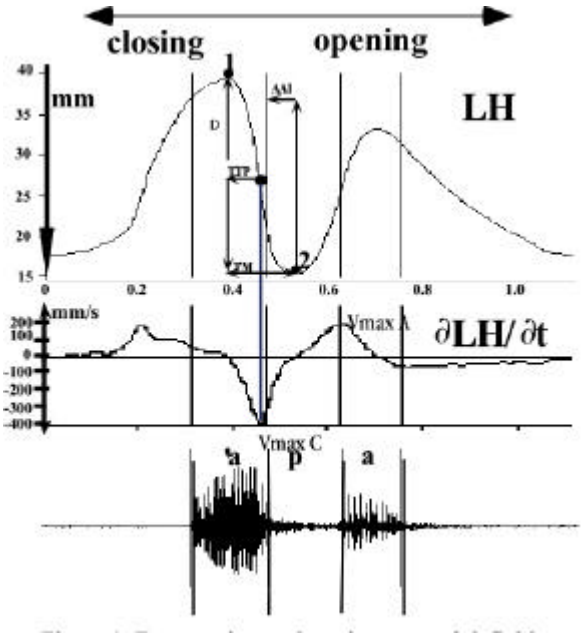


**Figure 2.12.1.** This image shows the several pieces of information one gets from Elite while a speaker has produced /'apa/. The curves correspond to the labial parameter "Lip Height"(LH). The top curve shows the value of the lip height during the production of /'apa/. The curve in the middle shows the velocity of the LH parameter (how fast this parameter is during the production of /'apa/). On the bottom is displayed the spectrogram of the audio signal.

***Which other modalities are available/visible in the resource but have not been annotated ?***

In addition to the lip movements, which have been annotated, cf. above, the synchronized audio is available.

## 2.12.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

## 2.12.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The resource may come in different formats. The video recording of the speaker's face is available at CNR of Padova (see address below). The temporal values of the articulatory parameters (jaw, labial aperture, lip height, lip width, upper and lower lip protrusion) are also available. The movements have been sampled at 100Hz and correspond to spatial displacements of the markers. The format of this file is as follows: on the first line is written the number, *n*, of samples of the recording /'VCV/, then follows a list of *n* spatial values for each of the articulatory parameters. Thus if there are *p* articulatory parameters, the files would be of *n***p*+1 lines:

151

16.658995

18.709934

20.408096

21.788244

22.978979

24.111238

25.231750

…

***How much data does the resource contain?***

The resource contains 1260 /'VCV/ symmetric sequences: each /'VCV/ sequence (C being one of the 21 Italian consonants and V = /a, i, u/) was pronounced 5 times by 4 speakers.

***Who created the resource and when?***

The resource was created by CNR di Padova, Istituto di Fonetica e di Dialettologia from 1991 to 1995.

***How was the resource created?***

The resource was created using the ELITE Motion Analyser technology from BTS. Their address is:
BTS S.p.A.
Via C. Colombo, 1A
20094 Corsico Milano
Italy
Tel +39-02458751 Fax +39-0245867074
http://www.bts.it

The 8 markers on the face are passive reflectors. They have a hemispherical shape of 1mm of diameter. Their elaboration over time is done in real-time with dedicated hardware. Two infrared CCD (Change Coupled Device) cameras are used to record the markers movements. The cameras are set up to have a frontal and a side view of the speaker's face. These cameras are able to detect the movement of the markers to a precision of 0.04mm. Marker detection is based on pattern recognition technique and provides the system with great flexibility allowing its use even in the presence of disturbances brighter than markers in indoor as well as in outdoor applications. In order to avoid any disturbances to the subject, infrared flashes are adopted. Recognised markers are displayed in real time on a monitor and their coordinates are sent, through a DMA (Direct Memory Access) interface, to the computer for further processing, together with the data provided by other analogue and/or digital instrumentation. Both cameras are connected to the A/D converter, which translates the raw camera analogue data to digital data appropriate for computer analysis.

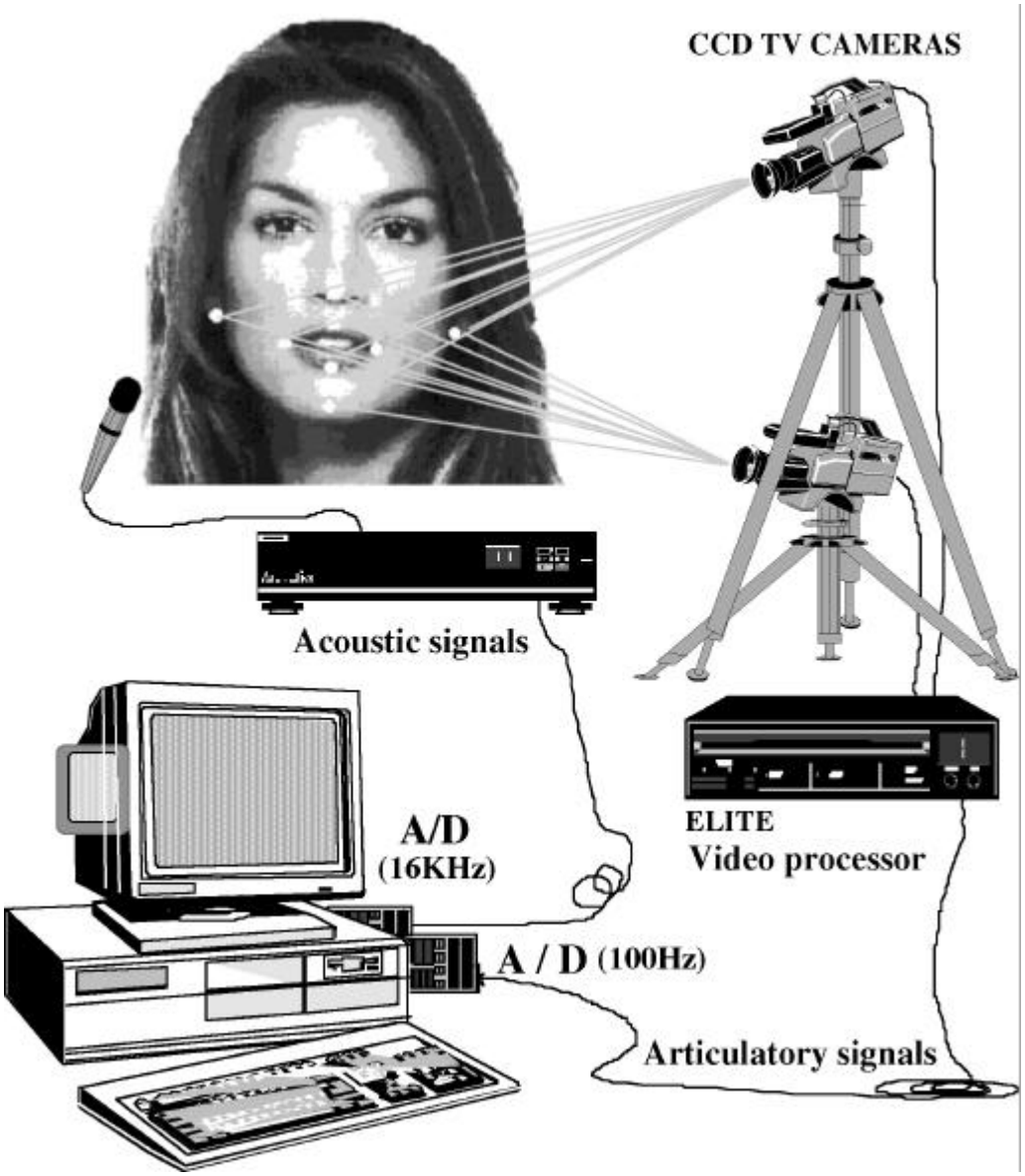


**Figure 2.12.2.** The setup which was used for creating the database.

Elite allows the camera data to be analysed in spatial and temporal terms.

- Spatial analysis allows the experimenter to determine the displacement of each individual marker in X, Y and Z coordinates, and the way in which each marker moves in relation to the other markers. For example, it is possible to determine the distance between the mid point of the upper lip and the mid point of the lower lip to assess the articulatory value of the lip height.
- Temporal analysis involves integration of the displacement profiles to obtain velocity and acceleration profiles. In this way details about the speed of movement and the way in which this speed changes in time can be assessed with an accuracy of ten milliseconds.

***What is the application area?***

One of the applications of this database is the study of lip movements during speech. Another is analysis of lip movements in order to drive a 3D talking head.

***What was the original purpose of creating the resource?***

The original purpose of creating the resource was to study lip shape characterisation during speech.

## 2.12.6 Accessibility

***How does one get access to the resource?***

One should contact
Dr. Emanuella Magno-Caldognetto
Istituto di Fonetica e Dialettologia
C.N.R. - Consiglio Nazionale delle Ricerche
Via G. Anghinoni, 10
35121 Padova, Italy.
Web site: http://nts.csrf.pd.cnr.it/IFD/Pages/emanuela.htm
Email: magno@csrf.pd.cnr.it

***Is the resource available for free or how much does it cost?***

Until now the resource has been distributed only under research collaboration and therefore for free.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added to the original resource in terms of an elaboration of phonetic and phonological data.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The main actor is collaborating with CNR in Padova to elaborate an Italian talking head. CNR provided a set of /'VCV/ data that had to be elaborated to be able to drive a 3D graphics model of a talking face. Therefore, the main actor had access to the resource while writing their contribution to 8.1. The verifying actor did not have access to the resource. Furthermore, Dr. Emanuella Magno-Caldognetto kindly validated the final description of the resource.

## 2.12.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for research on lip shape characterisation during speech to study articulatory movement for Italian, coarticulatory phenomena and also to elaborate an articulatory model to drive 3D talking faces.

***Who used the resource so far/who are the target users of the resource?***

So far the resource has only been used by the creators of the resource, namely CNR in Padova.

***Is the resource language dependent or language independent?***

The speech in the resource is in Italian.

## 2.12.8 Conclusion

***How interesting/important/high quality is the resource?***

This is the only Italian resource that contains information on lip shape during the production of /‘VCV/.

***What do the authors regret ( if anything) not to have done while building the resource?***

The authors regret not having done video recording of the subjects’ face while recording the markers movement with Elite. Having video data of the subjects’ lip movement could have been used as material on which various automatic recognition programs could have been applied to get information on kinematics behaviour and/or image elaboration of the lip. Such data elaboration could have been coupled with computer graphics technique that provides information on lip shape, as well as on teeth and tongue visibility.

# 3 Dynamic Facial Data Resources without Audio

## 3.1 LIMSI Gaze Corpus (CAPRE)

### 3.1.1 Description header

***Main actor***

LIMSI-CNRS : Jean-Claude MARTIN (martin@limsi.fr)

***Verifying actor***

IMS: Steve Berman (steve@ims.uni-stuttgart.de)

***Date of last modification of the description***

August 4th 2001

### 3.1.2 References

***Web site(s)***

Christophe Collet's website: http://www.limsi.fr/Individu/collet/

A short description of the corpus: http://www.limsi.fr/Individu/collet/Public_html/CapRe.html

Publications regarding the corpus:
http://www.limsi.fr/Individu/collet/Public_html/Publications/publi.html

***Short description***

A corpus of 40 video movies recording people sitting in front of a screen and gazing at different locations on the screen.

***Illustrative sample picture or video file***

**Figure 3.1.1** An example of still pictures taken from the resource.

***References to additional information on the reviewed resource***

C. Collet, A. Finkel, R. Gherbi (1998). CapRe: a Gaze tracking system in Human-Machine Interaction. Journal of Advanced Computational Intelligence, Vol.2 No.3, juin 1998, pp.77-81. Can be downloaded from: http://www.limsi.fr/Individu/collet/Public_html/Publications/publi.html

## 3.1.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

32 different humans have been recorded in the whole resource.

***How many humans are recorded at the same time?***

Only one human is recorded in the same frame.

***What is their profile?***

The subjects are people from LIMSI laboratory. 10 females and 23 males between 21 and 58 years old. No other information available.

***Which human body parts are visible in the resource?***

The face of the subjects are visible in the resource.

***Which modalities are annotated?***

Gaze (direction) and face movement have been annotated.

***Which other modalities are available/visible in the resource but have not been annotated?***

Facial expression is available in the resource but has not been annotated.

### 3.1.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans (none, graphical screen, computer pen, tactile screen, dataglove, loudspeakers…)?***

None.

### 3.1.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The file types included in the resource are movie files.

***How much data does the resource contain ?***

The resource contains 40 video sequences = 15 000 images of 32 subjects sitting in front of the screen.

***Who created the resource and when?***

The resource was created by Christophe Collet, Rachid Gherbi. A date has not been specified.

***How was the resource created?***

The subjects were instructed to click (and thus to look) at different locations on the screen. This procedure was the recorded.

***What is the application area?***

None.

***What was the original purpose of creating the resource?***

This work is related to a real-time vision system (CapRe), allowing computers to be aware of spontaneous user's actions in the context of man-machine interaction. It processes video image sequences, grabbed by a camera placed between the keyboard and the monitor. The system localizes and tracks face components. It performs processes that automatically detect the user, localize and then track his face, nose and eyes. These cases are performed by combining image processing techniques and pattern recognition methods. The tracking is based on a prediction-verification approach, using dynamic information. CapRe continuously and automatically adapts its recognition parameters to take into account the environment variations. The goal of this corpus was to proceed to a quantitative evaluation of the CapRe system.

### 3.1.6 Accessibility

***How does one get access to the resource?***

One can get access to the resource by contacting Christophe Collet, collet@limsi.fr, and Rachid Gherbi, gherbi@limsi.fr.

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added in the form of annotations of the bounding box of the face and of the gazed location on the screen.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

No. The review is based on web information and the reference listed above.

### 3.1.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

It is used by the developers of the Capre system.

***Who used the resource so far/who are the target users of the resource?***

Researchers in the processing of gaze or face.

***Is the resource language dependent or language independent?***

Language independent (no speech)

### 3.1.8 Conclusion

***How interesting/important/high quality is the resource?***

A very original resource on gaze.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 4 Static Facial Data Resources

## 4.1 3D_RMA: 3D database

### 4.1.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

May 30th, 2001.

### 4.1.2 References

***Web site***

The 3D_RMA: 3D database's own web site: http://www.sic.rma.ac.be/~beumier/DB/3d_rma.html

***Short description***

The resource contains data of 120 persons who were asked to pose twice in front of the system that was used to acquire the data. For each session, 3 shots were recorded with different (but limited) orientations of the head: straightforward / left or right / upward or downward. The original purpose of creating the resource was to make a validation of facial 3D face acquisition by structured light and recognition experiments by 3D comparison.

***Illustrative sample picture or video file***

It has not been possible for the reviewers to open the examples provided from the web site.

***References to additional information on the reviewed resource***

C. Beumier, M. Acheroy: Multi-Modal Database for Person Authentication. In the proceedings of the 10th International Conference on Image Analysis and Processing, Venice, Italy, 27-29 Sep, pp 704-708, 1999.

### 4.1.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

120 subjects are recorded in the whole resource.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

Among the 120 people, two thirds consisted of students from the same ethnic origins and with nearly the same age. The last third consisted of people of the academy, all aged between 20 and 60.

***Which human body parts are visible in the resource?***

Only the head of the subjects is visible in the resource.

***Which modalities are annotated?***

None.

***Which other modalities are available/visible in the resource but have not been annotated ?***

Facial expression is available as modality in the resource.

### 4.1.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 4.1.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The resource consists of 3D files. Each 3D file, with extension .xyz, is organized as a set of 3D points along stripes. Each stripe starts with a short integer indicating the number N of points in the stripe. Then N triplets of short (signed 2-byte) integers give the point coordinates. The resolution is 0.1 mm per unit. All short values are stored binary.

Ex: 2 x y z x y z 3 x y z x y z x y z 2 x y z x y z

The head is scanned (stripe order) from chin to forehead (increasing x) and the stripes are scanned left to right (increasing y). For technical reasons, the camera/projector head was rotated about 40°, so that the X and Y-axis do not correspond to respectively the vertical and horizontal axis in the files. No information is provided on how to open the 3D files.

***How much data does the resource contain?***

The resource contains data of 120 persons who were asked to pose twice in front of the system. For each session, 3 shots were recorded with different (but limited) orientations of the head: straightforward / left or right / upward or downward.

***Who created the resource and when?***

The resource was created between November 97 (session1) and January 98 (session2) at the
Signal and Image Centre (SIC) – Elec. Department
Royal Military Academy of Belgium
30 avenue de la Renaissance
B1000 Brussels
Belgium

***How was the resource created?***

SIC developed a 3D acquisition system based on structured light, which they used for the creation of the resource.

***What is the application area?***

No information available.

***What was the original purpose of creating the resource?***

The original purpose of creating the resource was:

- Validation of facial 3D face acquisition by structured light
- Recognition experiments by 3D comparison

The database has been acquired in the framework of the M2VTS project (see entry on the M2VTS Multimodal Face Database in this report for more information on the M2VTS project).

### 4.1.6 Accessibility

***How does one get access to the resource?***

The use of the 3D_RMA database is restricted to research purposes. The (re-) distribution of the database - or part of it - either in original form or modified is allowed only with the written permission of the concerned persons of the database (Prof. M. Acheroy, acheroy@elec.rma.ac.be; Charles Beumier, beumier@elec.rma.ac.be).

SIC distributes the database subject to the acceptance of the above conditions. By reception of the signed agreement, SIC will give access to the files for download (tar.gz 13MB).

***Is the resource available for free or how much does it cost?***

The resource is free for research purposes only.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

No value has been added to the original resource. But the database contains 3D, frontal, profile and speech.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

No, but contact was established to Charles Beumier, beumier@elec.rma.ac.be, who kindly answered questions. In addition to his help web information forms the basis of this description.

### 4.1.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for tests of 3D recognition. The quality is not high. It was to confirm the thesis that a 3D description of low quality already reveals much information. The resource has also been used for user authentication, lip tracking, face recognition.

***Who used the resource so far/who are the target users of the resource?***

The users of the resources were the partners in the M2VTS project.

***Is the resource language dependent or language independent?***

The resource is language independent.

### 4.1.8 Conclusion

***How interesting/important/high quality is the resource?***

The resource provides a set of 3D head data but no information on movement (head rotation, lip movement, facial expressions…).

***What do the authors regret ( if anything) not to have done while building the resource?***

Different problems encountered during the acquisition of the data. The subjects sometimes wore their spectacles, sometimes didn't, beards and moustaches were represented and some people smiled in some shots. Furthermore, the authors regret that only a few (14) women were available.

## 4.2 AR Face Database

### 4.2.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

May 23rd, 2001.

### 4.2.2 References

***Web site***

A description and examples from the database: http://rvl1.ecn.purdue.edu/~aleix/aleix_face_DB.html

***Short description***

This database contains more than 4,000 colour images corresponding to 126 people's faces (70 men and 56 women) and has been used for research within the area of face recognition and expression recognition devices.

***Illustrative sample picture or video file***

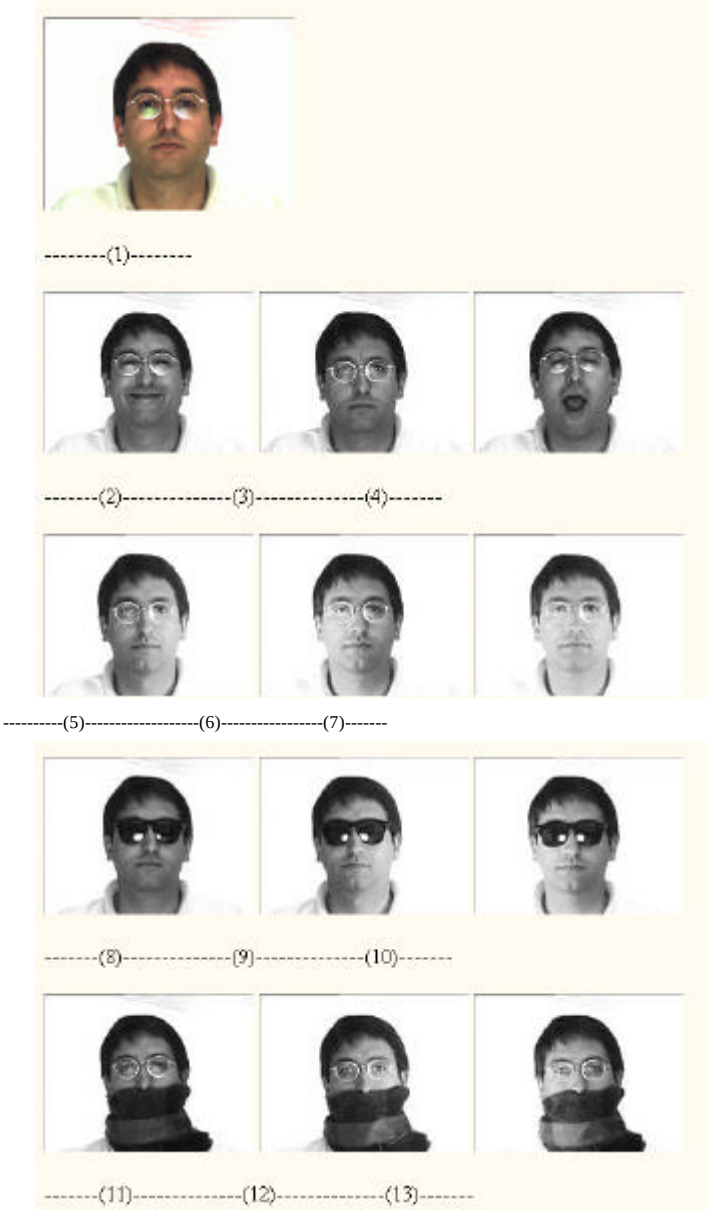


**Figure 4.2.1.** The example shows the different kinds of picture taken in the first of two sessions in the database.

See a movie example (due to compression the quality of the video is not very good): http://rvl1.ecn.purdue.edu/~aleix/arfd2.avi (1.97 Mb)

***References to additional information on the reviewed resource***

A.M. Martinez and R. Benavente: The AR Face Database. CVC Technical Report #24, June 1998.
The report can be obtained by contacting Dr. Aleix M. Martinez (aleix@ecn.purdue.edu).
More references can be found at Dr. Aleix Martinez' web page: http://rvl1.ecn.purdue.edu/~aleix/

### 4.2.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

126 humans have been recorded.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

70 of the 126 people in the resource are men and 56 are women. No other information is available.

***Which human body parts are visible in the resource?***

Only the faces are visible on the recordings in the resource – cf. figure 1.15.1.

***Which modalities are annotated?***

No information available.

***Which other modalities are available/visible in the resource but have not been annotated ?***

None.

### 4.2.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 4.2.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

All images are stored as RGB RAW files (pixel information). Images are of 768 by 576 pixels and of 24 bits of depth.

***How much data does the resource contain?***

The database contains more than 4,000 colour images corresponding to 126 people's faces (70 men and 56 women). Images feature frontal view faces with different facial expressions (neutral expression, smile, anger, scream), illumination conditions, and occlusions (sun glasses and scarf). The pictures were taken at the Computer Vision Center (CVC) under strictly controlled conditions. No restrictions on wear (clothes, glasses, etc.), make-up, hair style, etc. were imposed on the participants. Each person participated in two sessions, separated by two weeks (14 days) time. The same pictures were taken in both sessions.

1 : neutral expression
2 : smile

3 : anger
4 : scream
5 : left light on
6 : right light on
7 : all side lights on
8 : wearing sun glasses
9 : wearing sun glasses and left light on
10 : wearing sun glasses and right light on
11 : wearing scarf
12 : wearing scarf and left light on
13 : wearing scarf and right light on
14 to 26 : second session (same conditions as 1 to 13)
A total of 30 sequences of images were also grabbed to test dynamic systems. Each sequence consists of 25 colour images (same size as above).

CDs 1 to 8 contain the static images. CDs 9 and 10 contain the sequences.

***Who created the resource and when?***

The face database was created by Dr. Aleix M. Martinez and Robert Benavente at the CVC at the Purdue Robot Vision Lab at Purdue University, West Lafayette, IN 47907, USA. It has not been possible to find out when the resource was created.

***How was the resource created?***

No information available.

***What is the application area?***

The application area is research and education.

***What was the original purpose of creating the resource?***

According to the creators of the database there a was need for better databases in this research area. They feel that even with this database this is still the case.

### 4.2.6 Accessibility

***How does one get access to the resource?***

This database is publicly available from the above mentioned web site. It is free for researchers and institutions.
Permission to use but not reproduce the AR face database is granted to all researchers given that the following steps are properly followed:

1. Send an e-mail to Dr. Aleix M. Martinez (aleix@ecn.purdue.edu) before downloading the database.
2. All submitted papers (or any publicly available text) that uses mentions the AR face database must cite the following report: A.M. Martinez and R. Benavente: The AR face database. CVC Tech. Report \#24, 1998.
3. Permission is NOT granted to reproduce the database or post it into any web page that is not the AR face database web-page located at Purdue University.
4. Written permission must be obtained from Dr. Aleix M. Martinez if a faculty member desires to share the database with her/his co-workers or students. Even then, the database cannot be posted on a web-page accessible from outside the faculty research group.

5. No economical profit can be obtained from this database. Neither can the authors of the AR face database sell or commercialise its contents.

***Is the resource available for free or how much does it cost?***

The resource is available for free.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available***

No information available.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on web information and help from Dr. Aleix M. Martinez, who kindly answered questions.

### 4.2.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource can be used for face recognition and expression recognition devices. The resource may exclusively be used for non-profit research and education.

***Who used the resource so far/who are the target users of the resource?***

According to the creators of the resource it has been used extensively around the globe by many researchers.

***Is the resource language dependent or language independent?***

The resource is language independent.

### 4.2.8 Conclusion

***How interesting/important/high quality is the resource?***

The resource is interesting so far that the quality of the pictures is very high.

***What do the authors regret ( if anything) not to have done while building the resource?***

No information available.

# 4.3 AT&T Laboratories Database of Faces

## 4.3.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

June 12th, 2001.

## 4.3.2 References

***Web site***

The AT&T Database of Faces' own web site: http://www.uk.research.att.com/facedatabase.html
Download site of the database: ftp.uk.research.att.com/pub/data

***Short description***

The resource consists of ten different images of each of 40 distinct subjects. In compressed form the database contains 16MB. The database was used in the context of a face recognition project carried out in collaboration with the Speech, Vision and Robotics Group of the Cambridge University Engineering Department.

***Illustrative sample picture or video file***

**Figure 4.3.1.** One set of subject images from the database.

***References to additional information on the reviewed resource***

Ferdinando Samaria and Andy Harter: Parameterisation of a Stochastic Model for Human Face Identification. Proceedings of 2nd IEEE Workshop on Applications of Computer Vision, Sarasota FL, December 1994.

The paper can be downloaded at: ftp://ftp.uk.research.att.com:pub/data/att_faces.zip

### 4.3.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

40 humans have been recorded in the resource.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

The subjects are both male and female. No other information is available.

***Which human body parts are visible in the resource?***

Only the face of each subject is visible in the resource.

***Which modalities are annotated?***

None.

***Which other modalities are available/visible in the resource but have not been annotated ?***

Facial expression is available as modality in the resource but has not been annotated.

### 4.3.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 4.3.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The file type is .pgm and the pictures can be viewed with Paint Shop Pro under Windows or on UNIX (TM) systems using the 'xv' program. The size of each image is 92x112 pixels, with 256 grey levels per pixel. The images are organised in 40 directories (one for each subject), which have names of the form sX, where X indicates the subject number (between 1 and 40). In each of these directories, there are ten different images of that subject, which have names of the form Y.pgm, where Y is the image number for that subject (between 1 and 10). The files can be downloaded in compressed form and can be decompressed with WinZip.

***How much data does the resource contain?***

There are ten different images of each of the 40 different subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). In compressed form the database contains 16MB.

***Who created the resource and when?***

The set of face images was created between April 1992 and April 1994 at the AT&T laboratories, Cambridge UK.

***How was the resource created?***

By 35mm film, digitisation and conversion to .pgm format.

***What is the application area?***

Research into automatic face identification.

***What was the original purpose of creating the resource?***

The database was used in the context of a face recognition project carried out in collaboration with the Speech, Vision and Robotics Group of the Cambridge University Engineering Department.

### 4.3.6 Accessibility

***How does one get access to the resource?***

The database can be retrieved from: ftp.uk.research.att.com/pub/data

***Is the resource available for free or how much does it cost?***

The resource is available for free. When using these images, one should give credit to AT&T Laboratories Cambridge. One should contact Andy Harter (AHarter@uk.research.att.com) when using the database.

A convenient reference to the work using the database is the paper:

Ferdinando Samaria and Andy Harter: Parameterisation of a Stochastic Model for Human Face Identification. Proceedings of 2nd IEEE Workshop on Applications of Computer Vision, Sarasota FL, December 1994. The paper can be downloaded at:
ftp://ftp.uk.research.att.com:pub/data/att_faces.zip

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

No.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The verifying actor had access to the resource. Furthermore, the review is based on web information and information kindly provided by Andy Harter, who also validated the final description.

Andy Harter
Division Manager
AT&T Laboratories Cambridge
24A Trumpington Street
Cambridge CB2 1QA
ENGLAND
Email: aharter@uk.research.att.com
http://www.uk.research.att.com
Tel: +44 1223 343000
Fax: +44 1223 313542

### 4.3.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The database can be used for face recognition research.

***Who used the resource so far/who are the target users of the resource?***

There are no statistics on how many has used the database, since the database is freely available on the web.

***Is the resource language dependent or language independent?***

The resource is language independent.

### 4.3.8 Conclusion

***How interesting/important/high quality is the resource?***

The database contains a large amount of data. For each subject there are 10 different photos that may be very useful for face recognition projects.

***What do the authors regret ( if anything) not to have done while building the resource?***

No information available.

## 4.4 CMU Pose, Illumination, and Expression (PIE) database

### 4.4.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

June 11th, 2001.

### 4.4.2 References

***Web site***

The Pie database web site: http://www.ri.cmu.edu/projects/project_418.html

A short description of the database: http://www.ri.cmu.edu/pubs/pub_3462.html

ReadMe file: www.cs.cmu.edu/~simonb/pie_db

***Short description***

The resource contains 41,368 images of 68 people corresponding to images of persons performing 13 different poses under 43 different illumination conditions and with 4 different facial expressions. It was created in order to collect material for the design and evaluation of face recognition algorithms and has been used for facial expressions detection, temporal issue of facial expressions and other kinds of analysis of facial expressions.

***Illustrative sample picture or video file***

Several examples can be found at: http://www.cs.cmu.edu/~simonb/pie_db/talk/. The examples can be viewed with Windows MediaPlayer

***References to additional information on the reviewed resource***

T. Sim, S. Baker, and M. Bsat,: The CMU Pose, Illumination, and Expression (PIE) Database of Human Faces". Technical report CMU-RI-TR-01-02, Robotics Institute, Carnegie Mellon University, January, 2001.The report can be downloaded from: www.cs.cmu.edu/~simonb/pie_db

### 4.4.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

68 humans have been recorded in the whole resource.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

No information available.

***Which human body parts are visible in the resource?***

The head of each subject is visible in the resource.

***Which modalities are annotated?***

The resource has been used as a collection of material for the design and evaluation of face recognition algorithms, but has not been annotated.

***Which other modalities are available/visible in the resource but have not been annotated ?***

None, since the database is a collection of still pictures.

### 4.4.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 4.4.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The format of the resource is a database of a collection of still pictures.

***How much data does the resource contain?***

The resource contains 41,368 images of 68 people that correspond to images of persons performing 13 different poses under 43 different illumination conditions and with 4 different facial expressions. The resource was created to collect material for the design and evaluation of face recognition algorithms.

***Who created the resource and when?***

The resource was created between October and December 2000 at the Robotics Institute of the Carnegie Mellon University.

***How was the resource created?***

13 cameras and 21 flashes were set up in a room. The subject came in to the room and were asked to smile and talk. They were not told to say anything specific. Images were shot while the subjects smiled and talked.

***What is the application area?***

The application area is face recognition.

***What was the original purpose of creating the resource?***

The original purpose of creating the resource was to collect material for the design and evaluation of face recognition algorithms.

### 4.4.6 Accessibility

***How does one get access to the resource?***

One should contact Simon Baker (simonb@cs.cmu.edu) to get a copy of the database.

However, one should keep in mind that the database is 40+GB. Therefore, the only way to distribute it is via (E)IDE drive. Furthermore, one will need such a drive to store the database. The procedure is that people send Simon Baker the drive, who then takes care of copying the database onto the drive, where after the drive it returned.

***Is the resource available for free or how much does it cost?***

The resource in available for free, but one has to pay for the shipment of the (E)IDE drive.

If using the database one should refer to the following paper:

T. Sim, S. Baker, and M. Bsat, “The CMU Pose, Illumination, and Expression (PIE) Database of Human Faces”, tech. report CMU-RI-TR-01-02, Robotics Institute, Carnegie Mellon University, January 2001.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

The database shows expression variation, but this consists solely of:

- Single images of: neutral, smiling, blinking
- Short videos of: "talking". There is no control over or annotation of what the person of saying. It is just random lip motion.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

Due to the extensive size of the resource, the reviewers did not have access to the resource. But contact was established to Simon Baker, who has been helpful in answering questions about the resource and has validated the final description. Besides his information and input the description is based on web information.

### 4.4.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for facial expressions detection, temporal issue of facial expressions and other kinds of analysis of facial expressions.

***Who used the resource so far/who are the target users of the resource?***

The database has not been in use yet. It has been distributed to 5 groups funded under the DARPA Human Identification at a Distance Grant. The main use seems to be face recognition across pose, this be training images of frontal and profile and test images of ¾ view.

***Is the resource language dependent or language independent?***

The resource is language independent.

### 4.4.8 Conclusion

***How interesting/important/high quality is the resource?***

The resource contains a very large set of images.

However, the database was largely designed for Face Recognition, rather than for Expression or Speech processing. Due to its size and the difficult procedure of making and obtaining copies of the database, the creators of the database ask people to read the following article before requesting a copy of the database:

Terence Sim, Simon Baker and Maan Bsat, *The CMu Pose, Illimination, and Expression (PIE) Database of Human Faces,* www.cs.cmu.edu/~simonb/pie_db

***What do the authors regret ( if anything) not to have done while building the resource?***

No information available.

# 4.5 Cohn-Kanade AU-Coded Facial Expression Database

## 4.5.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

May 23rd, 2001.

## 4.5.2 References

***Web site***

The Robotics Institute's Face Databases web site: http://www.ri.cmu.edu/projects/project_420.html

The Facial Expression Analysis web page by Professor Jeffrey Cohn, Department of Psychology, University of Pittsburgh: http://www.cs.cmu.edu/~face

***Short description***

The database includes approximately 2000 image sequences. Image sequences have been annotated using the Facial Action Coding System (FACS). The database was created in order to check facial expression detection algorithms.

***Illustrative sample picture or video file***

**Figure 4.5.1.** Still picture sample of 1 of the 31% female caucasians that has been recorded in the resource.

***References to additional information on the reviewed resource***

None.

### 4.5.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

210 subjects were recorded in the database.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

The profile of the subjects is that their age range from 18 to 50 years old. 69% are female and 31% are male; and 81% Caucasian, 13% African, and 6% other groups.

***Which human body parts are visible in the resource?***

Only the face of the subjects is visible in the recordings.

***Which modalities are annotated?***

The facial expressions coded have been coded as AU using the Facial Action Coding System – FACS. For more information on FACS see ISLE report 9.1, "Survey of Annotation Schemes and Identification of Best Practice".

***Which other modalities are available/visible in the resource but have not been annotated ?***

Since one the face of the subjects is visible in the resource no other modalities are available.

### 4.5.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 4.5.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The still pictures format found in the database is .png. The still pictures can be opened with Paint Shop Pro.

***How much data does the resource contain?***

The Cohn-Kanade AU-Coded Facial Expression Database includes approximately 2000 image sequences.

***Who created the resource and when?***

A research group at Carnegie Mellon University (CMU) and the University of Pittsburgh made the resource. The research group among others included Jeffrey Cohn, Takeo Kanade, and Adena Zlochower. Most image sequences were collected from 1996 to 1998.

***How was the resource created?***

The observation room where the recordings were made, was equipped with a chair for the subject and two Panasonic WV3230 cameras, each connected to a Panasonic S-VHS AG-7500 video recorder with a Horita synchronized time-code generator. One of the cameras was located directly in front of the subject, and the other was positioned 30 degrees to the right of the subject.

***What is the application area?***

Facial expression analysis and recognition.

***What was the original purpose of creating the resource?***

The original purpose was to aid in developing and testing algorithms for facial expression analysis.

### 4.5.6 Accessibility

***How does one get access to the resource?***

One should contact:

Jeffrey Cohn (jeffcohn@pitt.edu)

Mailing address:
Carnegie Mellon University
Robotics Institute
5000 Forbes Avenue
Pittsburgh, PA 15213

***Is the resource available for free or how much does it cost?***

The resource is available for free. One must sign a letter of agreement governing the terms of the images use prior to receipt of the data. Once one has returned the letter of agreement one is given a login and a password for an ftp server, from which one can download the resource.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available***

The resource has been annotated using The FACS coding scheme. Since the verifying actor had access to the resource these annotation results have been made available on the password secured ftp server.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on web information and Jeffrey Cohn was contacted with a request for a copy of the database and with questions. The letter of agreement was signed by NISLab and consequently NISLab got a copy of the database. Furthermore, Jeffrey Cohn kindly provided information and made a validation of the final description of the resources.

### 4.5.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource can and has been used for facial expression detection, detection of temporal value of facial expression (onset – offset time of facial expression).

***Who used the resource so far/who are the target users of the resource?***

The resource was made and has been used by a research group including people from both CMU and Pittsburgh University.

***Is the resource language dependent or language independent?***

The resource is language independent.

### 4.5.8 Conclusion

***How interesting/important/high quality is the resource?***

This resource is interesting since it has coded with Aus (FACS).

***What do the authors regret ( if anything) not to have done while building the resource?***

The database was part of a much large data collection effort. In selecting image sequences for digitising and FACS coding, preference was given to complex facial expressions. More recent efforts have given preference to facial expressions consisting of single action units. In related research, the creators of the resource have emphasized naturally occurring facial expressions, and facial expressions recorded under conditions of varying illumination and pose.

# 4.6 FERET Database Demo

## 4.6.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

June 11th, 2001.

## 4.6.2 References

***Web site***

The FERET Database Demo's own web site:
http://vismod.www.media.mit.edu/vismod/demos/facerec/feret.html

***Short description***

The FERET database demo is an automatic system for recognition and interactive search in the FERET face database. Figure 1.19.1. shows the result of a typical Photobook similarity search on the FERET database. The user selected the face at the upper left. The remainder of the faces are the most similar faces from the 575 frontal views in the FERET database. Note that the first four images (in the top row) are all of the same individual. Also note that the database represents a realistic application scenario where position, scale, lighting and background are not uniform. Consequently, an automatic face processing system is used to correct for translation, scale, and contrast. Once the images are geometrically and photometrically normalized, they can be used in the standard eigenface technique. Eigenface is a related set of facial characteristics that a computer uses to recognize a person's face. The eigenface technique is a technique whereby each face image is deconstructed into separate eigenfaces, which enables a computer to distinguish the different facial characteristics of the subjects face.

***Illustrative sample picture or video file***

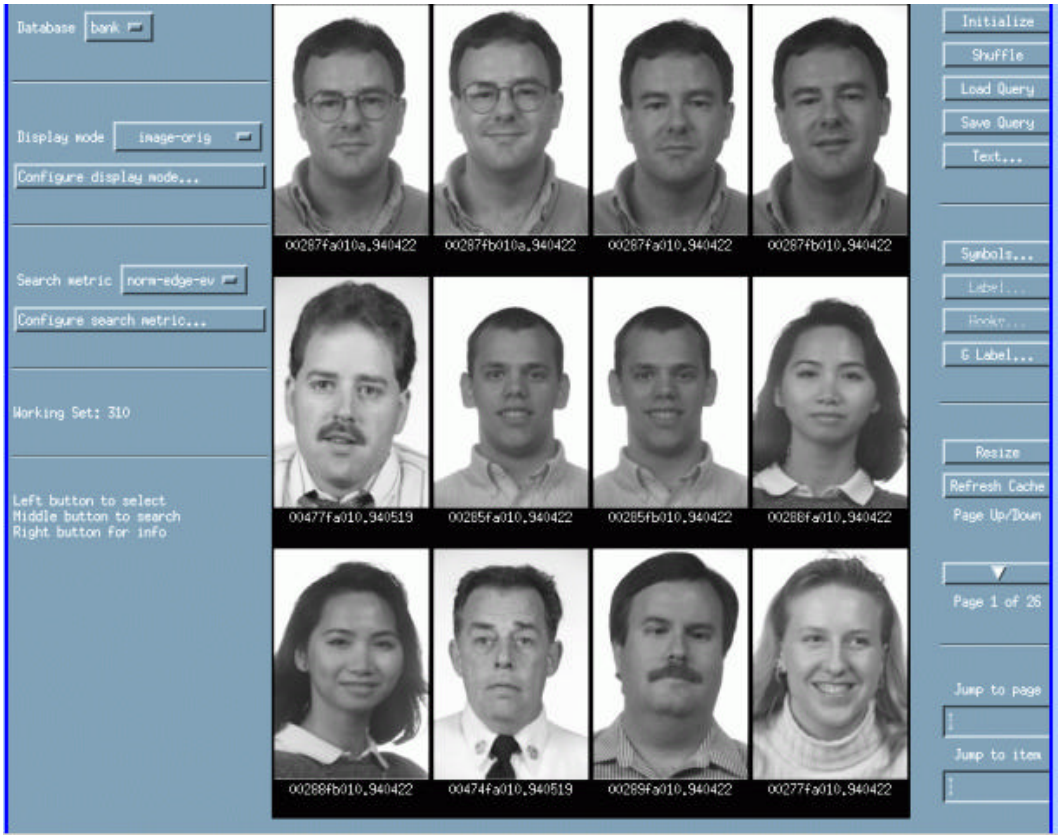


**Figure 4.6.1.** A typical Photobook similarity search on the FERET database.

***References to additional information on the reviewed resource***

None

## 4.6.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

310 people have been recorded in different poses and views (frontal and several profile views), under different lighting conditions, background, locations and times.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

The subjects are both male and female. No other information is available.

***Which human body parts are visible in the resource?***

Only the face of each subject is visible in the resource.

***Which modalities are annotated?***

None

***Which other modalities are available/visible in the resource but have not been annotated ?***

Facial expression is available as modality in the resource.

### 4.6.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 4.6.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

No information available.

***How much data does the resource contain?***

The resource contains 575 images of 310 people in different poses and views (frontal and several profile views), under different lighting conditions, background, locations and times.

***Who created the resource and when?***

No information available.

***How was the resource created?***

No information available.

***What is the application area?***

No information available.

***What was the original purpose of creating the resource?***

The resource was created for purposes of face recognition.

### 4.6.6 Accessibility

***How does one get access to the resource?***

One should contact:

Jonathan Phillips, jphillip@nvl.army.mil with a request for the resource. A link to an interactive web demo can be found on the FERET database demos own web site: http://vismod.www.media.mit.edu/vismod/demos/facerec/feret.html, but the link does not work

without a permission to access, and since the reviewers request for further information about the FERET database and a permission to try out the demo was never responded to, it has not been possible to try out the demo or to get hands-on experience with the database.

***Is the resource available for free or how much does it cost?***

The resource is available for free.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

No information available.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on web information. Jonathan Phillips has been contacted for information and with a request for permission to access the interactive web demo, which could improve the description, but he has not responded to the request. Therefore, many questions remain unanswered.

### 4.6.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used in the standard eigenface technique.

***Who used the resource so far/who are the target users of the resource?***

No information available.

***Is the resource language dependent or language independent?***

The resource is language independent.

### 4.6.8 Conclusion

***How interesting/important/high quality is the resource?***

The FERET database is a common database used to test image analysis and image recognition algorithms. The database contains a large amount of data.

***What do the authors regret ( if anything) not to have done while building the resource?***

No information available.

# 4.7 Psychological Image Collection at Stirling (PICS)

## 4.7.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

June 11th, 2001.

## 4.7.2 References

***Web site***

The image collection: http://pics.psych.stir.ac.uk/index.html, http://pics.psych.stir.ac.uk/cgi-bin/PICS/New/pics.cgi

***Short description***

This is a collection of images useful for research in Psychology, such as sets of faces and objects (such a bench, chair, lights, etc.).

***Illustrative sample picture or video file***

**Figure 4.7.1.** Example where the user has chosen one frontal face image of a male. Taken from the Aberdeen, set.

***References to additional information on the reviewed resource***

None.

## 4.7.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

No information available.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

No information available.

***Which human body parts are visible in the resource?***

Only the face of each subject is visible in the resource.

***Which modalities are annotated?***

None.

***Which other modalities are available/visible in the resource but have not been annotated ?***

Facial expressions are available as modality in the resource.

### 4.7.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 4.7.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The file type of the resource is .gif image format.

***How much data does the resource contain?***

The database contains several sets of face images and other objects. The part of the database that contains images of faces is further divided into sets:

**Figure 4.7.2. Set.** Aberdeen. Set size: 690 images (colour images). Set description: Colour faces from Ian Craw at Aberdeen. Here shown one example from set.

**Figure 4.7.3.** Set: nottingham_scans. Set size: 100 images (greyscale images). Set description: 50 males and 50 females in neutral, frontal pose. Here shown one example from set.

**Figure 4.7.4.** Set: UNFAM. Set size: 37 images (greyscale images). Set description: Faces grabbed from TV: 21 males and 16 females. Here shown one example from set.

**Figure 4.7.5.** Set: unfamiliar_faces. Set size: 20 images (greyscale images). Set description: Faces on a black background. Here shown one example from set.

**Figure 4.7.6.** Set: Stirling_faces. Set size: 331 images (greyscale images). Set description: 35 identities (18 female, 17 male), in 3 poses and 3 expressions. Here shown one example from set.

**Figure 4.7.7.** Set: nott-faces-originals. Set size: 384 images (greyscale images). Set description: Each face in four expressions with a frontal view, two expressions with a 3/4 view, and one frontal view wearing a bathing cap over hair. Here shown one example from set.

**Figure 4.7.8.** Set: more-unfam. Set size: 29 images (greyscale images). Set description: A selection of unfamiliar faces. Here shown one example from set.

The database contains also set of images on the following:

- Objects
- Drawings
- Textures
- Natural scenes

***Who created the resource and when?***

Ian Craw at Aberdeen University has created the set named Aberdeen. Information on the year of creation and information on the other sets is not available.

***How was the resource created?***

Some images were taken with a camera, but it is not specified which kind of camera was used, and whether light and other devices were used. Some of the pictures are still pictures captured from TV. It has not been specified exactly how this was done and which TV programs were used.

***What is the application area?***

It is a database of images that has been used in psychological research, in particular for visual perception, memory and processing.

***What was the original purpose of creating the resource?***

The database is a collection of images, which have been used, in Psychological research.

On the whole, this research is concerned with visual perception, memory and processing.

## 4.7.6 Accessibility

***How does one get access to the resource?***

The resource can be downloaded from: http://pics.psych.stir.ac.uk/index.html, http://pics.psych.stir.ac.uk/cgi-bin/PICS/New/pics.cgi

Contact person: Peter Hancock (p.j.b.hancock@stir.ac.uk)

***Is the resource available for free or how much does it cost?***

The resource is available for free.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

No.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The verifying actor had access to the database. Furthermore, the description of the resource is based on web information. Peter Hancock (p.j.b.hancock@stir.ac.uk) has been contacted with questions about the resource, but it has not been possible to obtain a response. Therefore, some questions remain unanswered.

## 4.7.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The database is a collection of images, which have been used in psychological research.

The original research connected to the database was concerned with visual perception, memory and processing.
The database can also be used for face recognition and expression recognition.

***Who used the resource so far/who are the target users of the resource?***

No information is available, since the creators have no statistical information on who has downloaded the resource.

***Is the resource language dependent or language independent?***

The resource is language independent.

### 4.7.8 Conclusion

***How interesting/important/high quality is the resource?***

The database is very large. In some sets, each subject has been photographed with different expressions, which makes the database useful for face and expression recognition.

***What do the authors regret ( if anything) not to have done while building the resource?***

No information available.

# 4.8 TULIPS 1.0

## 4.8.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

June 12th, 2001.

## 4.8.2 References

***Web site***

Site, where one gets access to both the audio and still pictures in the resource: ftp://mplab.ucsd.edu/pub/tulips1/

The read me file of the database: ftp://mplab.ucsd.edu/pub/tulips1/README

***Short description***

The database contains both raw acoustic signal traces, cepstral processed audio files and video files of subjects saying respectively the digits 1,2,3 and 4. The video files have been reduces to still pictures. As can be seen below, only the mouth of each subject is visible in the still pictures of the resource. The resource can be used to test lip-tracking algorithm.

***Illustrative sample picture or video file***

**Figure 4.8.1.** Example of one set of still pictures from the database. The subject is saying the digit "one".

***References to additional information on the reviewed resource***

Movellan J. R.: Visual Speech Recognition with Stochastic Networks. In G. Tesauro, D. Toruetzky, & T. Leen (eds.): Advances in Neural Information Processing Systems, Vol 7, MIT Press, Cambridge, 1995. The paper can be downloaded from http://cogsci.ucsd.edu/~movellan/.

### 4.8.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

Twelve humans have been recorded for creation of the resource,

***How many humans are recorded at the same time?***

Only one human is recorded in the same frame.

***What is their profile?***

The subjects are both male and female and are undergraduate students from the Cognitive Science Program at UCSD.

***Which human body parts are visible in the resource?***

Only the mouth of the subjects is visible in the resource.

***Which modalities are annotated?***

None.

***Which other modalities are available/visible in the resource but have not been annotated ?***

The available modalities are audio in the form of acoustic signals and the corresponding mouth movements.

The audio files have been cepstral processed.

### 4.8.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 4.8.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The database contains both raw acoustic signal traces (.au files) and cepstral processed audio files in .pgm format at 30 frames and video files in .pgm format . The video files have been reduces to still pictures.

The format of the .au files is as follows: The first 28 bytes in each file are reserved for header information, using the standard .au format. Signal information (video) starts on byte number 29 (byte 28 using zero-offset). Each byte encodes acoustic energy on each sample (1 byte per sample). The sampling rate is 11127 Hz. The subjects are saying respectively the digits 1,2,3 and 4. The audio files can be opened with Windows Media Player.

The processed audio files are in .pgm format (so they can actually be visualized) with 26 pixels arranged in the following order:

- 12 cepstral coefficients
- 1 log-power
- 12 cepstral derivatives
- 1 log-power derivative

The video files are in 100x75 pixel 8bit grey level. The file names found at the download site should be translated in the following manner:

Anthony12.00004 means this is the forth frame of subject Anthony saying the digit "1" for the second time. Each frame corresponds to 1/30 of a second.

***How much data does the resource contain?***

The database contains approximately 28 MB of files.

***Who created the resource and when?***

The database was compiled at Javier R. Movellan's laboratory at the Department of Cognitive Science, UCSD in 1995.

***How was the resource created?***

No information available.

***What is the application area?***

No information available.

***What was the original purpose of creating the resource?***

The original purpose of creating the resource was to test lip-tracking algorithm.

### 4.8.6 Accessibility

***How does one get access to the resource?***

The resource can be downloaded from the web at the following web site:

ftp://mplab.ucsd.edu/pub/tulips1/

The audio files can be played with Windows Media Player. The still picture files can be opened with Paint Shop Pro. It has not been possible to open the cepstral processed audio files or to find out exactly which software is needed for this.

One is requested to contact Javier R. Movellan, movellan@cogsci.ucsd.edu, if one uses the resource for any kind of research results.

***Is the resource available for free or how much does it cost?***

The resource is available for free.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

No information available.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The reviewers had access to the data resources and information through the web. An attempt to make contact to Javier R. Movellan, movellan@cogsci.ucsd.edu, who kindly referred the verifying actor to a better download site for the resource, but otherwise contributed with no new information. Therefore, some questions remain unanswered.

### 4.8.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource can be used to test lip-tracking algorithm.

***Who used the resource so far/who are the target users of the resource?***

Since the resources can be freely downloaded without any kind of registration from the web, it is assumed that there are no statistics available on this issue. It is furthermore assumed that the Javier R. Movellan's laboratory at the Department of Cognitive Science, UCSD has used the resource, but no evidence or references to this can be found. It is believed that researchers that work with the testing of lip tracking algorithms could find the resource useful.

***Is the resource language dependent or language independent?***

The acoustic signals of the resource are in English.

### 4.8.8 Conclusion

***How interesting/important/high quality is the resource?***

The database is small, but good for lip tracking algorithm.

***What do the authors regret ( if anything) not to have done while building the resource?***

No information available.

# 4.9 UMIST Face Database

## 4.9.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

June 11th, 2001.

## 4.9.2 References

***Web site***

Short description of the database: http://images.ee.umist.ac.uk/danny/database.html
Download site for the database: http://images.ee.umist.ac.uk/danny/face.tar.gz
Download site for the database: http://images.ee.umist.ac.uk/danny/cropped.tar.gz

***Short description***

The resource contains 564 images of 20 people and was created with the purpose of examining pose varying face recognition.

***Illustrative sample picture or video file***


**Figure 4.9.1.** Sequence of images from subject 1a.

***References to additional information on the reviewed resource***

Daniel Graham & Nigel M Allinson: Automatic Face Representation and Classification. BMVC, 1999 (http://citeseer.nj.nec.com/57512.html)

Daniel B Graham and Nigel M Allinson: Characterizing Virtual Eigensignatures for General Purpose Face Recognition. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman-Soulie and T. S. Huang (eds): Face Recognition: From Theory to Applications. NATO ASI Series F, Computer and Systems Sciences, Vol. 163, pp 446-456, 1998.

## 4.9.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

20 people have been recorded in the whole resource.

***How many humans are recorded at the same time?***

Only one subject is visible in the same frame.

***What is their profile?***

The subjects were a very mixed group of people, containing both female and male Asians and Caucasians, with or with glasses and with or without beards.

***Which human body parts are visible in the resource?***

The faces of the subjects are visible in the resource.

***Which modalities are annotated?***

None.

***Which other modalities are available/visible in the resource but have not been annotated ?***

Facial expression is available as modality in the resource.

### 4.9.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 4.9.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The files are all in .pgm format, approximately 220 x 220 pixels in 256 shades of grey. They can be viewed with Paint Shop Pro in Windows.

***How much data does the resource contain?***

The resource contains 564 images of 20 people.
Pre-cropped versions of the images are also available.
The database is available as a gzipped tar-file (29Mb).
The pre-cropped database is available as a gzipped tar-file (4.5Mb).

***Who created the resource and when?***

The resource was created over a period of six months from February to August 1997 by:
Daniel B. Graham and Nigel M. Allinson at Image Engineering and Neural Computing Group, Department of Electrical Engineering and Electronics, Manchester.

***How was the resource created?***

The images were captured on a Silicon Graphics Indy. The subjects were asked to rotate their head from one side to the other and during this process pictures were taken. Pictures were saved as 8-bit gray levels in standard PGM format in a C-program written by Daniel Graham.

***What is the application area?***

The resource can be used to establish pose tolerance, etc.

***What was the original purpose of creating the resource?***

The original purpose was to examine pose varying face recognition – it is one of the few databases that contain free form (horizontal) pose variation.

## 4.9.6 Accessibility

***How does one get access to the resource?***

The database can be downloaded from the web. One can use the database with some restrictions:

- UMIST grants the right to use the face database with the following restrictions: Only images of subject 1a, cf. figure 1.22.1., may be published (and then only with written permission). This is not out of vanity, but for legal reasons.
- The user must reference the following paper:

Daniel B Graham and Nigel M Allinson: Characterizing Virtual Eigensignatures for General Purpose Face Recognition. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman-Soulie and T. S. Huang (eds): Face Recognition: From Theory to Applications. NATO ASI Series F, Computer and Systems Sciences, Vol. 163, pp 446-456, 1998.

- The user must notify Daniel B. Graham or Nigel M. Allinson if s/he intends to use the database.

***Is the resource available for free or how much does it cost?***

The resource is available for free.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

No.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The description is based on web information, hands-on experience with the database and help from Daniel B. Graham, who has kindly provided information and made a validation of the final description.

## 4.9.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for face recognition.

***Who used the resource so far/who are the target users of the resource?***

The resource has been downloaded over 100 times now by people in Computer Vision labs throughout the world. Furthermore it is found on the Face Recognition Homepage, a homepage for researchers working with face recognition. Links to different face recognition resources can be found on this

homepage as well as links to the homepage of different researchers working with face recognition. (http://www.cs.rug.nl/~peterkr/FACE/face.html)

***Is the resource language dependent or language independent?***

The resource is language independent.

### 4.9.8 Conclusion

***How interesting/important/high quality is the resource?***

The main interest of this database is that it contains many poses of the each person.

***What do the authors regret ( if anything) not to have done while building the resource?***

The creators of the database would have preferred a more controlled environment and a homogenous race/gender set. Additionally, although not essential, the creators would have liked to do all the recordings on the same day.

# 4.10 University of Oulu Physics-Based Face Database

## 4.10.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

June 27th, 2001.

## 4.10.2 References

***Web site***

The University of Oulu Physics-Based Face Database's own web site: http://www.ee.oulu.fi/research/imag/color/pbfd.html

***Short description***

The University of Oulu Physics-Based Face Database is a face database collected at the Machine Vision and Media Processing Unit, University of Oulu, which contains colour images of faces under different illuminants and camera calibration conditions as well as skin spectral reflectance measurements of each person. In addition, the camera channel responses and the spectral power distribution of illuminants used are provided, thus the term physics-based. The database may be of general interest to face recognition researchers and of specific interest to colour researchers.

***Illustrative sample picture or video file***

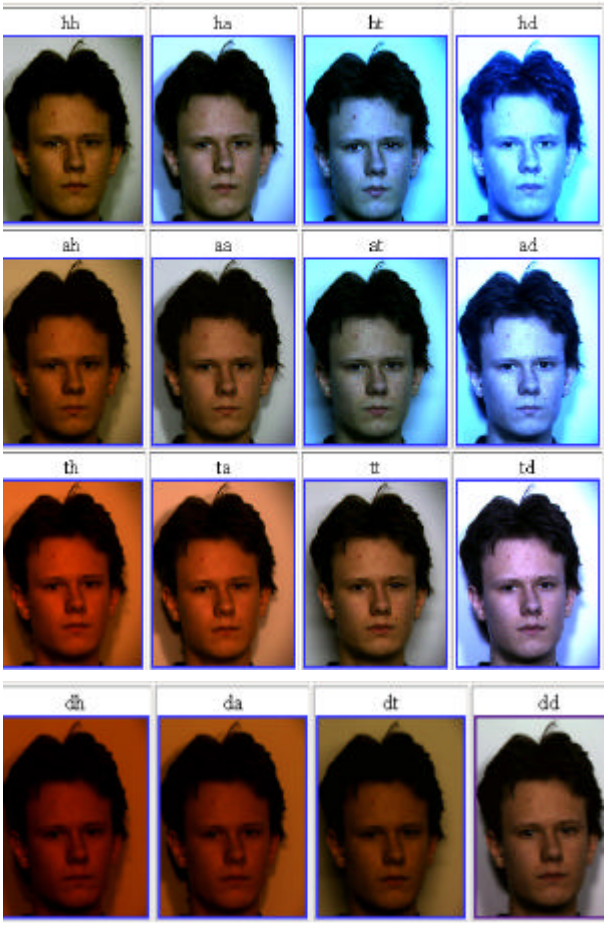


**Figure 4.10.1.** One set of images from the database. The letters above the images stands for: horizon (denoted as h) and is light at sunset or sunrise, incandescent (CIE Illuminant A) (a), TL84 fluorescent (t), and daylight (CIE D65) (d).

***References to additional information on the reviewed resource***

E. Marszalec, B. Martinkauppi, M. Soriano, M. Pietikäinen: A physics-based face database for color research. Journal of Electronic Imaging Vol. 9 No. 1 pp. 32-38, 2000.

M. Soriano, E. Marszalec, M. Pietikäinen: Color correction of face images under different illuminants by RGB eigenfaces. Proceedings from the 2nd Audio- and Video-Based Biometric Person Authentication Conference (AVBPA99), March 22-23, Washington DC USA pp. 148-153, 1999. Can be downloaded as a postscript file at: ftp://ftp.ee.oulu.fi/pub/tklab/msoriano/myAVBPA2.ps

M. Soriano, B. Martinkauppi, E. Marszalec, M. Pietikäinen: Making saturated images useful again. Proceedings from the SPIE Europto Conference on Polarization and Color Techniques in Industrial Inspection, June 17-18, Munich, Germany, pp. 113-121, 1999.

## 4.10.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

120 humans, 30 females and 95 males have been recorded in the whole resource.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

The photos are frontal face photographs.

***Which human body parts are visible in the resource?***

Only the face and neck of each subject is visible in the resource.

***Which modalities are annotated?***

Skin type and spectral reflectance, with or without eyeglasses, camera calibration illuminant and current illuminant have been annotated in the resource.

***Which other modalities are available/visible in the resource but have not been annotated ?***

Facial expression is available as modality in the resource.

### 4.10.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 4.10.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

Each image is 428x569 in 24-bit RGB, stored in .bmp format and is compressed using the command 'compress' in UNIX. For images of each person, a separate folder is created. The spectral reflectance of skin is its own folder. Also the SPD's of illuminants and camera responses have separate folders.

***How much data does the resource contain?***

The database contains 125 different faces each in 16 different camera calibrations and illumination conditions, an additional 16 if the person has glasses (totalling 2112 images).

***Who created the resource and when?***

The Machine Vision and Media Processing Unit, Infotech Oulu and Department of Electrical Engineering, University of Oulu, Finland. The database was created 1997-2000.

***How was the resource created?***

Faces were taken in frontal position captured under horizon, incandescent, fluorescent and daylight illuminant.

An image series of one person contains 16 frontal views each of which are taken under different illuminant calibration conditions by a 3CCD Sony DXC-755P camera.

Four illuminants provided by Macbeth Luminare were used for image acquisition: horizon (denoted as h), incandescent (CIE Illuminant A) (a), TL84 fluorescent (t), and daylight (CIE D65) (d).

Images of faces were taken under dark room conditions. A grey screen was placed behind the sitting subject and black curtains surrounded the sides. The camera was first calibrated under one illuminant then images were taken without changing the camera settings under this and the three other illuminants. The same procedure was used for all illuminants resulting in 16 conditions.

Spectral reflectance of facial skin was measured in three points (right cheek of image (cheek 1 in plot), forehead and left cheek of image (cheek 2 in plot)) from 400nm to 700nm by 10nm steps with a Minolta CM-2002 spectrophotometer. In addition, spectral sensitivities of the R, G and B channels of the 3CCD Sony DXC-755P and the spectral power distribution of the four illuminants were included from 400nm to 700nm in 10nm steps.

***What is the application area?***

Scientific research and possible commercial applications of results.

***What was the original purpose of creating the resource?***

For face recognition under varying illuminant spectral power distribution:

-to investigate skin appearance under different color camera white balance and illumination conditions

-To create a face database for research which does not only include images of the object (face) but also spectral power distribution of illuminants and spectral reflectance of skin.

## 4.10.6 Accessibility

***How does one get access to the resource?***

The UOPB Face Database is available for research and verification purposes upon e-mail request to Professor Matti Pietikäinen. His contact address is:

Prof. Matti Pietikäinen
Machine Vision and Media Processing Unit
Department of Electrical Engineering and Infotech Oulu
PO Box 4500 FIN-90014 University of Oulu, FINLAND
TEL: +358 8 553 2782
FAX: +358 8 553 2612
e-mail: mkp@ee.oulu.fi

***Is the resource available for free or how much does it cost?***

The complete database is sent in 2 CDs for a fixed handling fee of US $50 to cover copying and postage costs (this is non-profit). Invoice will be delivered with the CD's.
For legal reasons and for the privacy of the database participants, only faces 1, 3, 14, 25, 94 and 111 may be used for presentation or publication. Also, one should cite one of the above four publications to refer to the database.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

The database also includes 3 spectral reflectance of skin per person measured from both cheeks and forehead, and RGB spectral response of camera used and spectral power distribution of illuminants. However, this has not been used for annotation.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on web information. Professor Matti Pietikäinen was contacted with a request for the database and with several questions, and Birgitta Martinkauppi replied on his behalf and kindly validated the description.

## 4.10.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for face recognition research, PCA color correction of facial images, adaptive skin color modeling, color-based face tracking and spectral studies of skin color signals.

***Who used the resource so far/who are the target users of the resource?***

Target users are color science researchers.

Since April 2001 the University of Oulu Physics-Based Face Database have been delivered to 30 universities and companies in 5 continents.

***Is the resource language dependent or language independent?***

The resource is language independent.

## 4.10.8 Conclusion

***How interesting/important/high quality is the resource?***

According to the creators of the database this is the first database, which has systematically taken images of objects (faces) under different illuminants with different camera calibration. Furthermore, also the reflectance of faces at three point is measured, SPD's of illuminants are given and spectral responses of the camera are known. Therefore the Physics-based face database offers a unique resource for those interested in skin color and varying illuminant spectral power distribution.

***What do the authors regret ( if anything) not to have done while building the resource?***

The creators of the resource regret not having used the full sensitivity range of the camera to avoid clipping in the extreme illumination and camera calibration conditions.

# 4.11 VASC – CMU Face Detection Databases

## 4.11.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

June 11th, 2001.

## 4.11.2 References

***Web site***

The VASC – CMU Face Detection Databases' own web site:
http://www.ri.cmu.edu/projects/project_419.html

Description of suggested use and data formats:
http://www.vasc.ri.cmu.edu/IUS/eyes_usr17/har/har1/usr0/har/faces/test/

At http://www.ri.cmu.edu/projects/project_420.html a complete list of CMU databases related to face research can be found.

***Short description***

The VASC – CMU Face Databases consist of three datasets for the testing and training of face detectors:

- The combined MIT / CMU test set with ground truth for frontal face detection at http://www.vasc.ri.cmu.edu/idb/html/face/frontal_images/index.html. This particular test set was originally assembled as part of work in Neural Network Based Face Detection. It combines images collected at CMU and MIT.
- The CMU test set II with ground truth for frontal and non-frontal face detection at http://www.vasc.ri.cmu.edu//idb/html/face/profile_images/index.html. The image dataset is used by the CMU Face Detection Project and is provided for evaluating algorithms for detecting frontal and profile views of human faces. This test set was collected at CMU by Henry Schneiderman and Takeo Kanade.
- Frontal and non-frontal training images with ground truth collected by Henry Rowley, Shumeet Baluja, and Henry Schneiderman at http://www.vasc.ri.cmu.edu/idb/html/face/train_images/index.html. This dataset contains human face images (both frontal and profile views) with ground truth.

Ground truth consists of the annotated positions of pre-specified features (e.g. eyes, nose, mouth, etc) for each face.

***Illustrative sample picture or video file***

**Figure 4.11.1.** Four examples taken from the combined MIT / CMU test set with ground truth for frontal face detection.

***References to additional information on the reviewed resource***


H. Rowley, S. Baluja, and T. Kanade: Rotation Invariant Neural Network-Based Face Detection. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 1998.

H. Rowley, S. Baluja, and T. Kanade: Neural Network-Based Face Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 1, pp. 23-38, January 1998.

H. Rowley, S. Baluja, and T. Kanade: Rotation Invariant Neural Network-Based Face Detection. Technical report CMU-CS-97-201, Computer Science Department, Carnegie Mellon University, December 1997.

H. Rowley, S. Baluja, and T. Kanade: Neural Network-Based Face Detection. Computer Vision and Pattern Recognition '96, June 1996.

H. Rowley, S. Baluja, and T. Kanade: Neural Network-Based Face Detection. DARPA Image Understanding Workshop, February 1996.

H. Rowley, S. Baluja, and T. Kanade: Human Face Detection in Visual Scenes. Advances in Neural Information Processing Systems 8, pp. 875 – 881, 1996.

H. Schneiderman: A Statistical Approach to 3D Object Detection Applied to Faces and Cars. Doctoral Dissertation, Technical report 00-06, Robotics Institute, Carnegie Mellon University, May 2000.

H. Schneiderman and T. Kanade: A Statistical Model for 3D Object Detection Applied to Faces and Cars. Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, June 2000.

H. Schneiderman and T. Kanade: Probabilistic Modelling of Local Appearance and Spatial Relationships for Object Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '98), pp. 45-51, July 1998.

All these papers can be downloaded in various formats at:
http://www.ri.cmu.edu/projects/project_419.html

## 4.11.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

Approximately 3,000.

***How many humans are recorded at the same time?***

Some frames include multiple humans, cf. figure 1.24.1.

***What is their profile?***

No information available.

***Which human body parts are visible in the resource?***

The resource includes pictures where everything ranging from only the face to the whole body is visible, cf. figure 1.24.1.

***Which modalities are annotated?***

The images are annotated with locations of specific features (e.g. eyes, nose, mouth). The annotation results and files are available from the databases own website.

***Which other modalities are available/visible in the resource but have not been annotated ?***

Both hand gestures, arm gestures, body posture and facial expression are available as modalities in the resource, cf. figure 1.24.1.

## 4.11.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

Interactive media may be visible and used by the humans in some of the pictures in the resource, but there is no information available on this issue.

## 4.11.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

All images are in .gif and .pgm formats and in greyscale. They can be opened with Netscape, Explorer, Microsoft Photo Editor, etc.

***How much data does the resource contain?***

The combined MIT / CMU test set with ground truth for frontal face detection contains 130 images divided in to three test sets: test set A, B and C (with 511 faces) in total.

The CMU test set II with ground truth for frontal and non-frontal face detection contains 208 images with 441 faces.

The frontal and non-frontal training images set with ground truth contains 1,589 frontal face images and 842 non-frontal images.

***Who created the resource and when?***

The combined MIT / CMU test set was created by MIT and CMU.

Test Set B was provided by Kah-Kay Sung and Tomaso Poggio at the AI Lab at MIT in 1995, and Test Sets A and C were collected at CMU (by Henry A. Rowley, Shumeet Baluja, and Takeo Kanade) in 1995-1996. It is uncertain whether they themselves photographed the subjects or whether the subjects supplied them with photos.

The CMU test set II with ground truth for frontal and non-frontal face detection was collected at CMU by Henry Schneiderman and Takeo Kanade from 1998-2000

Henry Rowley, Shumeet Baluja, and Henry Schneiderman collected the frontal and non-frontal training images with ground truth. The year of collection is not specified.

***How was the resource created?***

All of the above data sets were created by finding useful pictures on the web and by either taking photographs of the subjects or by making the subjects provide photographs. The photographs were originally collected using a wide variety of formats and cameras. They have been converted into the format of the resource using non-loss methods.

***What is the application area?***

Computer vision.

***What was the original purpose of creating the resource?***

The original purpose of creating the resource was to train and test face detection algorithms.

## 4.11.6 Accessibility

***How does one get access to the resource?***

The combined MIT / CMU test set at http://www.vasc.ri.cmu.edu/idb/html/face/frontal_images/index.html can be directly downloaded from the web site. The images are available individually or collectively in a tar file. The latter file will unpack into four directories:

- Test containing the images in Test Set A
- Test-low containing Test Set B
- New test containing Test Set C
- Rotated containing the Rotated Test Set

The CMU test set II at http://www.vasc.ri.cmu.edu//idb/html/face/profile_images/index.html is provided as ground truth for face location in two files. (Note: Some faces are listed in both files if they are between frontal view and profile view). The two files are available from the web site.

The frontal and non-frontal training images with ground truth collected by Henry Rowley, Shumeet Baluja, and Henry Schneiderman at http://www.vasc.ri.cmu.edu/idb/html/face/train_images/index.html can be obtained by completing and faxing a usage agreement to Henry Schneiderman at +001 412-268-6436. He will make the dataset available via a password protected account.

***Is the resource available for free or how much does it cost?***

All the datasets are available for free, but the creators want the users the reference specific papers when using the different datasets. More information and the specific papers can be found at the web sites of the datasets.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

The images are annotated with locations of specific features (e.g. eyes, nose, mouth).

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The description is based on web information and information kindly provided by Henry Schneiderman (hws@cs.cmu.edu), who also validated the final description. The verifying actor had access to the combined MIT / CMU test set and to the two files of ground truth for face location in the CMU test set II, since both of these datasets are directly available on the web. The verifying actor did not have access to the frontal and non-frontal training images with ground truth collected by Henry Rowley, Shumeet Baluja, and Henry Schneiderman.

## 4.11.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for face detection, for facial feature detection, for detection algorithms to associate names with faces and to locate facial features to be used by search engines for images on the web.

***Who used the resource so far/who are the target users of the resource?***

The resource has been created for and used by the CMU group and MIT group.

***Is the resource language dependent or language independent?***

The resource is language independent.

## 4.11.8 Conclusion

***How interesting/important/high quality is the resource?***

The resource provides large sets of images with a variety of situations (close up views of faces, groups of people, several nationalities, etc.)

***What do the authors regret ( if anything) not to have done while building the resource?***

The creators of the three datasets regret that they did not annotate more facial features.

# 4.12 Visible Human Project

## 4.12.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

June 12th, 2001.

## 4.12.2 References

***Web site***

The Visible Human's own web site: http://www.nlm.nih.gov/research/visible/visible_human.html
Several conferences on the visible human project have taken place. Information can be found on the mentioned website.

***Short description***

The Visible Human Project® is an outgrowth of the NLM's (National Library of Medicine) 1986 long-term range plan. It is the creation of complete, anatomically detailed, three-dimensional representations of the normal male and female human bodies. The resource primarily contains static pictures.

***Illustrative sample picture or video file***

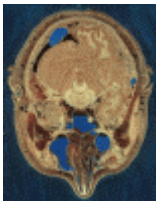

**Figure 4.12.1.** Human male head section, including cerebellum, cerebral cortex, brainstem, nasal passages (from Head subset).

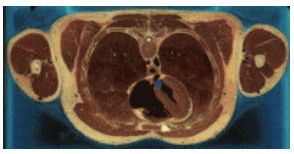

**Figure 4.12.2.** Human male thorax, including heart (with muscular left ventricle), lungs, spinal column, major vessels, musculature (from Thorax subset).

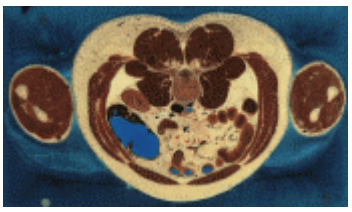

**Figure 4.12.3.** Human male abdomen, including large and small intestines, spinal column, musculature, subcutaneous fat (from Abdomen subset).

**Figure 4.12.4.** Human male upper thigh below femoral head, including prominent musculature, part of male reproductive system (from Pelvis subset).

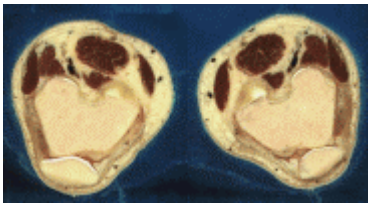

**Figure 4.12.5.** Human male knee, including patella (from Thigh subset).

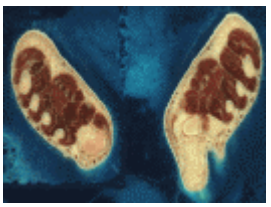

**Figure 4.12.6.** Human male feet (from Feet subset).

See also “From head to toe” - an animated trip through the Visible Human male cryosections [colour MPEG, 770810 bytes]. To do this you must have a Windows Media Player.

***References to additional information on the reviewed resource***

The Visible Human Project® Conference proceedings, 1996.

The Second Visible Human Project® Conference proceedings, 1998

The Visible Human Project®: From Data to Knowledge: An update of ongoing National Library of Medicine VHP initiatives.

The papers can be downloaded from this website:
http://www.nlm.nih.gov/research/visible/visible_human.html

### 4.12.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

Two humans have been recorded in the whole resource, a normal adult female and a normal adult male.

***How many humans are recorded at the same time?***

Only one body part of a subject is visible in the same frame.

***What is their profile?***

No information available.

***Which human body parts are visible in the resource?***

Full recording from head to toe divided into six primary sections: head, thorax, abdomen, thighs, knee and feet.

***Which modalities are annotated?***

None.

***Which other modalities are available/visible in the resource but have not been annotated ?***

MRI images, CT scans and full colour cryosections. MRI stands for Magnetic Resonance Imaging and is a method used by physicians to look inside the human body to obtain diagnostic information. Incorporating an advanced technology, MRI produces images of the anatomy without the use of radiation found in x-ray and CT scanning. CT scan stands for Computed Axial Tomography and is also called a CAT Scan.

### 4.12.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 4.12.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The format of the resource is UNIX compressed .raw files and that can be converted for viewing with Netscape, Internet Explorer and various photo editors.

***How much data does the resource contain?***

The male dataset consists of axial MRI images of the head and neck taken at 4 mm intervals and longitudinal sections of the rest of the body also at 4 mm intervals. The MRI images are 256 pixels by 256 pixels resolution. Each pixel has 12 bits of grey tone resolution.

The CT data consists of axial CT scans of the entire body taken at 1 mm intervals at a resolution of 512 pixels by 512 pixels where each pixel is made up of 12 bits of grey tone. The axial anatomical images are 2048 pixels by 1216 pixels where each pixel is defined by 24 bits of colour, each image consisting of about 7.5 megabytes of data. The anatomical cross-sections are also at 1 mm intervals and coincide with the CT axial images. There are 1871 cross-sections for each mode, CT and anatomy, obtained from the male cadaver - approximately 15 gigabytes of data.

The dataset from the female cadaver has the same characteristics as the male cadaver with one exception. The axial anatomical images were obtained at 0.33 mm intervals instead of 1.0 mm intervals. This results in over 5,000 anatomical images. The female dataset is about 40 gigabytes in size. Spacing in the "Z" direction was reduced to 0.33 mm in order to match the pixel spacing in the "XY" plane, which is 0.33 mm. This enables developers who are interested in three-dimensional reconstructions to work with cubic voxels.

A recent addition to the male dataset is the inclusion of higher resolution anatomical images. 70mm film taken during the original data collection phase has been digitised at a resolution of 4096 pixels by 2700 pixels where each pixel is made up of 24 bits of colour. As with the original anatomical images, there are a total of 1871 of these high-resolution images.

***Who created the resource and when?***

The resource was created by:

Visible Human Project®
National Library of Medicine
Building 38A, Room B1N-30
8600 Rockville Pike
Bethesda, MD 20894

The project started in 1986.

***How was the resource created?***

The resource was created using the two subjects' full bodies in three modes: traditional medical diagnostic procedures (MRI and CT scans), and colour cryosections. Croysections were derived from frozen cadavers milled in 1mm or .33mm increments, male and female respectively, and digitally photographing the remaining cadaver.

***What is the application area?***

The primary application areas are education, computer graphics, 3D rendering, and medical and surgical simulations.

***What was the original purpose of creating the resource?***

The long-term goal of the Visible Human Project® is to produce a system of knowledge structures that will transparently link visual knowledge forms to symbolic knowledge formats such as the names of body parts.

## 4.12.6 Accessibility

***How does one get access to the resource?***

A single license agreement covering use of both the male and female Visible Human Project® datasets is available, as either a text file or a WordPerfect file. The Visible Human Project® asks the user to make two copies of the agreement and have both signed as originals by the appropriate officials. The agreement requires the user to include a brief statement explaining the intended use of the dataset. Both signed copies of the agreement and the statement of the intended use of the data should be sent to:

Visible Human Project®
National Library of Medicine
Building 38A, Room B1N-30
8600 Rockville Pike
Bethesda, MD 20894

The agreement will be signed at the NLM and one of the originals will be returned the user. When the agreement is returned the user will be sent an account and password to the Visible Human Project® FTP site (if you wish to download all or part of the dataset via the internet), and information on how to purchase the dataset on 8 mm Exabyte or 4 mm DAT tape along with the agreement. The dataset is distributed from the FTP site and on tape in a Z-compressed UNIX TAR format. There are 6 sets of tapes of anatomical images corresponding to 6 body regions: head, thorax, abdomen, pelvis, thighs, and feet. A 7th tape contains all the MR and CT images.

If you have any questions please contact:

Michael J. Ackerman, Ph.D.
Project Officer
vhp@nes.nlm.nih.gov

***Is the resource available for free or how much does it cost?***

FTP access to the dataset is free. For tapes, each costs $150 in the US, Canada and Mexico, $300 elsewhere. A complete set of tapes for either the male or the female costs $1,000 in the US, Canada and Mexico, $2,000 elsewhere.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

In the second phase of the project segmentation, classification and the building of a prototype database (AnatLine) of the thorax region of the Visible Male has been completed, with AnatLine currently in beta testing. A future web-based atlas of the head and neck region is under development, the intent of which is to be a model for a new wave of educational applications. It will consist of six functional anatomy-teaching modules. The Visible Human Project Imaging Processing Tools have as their goal to create a self-sustaining development effort to support image analysis research in segmentation, classification and deformable registration of medical images. Use of the Next Generation Internet (NGI) is being investigated by employing the Visible Human's large image data sets in real time 2D and 3D visualisations under haptic control.

Some tools can be used in connection with the Visible Human Data Sets. Information on these can be found at:

http://www.nlm.nih.gov/research/visible/tools.html

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

No, due to the cost no access to the resource was obtained. Therefore, the review is based on web-information with the content edited and validated by Richard A. Banvard (vhp@nes.nlm.nih.gov) of the National Library of Medicine.

## 4.12.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for studies of anatomy, creation of synthetic models and test image segmentation algorithms.

***Who used the resource so far/who are the target users of the resource?***

The primary target users of the resource include anyone interested in studies of human anatomy. Furthermore, educators from elementary school to medical school, anatomists, computer scientist, graphics artist, animators, students and a host of other professionals are using it.

***Is the resource language dependent or language independent?***

The resource is language independent.

## 4.12.8 Conclusion

***How interesting/important/high quality is the resource?***

The database provides complete anatomic information on the normal male and female body.

***What do the authors regret ( if anything) not to have done while building the resource?***

The female cadaver was that of a post-menopausal woman, many have suggested that it would have been better to use that of a pre-menopausal subject.

# 4.13 Yale Face Database

## 4.13.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

May 30th, 2001.

## 4.13.2 References

***Web site***

The Yale Face Database's own web site: http://cvc.yale.edu/projects/yalefaces/yalefaces.html

Download site of the database: ftp://plucky.cs.yale.edu/CVC/pub/images/yalefaces/

***Short description***

The goal of the research behind the creation of this database is image analysis and image understanding, in particular face recognition.

***Illustrative sample picture or video file***

**Figure 4.13.1.** Four examples from the database. Upper left: happy face. Upper right: normal face. Lower left: sad face. Lower right: sleepy face.

***References to additional information on the reviewed resource***

P. N. Belhumeur and J. P. Hespanha and D. J. Kriegman: Eigenfaces vs.~{Fisherfaces}: Recognition using class specific linear projection. PAMI: Special Issue on Face Recognition. 19 (7), pp. 711-720, 1997.

### 4.13.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

15 subjects have been recorded in the whole resource.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

No information available.

***Which human body parts are visible in the resource?***

Only the face is visible in the resource, cf. figure 1.26.1.

***Which modalities are annotated?***

None.

***Which other modalities are available/visible in the resource but have not been annotated ?***

Facial expression is available as modality in the resource.

### 4.13.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 4.13.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The resource is in .gif format. The files are downloaded as files without extension. One must provide the files with .gif as extension, where after they can be viewed in Explorer, Netscape, Microsoft Photo Editor, and other photo editors.

***How much data does the resource contain?***

The data resource (size 6.4MB) contains 165 greyscale images in .gif format of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: centre-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink.

***Who created the resource and when?***

The resource was created by Peter Belhumeur and Athos Georghiades at the
Center for Computational Vision and Control
Yale University
51 Prospect Street, P.O. Box 208285
New Haven, CT 06520-8285
It has not been possible to determine when it was created.

***How was the resource created?***

No information available.

***What is the application area?***

No information available.

***What was the original purpose of creating the resource?***

The resource was created for the purpose of making face recognition.

## 4.13.6 Accessibility

***How does one get access to the resource?***

The resource can be downloaded from the web using the above ftp web site.

***Is the resource available for free or how much does it cost?***

The database is publicly available for non-commercial use.

When using the database, one should refer to:

P. N. Belhumeur and J. P. Hespanha and D. J. Kriegman: Eigenfaces vs.~{Fisherfaces}: Recognition using class specific linear projection. PAMI: Special Issue on Face Recognition. 19 (7), pp. 711-720, 1997.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

No information available.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on web information. Prof. David Kriegman (kriegman@yale.edu) and Prof. Peter Belhumeur (belhumeur@yale.edu) were contacted in order to gain more information, but did not reply.

### 4.13.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has so far been used for research on face recognition.

***Who used the resource so far/who are the target users of the resource?***

No information available.

***Is the resource language dependent or language independent?***

The resource is language independent.

### 4.13.8 Conclusion

***How interesting/important/high quality is the resource?***

The database provides picture variations of different individuals under different light conditions, with or without artifacts, and showing different emotions.

***What do the authors regret ( if anything) not to have done while building the resource?***

No information available.

# 4.14 Yale Face Database B

## 4.14.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

May 30th, 2001.

## 4.14.2 References

***Web site***

Download site for the database: ftp://plucky.cs.yale.edu/CVC/pub/images/yalefacesB/

***Short description***

The database consists of face images of subjects placed in a single light source. Each subject is seen under various different viewing conditions. The resource could be useful to people working with computer vision, image processing, pattern recognition, etc.

***Illustrative sample picture or video file***

**Figure 4.14.1.** Six images showing different poses of the same person.

***References to additional information on the reviewed resource***

Georghiades, A.S. and Belhumeur, P.N. and Kriegman, D.J.: From Few To Many: Generative Models For Recognition Under Variable Pose and Illumination. From the IEEE International Conference on Automatic Face and Gesture Recognition, pp. 277-284, 2000.

## 4.14.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

10 subjects are recorded in the whole resource.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

No information available.

***Which human body parts are visible in the resource?***

Only the face and the upper part of the torso are visible in the resource, cf. figure 1.27.1.

***Which modalities are annotated?***

None.

***Which other modalities are available/visible in the resource but have not been annotated ?***

Facial expression is available as modality in the resource.

## 4.14.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

## 4.14.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The downloadable images are 8-bit (grey scale) and stored in .pgm format. The size of each image is 640(w) x 480 (h). The images can be viewed with PaintShop Pro.

***How much data does the resource contain?***

The database contains 5760 single light source images of 10 subjects each seen under 576 viewing conditions (9 poses x 64 illumination conditions). For every subject in a particular pose, an image with

ambient (background) illumination was also captured. Hence, the total number of images is (5760+90=) 5850 but the image with ambient illumination is usually quite dark. The total size of the database is about 1GB. The 65 (64 illuminations + 1 ambient) images of a subject in a particular pose have been "tarred" and "gzipped" into a single file.

***Who created the resource and when?***

The resource was created by Peter Belhumeur and Athos Georghiades in June 1999 at the
Center for Computational Vision and Control
Yale University
51 Prospect Street, P.O. Box 208285
New Haven, CT 06520-8285

***How was the resource created?***

The images were captured with a Sony XC-75 camera with a linear response function.

***What is the application area?***

No information available.

***What was the original purpose of creating the resource?***

The original purpose of creating the resource was face recognition under various poses and illumination.

## 4.14.6 Accessibility

***How does one get access to the resource?***

The database can be freely downloaded from:

http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html

***Is the resource available for free or how much does it cost?***

The resource is free for research purposes only. If one uses the database one must refer to:

Georghiades, A.S. and Belhumeur, P.N. and Kriegman, D.J.: From Few To Many: Generative Models For Recognition Under Variable Pose and Illumination. From the IEEE International Conference on Automatic Face and Gesture Recognition, pp. 277-284, 2000.

Without permission from Yale, images from the database may not be incorporated into a larger database for public distribution.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

The coordinates of the centre of the face in each image are provided for all poses but the frontal one. For the frontal pose, the coordinates of the eyes and centre of the mouth are provided. See the README file at ftp://plucky.cs.yale.edu/CVC/pub/images/yalefacesB/ for more information.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The description is based on web information and information kindly provided by Athos Georghiades (athinodoros.georghiades@yale.edu). The verifying actor had access to the resource over the web.

## 4.14.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for face recognition under various poses and illuminations. It was primarily gathered as a test set for recognition under variations in illumination and pose, but can also be used for face detection, although there is only one face per image. This database can also be used in testing image compression algorithms as well as in psychophysical experiments.

***Who used the resource so far/who are the target users of the resource?***

To the knowledge of the creators the resource has so far only been used by the creators themselves. The creators of the resource believe that the resource could be useful to people working with computer vision, image processing, pattern recognition, etc.

***Is the resource language dependent or language independent?***

The resource is language independent.

## 4.14.8 Conclusion

***How interesting/important/high quality is the resource?***

The resource offers a large variety of photos of each person under different conditions.

***What do the authors regret ( if anything) not to have done while building the resource?***

No information available.

# 5 Lesser Known/Used Facial Data Resources

## 5.1 3D Surface Imaging in Medical Applications

### 5.1.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Date of last modification of the description***

June 12th, 2001.

### 5.1.2 References

***Web site***

Medical Facial Surface Scanner Siggraph '91 Position Paper: http://www.ncsa.uiuc.edu/SDG/DigitalGallery/FACE.html

### 5.1.3 Description

**Figure 5.1.1.** The following Images were created with a new 3D Surface Imaging System to digitise human head forms.

The digitised 3D data is being used in many medical applications. Various techniques are being developed by using image processing, computer graphics, computer-aided design and engineering and other software packages to refine and manipulate the data. All the images were created on a Silicon Graphics WorkStation.

# 5.2 ATR Database for Talking Face

## 5.2.1 Description header

***Main actor***

LIMSI: Jean-Claude MARTIN (martin@limsi.fr)

***Date of last modification of the description***

August 4th, 2001 (nakamura@slt.atr.co.jp has been contacted by email on August 8th)

## 5.2.2 References

Nakamura, S. et al: Multimodal Corpora for Human-Machine Interaction Research, Proc. of ICSLP, Volume IV, pp. 25-28, 2000

***Web site***

Contact nakamura@slt.atr.co.jp

## 5.2.3 Description

Speech and 3D lip position data for male and female speakers of Japanese are recorded at 125Hz using the OPTOTRAK 3D position sensoring system. These positions are transformed into the visual parameters height (X), width of the outer lip contour (Y) and protrusion (Z) based on five parameters of the 3D lip model. The audio parameter has 33 dimensions of 16-order MelCepstral coefficients, their delta coefficients and the delta log power.

# 5.3 Audio-Visual Speech Processing Project

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Date of last modification of the description***

June 12th, 2001.

## 5.3.1 References

***Web site***

The website of the project: http://amp.ece.cmu.edu/projects/AudioVisualSpeechProcessing/

## 5.3.2 Description

Within the projects an audio-visual data corpus was collected, which is available to the public. The corpus contains data from 10 subjects (7 males and 3 females) saying 78 isolated words commonly used for time, such as, "Monday", "February", "night", etc. Each word repeated 10 times.

The collecting procedure of the database included setting the recordings up in a soundproof studio to collect noise-free audio data, and use of controlled light and blue-screen background to collect the image data.

A SONY digital camcorder with tie-clip microphone was used to record the data on DV tapes. The data on DV tapes was transferred to a PC by the Radius MotoDV program and stored as Quicktime files. Since the data on the DV tapes was already digital, there's no quality loss occurred when it was transferred to PC.

Sample QuickTime videos can be found at:

http://amp.ece.cmu.edu/projects/AudioVisualSpeechProcessing/anne.ram

http://amp.ece.cmu.edu/projects/AudioVisualSpeechProcessing/betty.ram

http://amp.ece.cmu.edu/projects/AudioVisualSpeechProcessing/chris.ram

http://amp.ece.cmu.edu/projects/AudioVisualSpeechProcessing/gavin.ram

the QuickTime files have been converted to streaming RealVideo sequences with lower quality, to make online viewing easier.

Altogether the resource contains 100 such QuickTime files, each one containing one subject articulating all the words in the vocabulary. Each file is about 450MBytes. Both the DV raw data and the Quicktime files are available upon request and can be downloaded at: http://amp.ece.cmu.edu/projects/AudioVisualSpeechProcessing/download.htm

# 5.4 Facial Feature Recognition using Neural Networks

## 5.4.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Date of last modification of the description***

June 12th, 2001.

## 5.4.2 References

***Web site***

The project's own website: http://www.cs.berkeley.edu/~debevec/face_recognition.html

Report from the Facial Feature Recognition using Neural Networks Project: http://www.cs.berkeley.edu/~debevec/face_recognition_report.html

## 5.4.3 Description

The project was developed in the fall of 1992 by Paul Debevec and deals with the recognition of human faces. The project was designed for a class project in artificial intelligence. One stage of recognizing a face is to figure out how the face is oriented, and to do this one needs to know where the facial features are in the image. For the project, he implemented a neural network to locate the eyes, nose, and mouth of facial images ("mug shots") taken from a high school yearbook.

His first thought was to train the neural network by showing it an 8 × 8 pixel window around the left eye of some training image, and then influence it (via back-propagation) so that it should recognize this small image patch as a left eye. Then he would show it a right eye, then a nose, and then a mouth, and keep this up through the whole testing set until the weights in the network converged to stable values. He implemented this approach with 64 input units (for the 8 × 8 patch of grey values), 9 units in the hidden layer, and four output units, one for each feature.

He realized that the neural network would probably have a hard time telling the difference between a left eye and a right eye if it only got to see an 8 × 8 pixel region of the image. Thus, instead of sending in an 8 × 8 window around the feature, he sent in an 8 × 8 sub-sampled version of the log-polar map of the image centred on the feature. This had the effect of including detailed information from the local area about the feature, as well as coarser information about the rest of the image around the feature. In this way, the neural network had a chance of telling the difference between a left eye and right eye by noticing their location relative to the rest of the face.

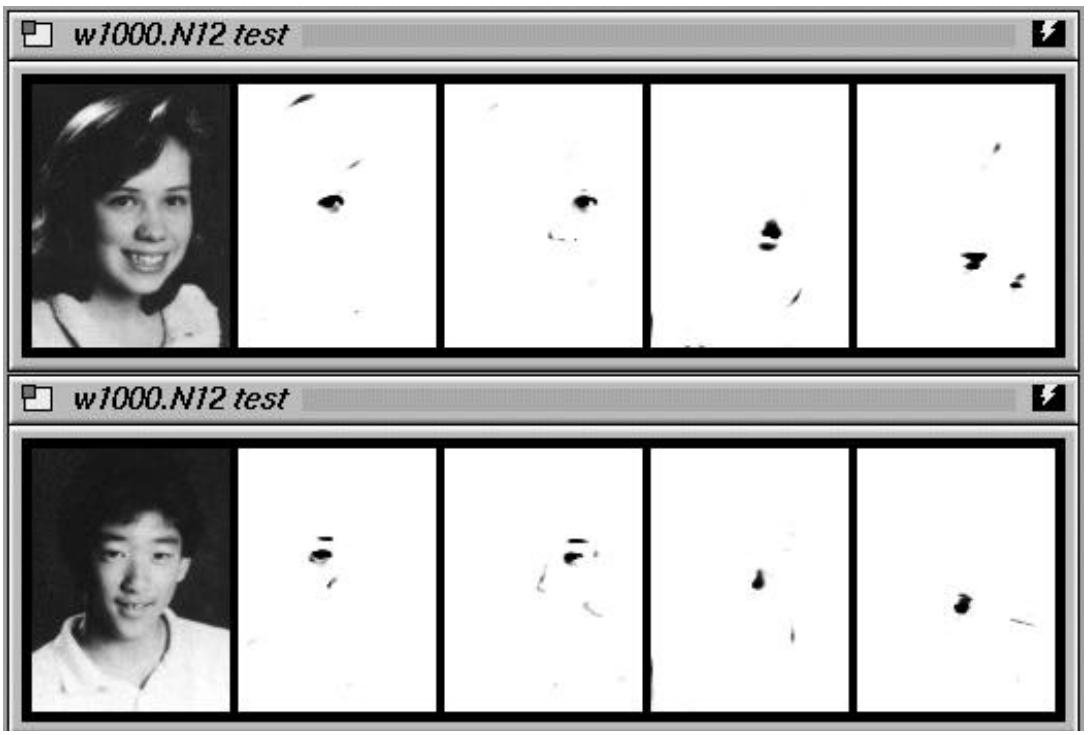


**Figure 5.4.1.** Two examples with the corresponding right eye, left eye, nose and mouth images. Taken from the high school yearbook mentioned below.

The images Paul Debevec used for the network were the underclass portraits of the 1987 University High School Yearbook. These were chosen to provide a reasonably homogeneous set of images that had similar contrast, lighting, and subject orientation. They were scanned in on a Macintosh using a Microtek flatbed scanner at 96 horizontal by 128 vertical pixels of resolution, in greyscale with 8 bits/pixel. This resolution was the minimum necessary to clearly see details of all the facial features. Paul Debevec used 101 of these images for the network, with 97 in the training set and four in the testing set.

More images can be found at the project's own website.

# 5.5 Image Database of Facial Actions and Expressions

## 5.5.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

June 27th, 2001.

## 5.5.2 References

***Web site***

http://mambo.ucsc.edu/psl/joehager/images.html

***Short description***

This database contains approximately 7000 still pictures which have been annotated with the Facial Action Coding System (FACS). The database was created for training neural networks to classify facial behaviours based on FACS (for more information on FACS see ISLE report 9.1, "Survey of Annotation Schemes and Identification of Best Practice")

***Illustrative sample picture or video file***

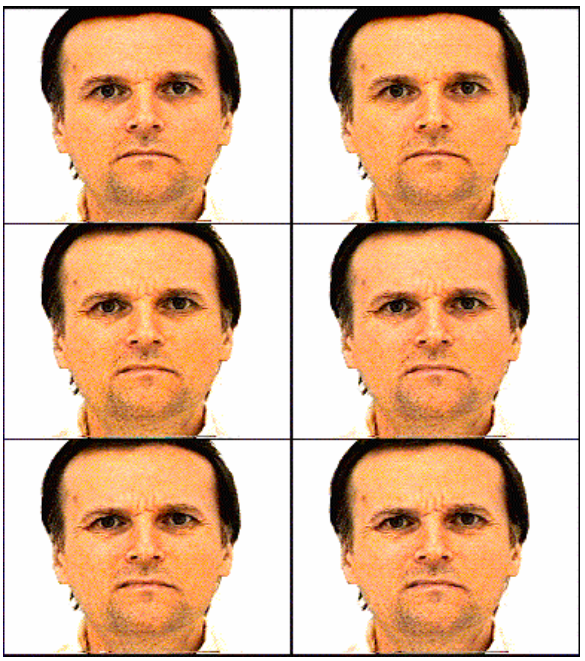

**Figure 5.5.1.** This illustrative example shows the AU's 4+6. The description of these is according to FACS: AU 4 Draws the inner corners of the eyebrows medially and down (action of corrugator and procerus muscles). AU 6 lifts the infraorbital triangle and produces crowsfoot wrinkles at the outer margins of the eye (action of orbicularis oculi, outer part).

***References to additional information on the reviewed resource***

None.

## 5.5.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

No information available.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

No information available.

***Which human body parts are visible in the resource?***

Only the faces of the subjects are visible in the resource.

***Which modalities are annotated?***

The annotated modalities are facial expression annotated with FACS. More information on FACS can be found in ISLE report 9.1, "Survey of Annotation Schemes and Identification of Best Practice".

***Which other modalities are available/visible in the resource but have not been annotated ?***

No information available.

### 5.5.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 5.5.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The format of the resource is a database of approximately 7000 still pictures taken from VHS video recordings. The video sequences have then been digitised into images of 24-bit colour, 640x480 pixels, and saved in TIFF format.

***How much data does the resource contain?***

The resource consist of a representation of subjects yielding between about 6 to 18 examples of the 150 different requested actions facial actions in the FACS system. Approximately 7,000 colour images are included in the database.

***Who created the resource and when?***

No information available.

***How was the resource created?***

Subjects contract specific facial muscles, singly and in combination, on request repeatedly and over multiple sessions. Few individuals can perform the requested actions accurately, and they must know FACS to understand the complicated requests and interpret feedback about what to do or not to do. The expressions depicted in the images are, thus, deliberately produced behaviours made on request.

The subjects' performances were videotaped (in NTSC format) under standard conditions, which include relatively flat lighting. The videotapes were carefully examined using FACS to locate performances that contain the correct facial actions of the FACS system and nothing else, and to determine the exact location of each frame to be digitised.

Then, sequences of individual frames from each correct performance were digitised in colour. To examine the motion or time course of the actions, several images that sample the increase in the strength of contraction of the muscles were digitised. Video frames showing low, medium, and high intensity action were digitised, if possible, with a frame grabber capturing full frame (1/30th sec.) stills. The time difference between the frames representing different intensities varies, depending on how fast the subject moved. These intervals were recorded in a relational database in SMPTE time code, along with other information about the images and the project. Two consecutive frames at each intensity level were digitised to represent movement over small time intervals.

***What is the application area?***

The application area is training of neural networks for classification of facial behaviour.

***What was the original purpose of creating the resource?***

This data resource was created in order to train neural networks to classify facial behaviours based on FACS.

## 5.5.6 Accessibility

***How does one get access to the resource?***

No information available.

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added in terms of codification of the facial expression using FACS, Facial Action Coding System developed by P. Ekman and W. Friesen.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on web information.

## 5.5.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The data has been used to train neural networks to classify facial behaviours based on FACS.
Since the data also contains some temporal data one could use it as precious data resource on the duration of muscle contraction and relaxation to produce an expression.

***Who used the resource so far/who are the target users of the resource?***

No information available.

***Is the resource language dependent or language independent?***

The resource is language independent since no speech is included.

## 5.5.8 Conclusion

***How interesting/important/high quality is the resource?***

The resource is carefully annotated using FACS.

***What do the authors regret ( if anything) not to have done while building the resource?***

No information available.

# 5.6 JAFFE Facial Expression Image Database

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Date of last modification of the description***

June 12th, 2001.

## 5.6.1 References

***Web site***

The website of the database: http://www.mic.atr.co.jp/~mlyons/jaffe.html

***References to additional information on the reviewed resource***

Michael J. Lyons, Julien Budynek and Shigeru Akamatsu. *Automatic Classification of Single Facial Images*. In IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (12): 1357-1362. 1999. Can be downloaded as pdf file at: http://www.mic.atr.co.jp/~mlyons/pub_pdf/michael.pdf

Michael J. Lyons, Shigeru Akamatsu, Miyuki Kamachi and Jiro Gyoba: Coding Facial Expressions with Gabor Wavelets. In proceedings from the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara Japan, IEEE Computer Society, pp. 200-205. 1998. Can be downloaded as pdf file at: http://www.mic.atr.co.jp/~mlyons/pub_pdf/fg98-1.pdf

## 5.6.2 Description

The JAFFE database consists of 213 images of Japanese female subjects posing 6 basic facial expressions as well as a neutral pose. Ratings on emotion adjectives are also available, free of charge, for research purposes.

## 5.6.3 Contact person

Please e-mail Michael Lyons (mlyons@mic.atr.co.jp) for information on how to obtain access to the images.

# 5.7 Multi-modal dialogue corpus

## 5.7.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Date of last modification of the description***

June 12th, 2001.

## 5.7.2 References

***Web site***

Description of the corpus: http://www.sccs.chukyo-u.ac.jp/ICCS/olp/p3-14/p3-14.htm

## 5.7.3 Description

This is a corpus of unscripted, task-oriented dialogues, which has been digitally recorded, and transcribed to support the study of multi-modal dialogue. The particularity of present corpus is collection of facial-expressions and gesticulation in the high quality video format. The corpus uses the Face Task and the Travel Task in which speakers must collaborate both verbally and nonverbally to achieve their goal in the task. In all, the corpus includes 9 dialogues (approximately 94 minutes). Although the corpus has not enough size, it manipulates as much as possible the following variables: task, familiarity and sexuality of speakers, eye contact between speakers. The collection and publication of the corpus was made by the Japan Electronic Industry Development Association as a part of the research on multi-modal dialogue processing technology.

## 5.7.4 Contact persons

Takuya Kaneko and Shun Ishizaki

Graduate School of Media and Governance

Keio University

5322 Endo, Fujisawa, Kanagawa, 252-8520 JAPAN

takuya@sfc.keio.ac.jp

ishizaki@sfc.keio.ac.jp

# 5.8 Photobook

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Date of last modification of the description***

June 12th, 2001.

## 5.8.1 References

***Web site***

A description of Photobook: http://vismod.www.media.mit.edu/~tpminka/photobook/

## 5.8.2 Description

Photobook is a tool for performing queries on image databases based on image content. It works by comparing features associated with images, not the images themselves.

Photobook is available free by FTP at ftp://whitechapel.media.mit.edu/pub/photobook/. It runs under the UNIX/Linux operating system. The distribution contains very little feature extraction code; the user has to provide it him/herself. In particular, no face recognition code is provided. For face recognition code, the user should look at ftp://whitechapel.media.mit.edu/pub/eigenfaces/ or get the file at ftp://whitechapel.media.mit.edu/pub/face-recognition.tar.Z. Send and e-mail to Alex Pentland (sandy@media.mit.edu) for more face recognition info. For more info on heavy-duty image retrieval software that runs out of the box, check out the commercial products listed in the Northumbria report at http://www.unn.ac.uk/iidr/research/cbir/report.html - Heading34.

# 5.9 Video Rewrite

## 5.9.1 Description header

***Main actor***

UROME: Catherine Pelachaud (cath@dis.uniroma1.it) and Isabella Poggi (poggi@uniroma3.it)

***Date of last modification of the description***

June 12th, 2001.

## 5.9.2 References

***Web site***

Short description of Video Rewrite: http://graphics.stanford.edu/~bregler/videorewrite/

***Illustrative sample picture or video file***

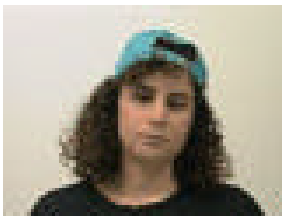

**Figure 5.9.1.** Still picture taken from the above website. The corresponding animation, which demonstrate the quality of the reconstructions from the Video Rewrite process can be found at: http://graphics.stanford.edu/~bregler/videorewrite/e_train.mov and requires a QuickTime Player to be viwed. All of the animations found at the website, http://graphics.stanford.edu/~bregler/videorewrite/, are synthesized from a video database of the subject speaking, and from new audio. Video Rewrite automatically rearranges the mouth and chin images, and stitches them into a background video. The resulting video shows the subject mouthing the words they never said.

## 5.9.3 Description

Video Rewrite uses existing footage to create automatically new video of a person mouthing words that s/he did not speak in the original footage. This technique is useful in movie dubbing, for example, where the movie sequence can be modified to sync the actors' lip motions to the new soundtrack.

Video Rewrite automatically labels the phonemes in the training data and in the new audio track. Video Rewrite reorders the mouth images in the training footage to match the phoneme sequence of the new audio track. When particular phonemes are unavailable in the training footage, Video Rewrite selects the closest approximations. The resulting sequence of mouth images is stitched into the background footage. This stitching process automatically corrects for differences in head position and orientation between the mouth images and the background footage.

Video Rewrite uses computer-vision techniques to track points on the speaker's mouth in the training footage, and morphing techniques to combine these mouth gestures into the final video sequence. The

new video combines the dynamics of the original actor's articulations with the mannerisms and setting dictated by the background footage.

Video Rewrite is the first facial-animation system to automate all the labelling and assembly tasks required to resync existing footage to a new soundtrack.

# 6 Gesture Data Resources

This introduction covers chapters 6 and 7 both of which have a primary focus on gesture data resources. Chapter 6 presents gesture data resources while Chapter 7 describes lesser known gesture data resources for which we have found little information.

Several descriptions of data resources were found on the web while others were found from paper proceedings of specialised conferences. None of the resources were available/downloadable from the web. In all descriptions, information is provided on how to contact the creators of a particular data resource and ask them how to get access to it.

Gesture data resources have been created for a number of different purposes, including, but not limited to:

- study of communicative and multimodal behaviour;
- multimodal and natural interactive systems specification and development;
- creation of synthetic characters;
- training of various recognisers.

Most of the resources are dynamic. Depending of the purpose, the data resource may include, e.g., human-human face-to-face communication or human-system interaction involving various modalities.

## 6.1 ATR Multimodal human-human interaction database

### 6.1.1 Description header

***Main actor***

IMS: Ulrich Heid (uli@IMS.Uni-Stuttgart.DE)

***Verifying actor***

LIMSI: Jean-Claude MARTIN (martin@limsi.fr)

***Date of last modification of the description***

August 4th, 2001 (nakamura@slt.atr.co.jp has been contacted by email on August 8th)

### 6.1.2 References

***Web site(s)***

Contact nakamura@slt.atr.co.jp

***Short description***

A collection of two-people conversation data which contains speech, movie, motion data (3D positions for 18 body parts) and speech transcripts.

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Nakamura, S. et al: Multimodal Corpora for Human-Machine Interaction Research. Proceedings of ICSLP, Volume IV, pp. 25-28, 2000.

And also:

Hayamizu, S., Hasegawa, O., Itou, K., Sakaue, K., Tanaka, K., Nagaya, S., Nakazawa, M., Endoh, T., Togawa, F., Sakamoto, K. and Yamamoto, K.: RWC Multimodal Database for Interactions for Integration of Spoken Language and Visual Information. Proceedings of ICSLP, pp. 2171-2174, 1996.

Hayamizu, S., Nagaya, S., Watanuki, K., Nakazawa, M., Nobe, S. and Yoshimura, T.: A Multimodal Database of Gestures and Speech. Proceedings of Eurospeech, pp. 2247-2250, 1999.

Kiyama, J., Watanuki, K. and Togawa, F.: Multimodal Interaction Database and Analysis Environment. Proceedings of 1997 RWC Symposium, pp. 23-30, 1097.

Seki, S. et al: Multimodal Agent Interface for Communication. Proceedings of 2000 RWC Symposium, pp. 135-138, 2000.

### 6.1.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

28 different humans have been recorded in the whole resource.

***How many humans are recorded at the same time?***

Two subjects are visible in the same frame.

***What is their profile?***

The speakers know each other, so that their speaking style is friendly, natural and interactive. No other information available.

***Which human body parts are visible in the resource?***

The face and upper body of each subject is visible in the resource.

***Which modalities are annotated?***

- Available modalities:
  - Raw data:
    - speech signal
    - digitised movie
    - data on the position of 18 points of the body,
      taken within the 3-dimensional space, by synchronized cameras
  - Thus the following modalities are available:
    - speech

- transcribed speech (indirectly, by use of the transcriptions produced in the analysis step)
- gesture (to be interpreted from the raw data)

- Annotated modalities:
  - From audio data, intensity and pitch of speech are calculated
  - Speech transcripts
  - From motion data,
    position and direction of the following is calculated:
    - right and left arm
    - right and left upper arm
    - right and left hand
    - head
    - body

***Which other modalities are available/visible in the resource but have not been annotated?***

No information available.

### 6.1.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 6.1.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The files are organised in a database.

***How much data does the resource contain?***

The resource contains 14 dialogues, 28 persons Each dialogue lasts 10 minutes per person.

***Who created the resource and when?***

The resource has been created by a consortium of ATR, Sharp and Tsukuba Electrotechnical Laboratories. An exact date is not indicated. The level of development described here corresponds to the status of the project at the point of publication.

***How was the resource created?***

Two persons are seated in different locations and talk to each other through 29 inch monitors. One camera records the upper body motions, the other camera records zooming out body motions. Each movie is recorded by analogue VTRs (betacam) controlled by a PC. The recorded movie data are then converted to a digital movie data. The image size of a movie is 320x240 and the sampling rate is 30 frames per seconds. The audio sampling rate is 16kHz. Motion sensing devices measures 3D positions

for the markers attached to the 18 positions on the upper body including head top, shoulders, elbows and hands. Each marker is taken by 5 infra-red cameras for each subject. The marker positions are traced optically at 60 frames / sec. From audio data, intensity and pitch of speech are calculated, and speech is transcribed.

***What is the application area?***

None

***What was the original purpose of creating the resource?***

The purpose of creating the database is to provide a source for analysing the relation between various modalities observed in one individual, and also the interaction between the two.

### 6.1.6 Accessibility

***How does one get access to the resource?***

No information available.

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Added value is provided by the transcriptions of speech, and by the
interpretative analyses of the raw data:

- From audio data, intensity and pitch of speech are calculated
- Speech transcripts
- From motion data,
  position and direction of the following is calculated:
  - right and left arm
  - right and left upper arm
  - right and left hand
  - head
  - body

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

No. The review is based on the references listed above.

### 6.1.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The purpose of the creating the resource is the modelling of face to face interaction in conversation, as an input to better man-machine communication in advanced interfaces.

***Who used the resource so far/who are the target users of the resource?***

With its rich motion data and the fact that all events are strictly time-aligned, any kind of research on the interaction of speech and gesture, as much as it can be derived from the data mentioned above, seems to be possible. First results of the use of the data seem to go into the direction of the identification of correlations between motion and speech cites the example on the strong correlation found in the data between turn taking and head motion): cross-modality interaction.

***Is the resource language dependent or language independent?***

No information available.

### 6.1.8 Conclusion

***How interesting/important/high quality is the resource?***

The resource is interesting for ISLE both methodologically and practically. Designed for the study of the correlation between speech and gesture, speech and bodily posture, it is a very good example of a resource for cross-modality research in the dialogue area. The language (Japanese) and the unclear access conditions may relativize its practical usability. It would however be advisable, in a cooperative effort (European/Japanese) to check possibilities to access the data and to use it, for example, for tests of the NITE tools.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.2 CHCC OGI Multimodal Real Estate Map

## 6.2.1 Description header

***Main actor***

LIMSI-CNRS : Jean-Claude MARTIN (martin@limsi.fr)

***Verifying actor***

IMS: Steve Berman (steve@ims.uni-stuttgart.de)

***Date of last modification of the description***

Tuesday, 26 June 2001 (Sharon Oviatt has been contacted by email on that date)

## 6.2.2 References

***Web site(s)***

Sharon Oviatt's website: http://www.cse.ogi.edu/CHCC/Personnel/oviatt.html
A list of downloadable publications is available as well as the description of some of the projects.
User-centered Modeling for Spoken Language and Multimodal Interfaces by Sharon Oviatt:
http://www.cse.ogi.edu/CHCC/Papers/sharonPaper/Ieee/Ieee.html

***Short description***

The resource is a video corpus of pen and speech human-computer interactions in the framework of a real estate map application.

***Illustrative sample picture or video file***

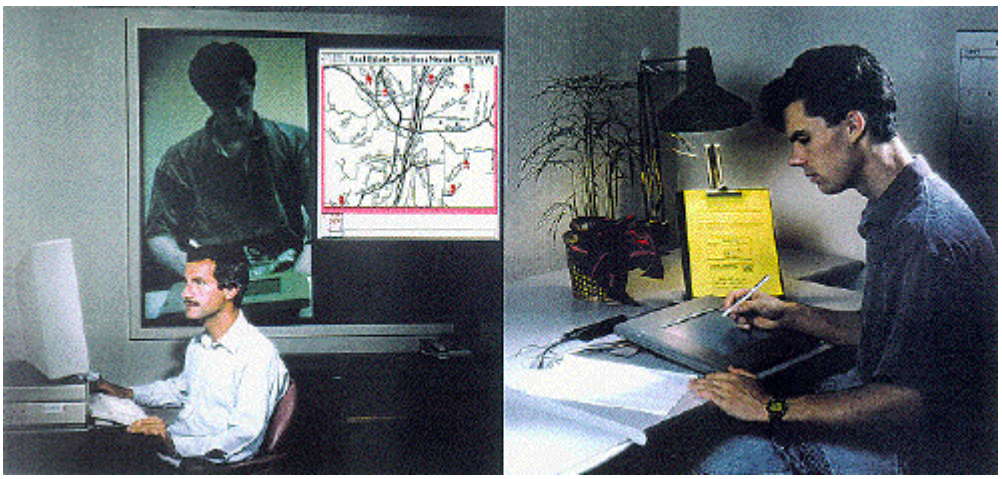

**Figure 6.2.1.** Examples of users using the system.

***References to additional information on the reviewed resource***

Oviatt, S. L. & Kuhn, K. Referential features and linguistic indirection in multimodal language, Proceedings of the International Conference on Spoken Language Processing, 1998.

Oviatt, S. L. (1996). User-centered modeling for spoken language and multimodal interfaces. In IEEE Multimedia, 3 (4) 26-35.

### 6.2.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

18 different subjects have been recorded in the whole resource.

***How many humans are recorded at the same time?***

Only one subject is visible in the same frame.

***What is their profile?***

The subjects were paid volunteers. No other information available.

***Which human body parts are visible in the resource?***

Head, hand, arms, upper body.

***Which modalities are annotated?***

The annotated modalities are speech and handwritten input.

***Which other modalities are available/visible in the resource but have not been annotated?***

Facial expression, hand-gesture, gaze, upper body posture are also available as modalities but have not been annotated.

### 6.2.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

A graphical screen is visible and is used by the subjects in the resource.

### 6.2.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

No information available.

***How much data does the resource contain ?***

The resource contains speech-only, pen-only and pen/speech input of respectively 216 tasks, approximately, 2700 utterances and 12000 words.

***Who created the resource and when?***

The resource was among others created by Sharon Oviatt. The date has not been specified.

***How was the resource created?***

The resource was created in a Wizard of Oz Setting. The subject's input was received by an informed assistant who performed the role of interpreting and responding as a fully functional system would.

***What is the application area?***

Service Transaction System (real estate selection).

***What was the original purpose of creating the resource?***

To compare the linguistic differences and relative ease of processing multimodal input compared with unimodal input.

### 6.2.6 Accessibility

***How does one get access to the resource?***

By contacting Sharon Oviatt, oviatt@cse.ogi.edu.

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Speech has been annotated. Handwriting has been automatically processed. Statistics have been computed on co-referring expressions, definite and indefinite reference, deictic expressions, linguistic indirection.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

No, the review is based on web information and the references listed above.

### 6.2.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for comparing linguistic features in multimodal and unimodal tasks for the same user.

***Who used the resource so far/who are the target users of the resource?***

The resource has been used by researchers from CHCC.

***Is the resource language dependent or language independent?***

The speech in the resource is in English.

### 6.2.8 Conclusion

***How interesting/important/high quality is the resource?***

A rich corpus of Human-Computer interaction including the drawing and writing modality via the pen.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.3 GRC Multimodal Dialogue during Work Meeting

## 6.3.1 Description header

***Main actor***

LIMSI-CNRS : Jean-Claude MARTIN (martin@limsi.fr)

***Verifying actor***

IMS: Steve Berman (steve@ims.uni-stuttgart.de)

***Date of last modification of the description***

7th of August 2001 (Creators were contacted by email on Thursday the 5th of April 2001, but no response was given.)

## 6.3.2 References

***Web site(s)***

The GRC web site: http://www.univ-nancy2.fr/RECHERCHE/Grc.html

***Short description***

The resource contains recordings and annotations of multimodal interaction (speech, hand, head and body gestures) between 3 engineers during a work meeting..

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Sannino, Annalisa. L'accomplissement interlocutoire et intergestuel d'une interaction en situation de travail in A.Trognon, and K. Kostulski. Communications interactives dans les groupes de travail. Collection language, cognition, interaction - Presses Universitaires de Nancy, 1998, pp. 123-155

## 6.3.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

Three different humans have been recorded in the whole resource.

***How many humans are recorded at the same time?***

Three different people are visible in the same frame.

***What is their profile?***

Three engineers who have to create a component of a mechanic system.

***Which human body parts are visible in the resource?***

The body, head, hands and arms of the subjects are visible in the resource.

***Which modalities are annotated?***

Speech, body and head posture, hand and arm gesture have been annotated.

***Which other modalities are available/visible in the resource but have not been annotated?***

Facial expressions (probably).

### 6.3.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 6.3.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The referenced paper mentions a video, but does not specify the format of this.

***How much data does the resource contain ?***

No information available.

***Who created the resource and when?***

The resource was created in 1995 by the GRC group.

***How was the resource created?***

No information available.

***What is the application area?***

The application area is work meeting analysis

***What was the original purpose of creating the resource?***

The original purpose was to study the patterns of multimodal communication during a work session about collaborative conception.

### 6.3.6 Accessibility

***How does one get access to the resource?***

By contacting Alain Trognon, Alain.Trognon@clsh.univ-nancy2.fr, Annalisa Sannino, Sannino@clsh.univ-nancy.fr or Dominique-Eve Weil, Dominique-Eve.WEIL@clsh.univ-nancy2.fr.

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added to the original resource in the form of informal and formal annotation of speech, gestures, environment.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

No, the review is based on web information and the reference listed above.

### 6.3.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for studying interaction and communication during working groups.

***Who used the resource so far/who are the target users of the resource?***

Several studies have used the corpus and are quoted in (Sannino 1998): Brassac, Grégori, Grosjean, Trognon and Kotulski).

***Is the resource language dependent or language independent?***

The speech in the resource is French.

### 6.3.8 Conclusion

***How interesting/important/high quality is the resource?***

An illustrative example of a resource on multimodal interaction between people during a meeting.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.4 LIMSI Multimodal Dialogues between Car Driver and Co-pilot Corpus

## 6.4.1 Description header

***Main actor***

LIMSI-CNRS: Jean-Claude MARTIN (martin@limsi.fr)

***Verifying actor***

IMS: Steve Berman (steve@ims.uni-stuttgart.de)

***Date of last modification of the description***

3rd April 2001 (authors of corpus have been contacted to check the description)

## 6.4.2 References

***Web site(s)***

An introduction to the corpus (in French):
http://www.limsi.fr/Individu/xavier/Articles/RapportInterneXBMDAideALaNavigation/NDXBMD.html

***Short description***

The resource contains multimodal dialogue between driver and co-pilot during real car driving tasks.

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Chalme, S., Briffault, X., Denis, and M., Gaunet, F.: Experiments For Designing Multimodal Dialogue Interfaces In Navigational Aid Systems : Real Versus Simulated Driving Situations. Dsc'99 (Driving Simulation Conference), Paris, Juillet 1999.

Denis M., Briffault, X.: Analyse Des Dialogues De Navigation À Bord D'un Véhicule Automobile. Le Travail Humain, Tome 63, N° 1/2000.

Briffault, X., and Denis, M.: Analyses D'un Corpus De Dialogues De Navigation À Bord D'un Vehicule Automobile. Technical Report, Limsi N°95-28, 1995.

### 6.4.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

54 subject have been recorded in the whole resource (27 drivers and 27 co-pilots).

***How many humans are recorded at the same time?***

2 humans are visible in the same frame.

***What is their profile?***

The subjects are people from LIMSI laboratory. No other information available.

***Which human body parts are visible in the resource?***

Face, arms and the hands of the subjects are visible in the resource.

***Which modalities are annotated?***

Some of the speech, gestures (hand, head, gaze), and part of the visual environment have been annotated.

***Which other modalities are available/visible in the resource but have not been annotated?***

Speech, gestures (hand, head, gaze), visual environment.

### 6.4.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 6.4.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The file types included in the resource are video Hi8.

***How much data does the resource contain ?***

The resource contains 27 dialogues of 90 minutes each.

***Who created the resource and when?***

The resource was created by Xavier Briffault and Michel Denis for Renault in 1995.

***How was the resource created?***

The experiment consisted in video and audio recording "multimodal" communication exchanges between a driver and a co-driver in a real situation of guided navigation.

***What is the application area?***

The application area is car navigation systems

***What was the original purpose of creating the resource?***

The objective was to investigate the components of the dialogue between the driver and the co-driver, the communication modalities used and the strategies of direction giving in order to provide guidelines for the specifications of an interactive navigational aid system for cars.

## 6.4.6 Accessibility

***How does one get access to the resource?***

The 27 video tapes (one dialogue pr. tape) and the transcriptions are available at the LIMSI library. One can contact Xavier Briffault, briffault@limsi.fr for access.

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added in the form of annotations a proposed coding scheme.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on hands-on experience with the resource, web information and the references listed above.

## 6.4.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for the study of multimodal communication between a driver and a co-pilot in different setting (co-pilot knows the route / co-pilot uses a map…).

***Who used the resource so far/who are the target users of the resource?***

The main results are that the dialogues were considered as multimodal ones, for they make use of gestures and pointing into a map, descriptions produced by the co-drivers are essentially incremental (instructions are given gradually, as one goes along), concerning the strategies of expressing spatial relations, the system of reference most often used is deictic - only a few examples have been noted in which an intrinsic system of reference is clearly used. The following conclusions are made: the communication between the driver and the system should contain more than one modality, a real

dialogue between the driver and the system is necessary, the main use made of a verbal modality is an incremental route description.

***Is the resource language dependent or language independent?***

Language dependent (French).

### 6.4.8 Conclusion

***How interesting/important/high quality is the resource?***

Large data resource with systematic annotation and documentation

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.5 LIMSI Pointing Gesture Corpus (PoG)

## 6.5.1 Description header

***Main actor***

LIMSI-CNRS : Jean-Claude MARTIN (martin@limsi.fr)

***Verifying actor***

IMS: Steve Berman (steve@ims.uni-stuttgart.de)

***Date of last modification of the description***

5th April 2001 (authors have been contacted to check the description)

## 6.5.2 References

***Web site(s)***

Contact information and description of corpus (in French):http://www.limsi.fr/Individu/braffort/

***Short description***

The corpus is a video corpus of pointing gestures and "constrained" speech on a building map.

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Braffort, A. and Gherbi, R.: Video-Tracking and recognition of pointing gestures using Hidden Markov Models. IEEE INES'98. Budapest, 1998.

Gherbi, R., Braffort, A.: Pointing gesture interpretation in a multimodal context. 3rd International Gesture Workshop, Gif-sur-Yvette (France), Springer, 1999.

## 6.5.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

12 different humans have been recorded in the whole resource.

***How many humans are recorded at the same time?***

Only one subject is visible in the same frame.

***What is their profile?***

The profile of the subjects is that they are people from the LIMSI laboratory. No other information available.

***Which human body parts are visible in the resource?***

The head and the body of each subject is visible in the resource.

***Which modalities are annotated?***

Speech and pointing hand gesture on a map have been annotated.

***Which other modalities are available/visible in the resource but have not been annotated?***

No information available.

### 6.5.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 6.5.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The file types included in the resource are VHS video tapes.

***How much data does the resource contain ?***

The resource contains 16 different commands recorded for each of the 12 subjects.

***Who created the resource and when?***

The resource has been created by Annelies Braffort, braffort@limsi.fr, and Rachid Gherbi, gherbi@limsi.fr. No date has been specified.

***How was the resource created?***

Twelve subjects have been recorded while making pointing gestures on a map and speaking constrained commands (i.e. “who is in this office?”). The aim of this corpus was to identify the features (handshape, dynamics, …) of pointing gestures in a multimodal context in order to design a gesture recognition system (Braffort & Gherbi 1998, Gherbi & Braffort 1999). Furthermore the aim was to develop a working methodology that goes from the design to the evaluation of recognition systems dedicated to the human gestures and focused on applications in the area of human-computer multimodal interaction. The methodology description is based on the conception and the development of a system named PoG (Pointing Gesture). This system grabs, recognizes and interprets co-verbal deictic gestures. PoG was integrated in a multimodal application combining gesture and speech in order to illustrate the contributions of the simultaneous use of gesture and verbal modalities, so that it

offers to the user a more natural interaction with the machine. This study was carried out within the framework of the European project Esprit "Chameleon".

***What is the application area?***

The application area is information kiosks about a building.

***What was the original purpose of creating the resource?***

No information available.

## 6.5.6 Accessibility

***How does one get access to the resource?***

By contacting Annelies Braffort, braffort@limsi.fr, or Rachid Gherbi, gherbi@limsi.fr.

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

No information available.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on hands-on experience as well as web information and the references listed above.

## 6.5.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for specification of a recognition system.

***Who used the resource so far/who are the target users of the resource?***

No information available.

***Is the resource language dependent or language independent?***

Language dependent (French)

### 6.5.8 Conclusion

***How interesting/important/high quality is the resource?***

The original point in this resource is that speech was constrained and pointing gestures unconstrained.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.6 McGill University, School of Communication Sciences & Disorders, Corpus of gesture production during stuttered speech

## 6.6.1 Description header

***Main actor***

LIMSI: Jean-Claude MARTIN (martin@limsi.fr)

***Verifying actor***

NISLab: Malene Wegener Knudsen (mwk@nis.sdu.dk), Laila Dybkjær (laila@nis.sdu.dk) and Niels Ole Bernsen (nob@nis.sdu.dk)

***Date of last modification of the description***

9th of August 2001 (rachel.mayberry@mcgill.ca has been contacted on August 8th to check the description)

## 6.6.2 References

***Web site(s)***

Rachel Mayberry's own web site: http://www.mcgill.ca/scsd/faculty/mayberry/

Research interests, List of publications, contact information (rachel.mayberry@mcgill.ca)

***Short description***

A video corpus of gesture and stuttered speech during description of a cartoon.

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Mayberry, R.I. and Jaques, J.: Gesture production during stuttered speech: Insights into the nature of gesture-speech integration. In D. McNeill (Ed.): Language and Gesture: Window into Thought and Action (pp. 199-213). Cambridge: Cambridge University Press. Recorded. 2000.

## 6.6.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

12 different humans have been recorded in the whole resource (six stuttering and six non stuttering).

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

Five subjects were males and one was female; they ranged in age from 21 to 51 years with a mean of 36 years. Six additional subjects with no history of stuttering were matched by age, sex, and highest level of education of the subjects who stuttered (4 undergraduate degree and 2 high school diploma).

***Which human body parts are visible in the resource?***

No information available.

***Which modalities are annotated?***

Gesture, speech (including disfluencies) and temporal concordance between speech and gesture have been annotated.

***Which other modalities are available/visible in the resource but have not been annotated?***

Not described in the paper (maybe facial expression, body and head movement ?).

### 6.6.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None

### 6.6.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

No information available.

***How much data does the resource contain?***

No information available.

***Who created the resource and when?***

The authors of the references paper created the resource sometime before 2000.

***How was the resource created?***

The creation process is thoroughly described in the references paper.

***What is the application area?***

McNeill (1992) protocol (description of a cartoon).

***What was the original purpose of creating the resource?***

The original purpose was to study relations between gesture and stuttered speech

## 6.6.6 Accessibility

***How does one get access to the resource?***

Contact rachel.mayberry@mcgill.ca

***Is the resource available for free or how much does it cost?***

No information available

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Yes.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

No the review is based on the reference and web site listed above.

## 6.6.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for studying relations between gesture and stuttered speech.

***Who used the resource so far/who are the target users of the resource?***

The authors

***Is the resource language dependent or language independent?***

The resource is language dependent due to the speech being English.

## 6.6.8 Conclusion

***How interesting/important/high quality is the resource?***

This resource is an example of the usefulness of multimodal resources in studies on handicaps.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.7 MPI Experiments with Partial and Complete Callosotomy Patients Corpus

## 6.7.1 Description header

***Main actor***

MPI: Gijs van Elswijk (gijsve@mpi.nl), Peter Wittenburg (pewi@mpi.nl)

***Verifying actor***

MPI: Sotaro Kita (kita@mpi.nl)

***Date of last modification of the description***

02/05/01

## 6.7.2 References

***Web site(s)***

Not available.

***Short description***

The resource contains data from experiments with partial and complete split-brain patients.

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Lausberg, Hedda, & Kita, Sotaro: Hemispheric specialization in spontaneous gesticulation investigated in split-brain patients. Proceedings of ORAGE 2001. (submitted).

## 6.7.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

20 different humans have been recorded in the whole resource.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

The age of the subject is around 40/50. No other information available.

***Which human body parts are visible in the resource?***

The whole body of each subject is visible in the resource.

***Which modalities are annotated?***

Gesture has been annotated.

***Which other modalities are available/visible in the resource but have not been annotated?***

Due to the whole body being visible several other modalities are available in the resource.

### 6.7.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

A computer screen is visible and used by the subjects in the resource.

### 6.7.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The file types included in the resource are video; MPEG1. The files are organised in the MPI Gesture Database structure (Oracle), see DB-gesture note.

***How much data does the resource contain ?***

The resource contains recordings of 20 people subject and several (up to 20) MPEG movies for each person has been made.

***Who created the resource and when?***

The resource was created by Hedda Lausberg and Sotaro Kita in 2000.

***How was the resource created?***

The resource was created using a DV and MediaTagger.

***What is the application area?***

The application area is scientific research.

***What was the original purpose of creating the resource?***

The original purpose of creating the resource was to make scientific research.

### 6.7.6 Accessibility

***How does one get access to the resource?***

By contacting Sotaro Kita, Stephen C. Levinson, and Hedda Lausberg , hedda.lausberg@mpi.nl.

***Is the resource available for free or how much does it cost?***

Not available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added in the form of gesture annotations.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on hands on experience with the resource and the reference listed above.

### 6.7.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for collaborative research only.

***Who used the resource so far/who are the target users of the resource?***

The resource has been used by researchers at MPI for Psycholinguistics and their collaborators.

***Is the resource language dependent or language independent?***

Language dependent: Quebec French, American English

### 6.7.8 Conclusion

***How interesting/important/high quality is the resource?***

An original resource on gesture with split brain patients.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.8 MPI Historical Description of Local Environment Corpus

## 6.8.1 Description header

***Main actor***

MPI: Gijs van Elswijk (gijsve@mpi.nl), Peter Wittenburg (pewi@mpi.nl)

***Verifying actor***

MPI: Sotaro Kita (kita@mpi.nl)

***Date of last modification of the description***

02/05/01

## 6.8.2 References

***Web site(s)***

Not available.

***Short description***

The resource contains data from historical descriptions of local environment.

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Not available.

## 6.8.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

More than two subjects have been recorded in the whole resource.

***How many humans are recorded at the same time?***

2 or 3 subjects are visible in the same frame.

***What is their profile?***

The age of the subjects are above 30. No other information available.

***Which human body parts are visible in the resource?***

The whole body of each subject is visible in the resource.

***Which modalities are annotated?***

Speech and gesture (for Dutch, Italian, Lao) have been annotated.

***Which other modalities are available/visible in the resource but have not been annotated?***

Due to the whole body of each subject is visible several other modalities are available.

### 6.8.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 6.8.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The file types included are video; QuickTime with Cinepak compression for Dutch, Arrernte, and Italian, and MPEG1 for the others. The files are organised in the MPI Gesture Database structure (Oracle), see database-gesture note.

***How much data does the resource contain ?***

The resource contains 11 dialogues.

***Who created the resource and when?***

The resource was created by Sotaro Kita, David Wilkins, Jan Peter de Ruiter, Nick Enfield, Chiara Piccini and Isabella Rega in 1994.

***How was the resource created?***

The resource was created by using VHS-C for the Italian data and Hi-8 camera for the others and MediaTagger for the annotations.

***What is the application area?***

The application area is scientific research.

***What was the original purpose of creating the resource?***

The original purpose of creating the resource was to do scientific research.

### 6.8.6 Accessibility

***How does one get access to the resource?***

By contacting Sotaro Kita, kita@mpi.nl and Stephen C. Levinson. Nick Enfield, nick.enfield@mpi.nl for the Lao data.

***Is the resource available for free or how much does it cost?***

Not available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added in the form of speech and gesture annotations.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on hands-on experience with the resource.

### 6.8.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for collaborative research only.

***Who used the resource so far/who are the target users of the resource?***

The resource has been used by researchers at MPI for Psycholinguistics and their collaborators.

***Is the resource language dependent or language independent?***

Language dependent: Dutch, Italian, Japanese, Lao from Laos, Arrernte from Australia

### 6.8.8 Conclusion

***How interesting/important/high quality is the resource?***

Interesting for research on the multilingual aspects of speech and gesture.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.9 MPI Living Space Description Corpus

## 6.9.1 Description header

***Main actor***

MPI: Gijs van Elswijk (gijsve@mpi.nl), Peter Wittenburg (pewi@mpi.nl)

***Verifying actor***

MPI: Sotaro Kita (kita@mpi.nl)

***Date of last modification of the description***

02/05/01

## 6.9.2 References

***Web site(s)***

Not available.

***Short description***

The resource contains data from dialogues on living space description in German.

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Seyfeddinipur, Mandana, & Kita, Sotaro: Gesture and speech dysfluencies. Proceedings of the 2nd Annual Meeting of the Berkeley Linguistics Society. (to appear)

Seyfeddinipur, Mandana, & Kita, Sotaro: Gesture and dysfluencies in speech. Proceedings of ORAGE 2001. (submitted)

## 6.9.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

More than two different humans have been recorded in the whole resource.

***How many humans are recorded at the same time?***

2 different subjects are visible in the same frame.

***What is their profile?***

The subjects are aged 20-40. No other information available.

***Which human body parts are visible in the resource?***

The whole body of each subject is visible in the resource.

***Which modalities are annotated?***

Speech and gesture have been annotated.

***Which other modalities are available/visible in the resource but have not been annotated?***

Due to the whole body of each subject being visible in the resource several other modalities are available in the resource.

### 6.9.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 6.9.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The included file types are video; QuickTime with Cinepak compression. The files are organised in the MPI Gesture Database structure (Oracle), see DB-gesture note.

***How much data does the resource contain ?***

The resource contains around 11 Gigabyte.

***Who created the resource and when?***

The resource was created by Mandana Seyfeddinipur and Sotaro Kita in 2000.

***How was the resource created?***

The annotations were created using MediaTagger.

***What is the application area?***

The application area is scientific research.

***What was the original purpose of creating the resource?***

The original purpose of creating the resource was to do scientific research.

### 6.9.6 Accessibility

***How does one get access to the resource?***

By contacting Sotaro Kita, kita@mpi.nl, Stephen C. Levinson, and Mandana Seyfeddinipur, mandana.seyfeddinipur@mpi.nl.

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added in the form of speech and gesture annotations.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on hands-on experience with the resource and the references listed above.

### 6.9.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for collaborative research only.

***Who used the resource so far/who are the target users of the resource?***

The resource has been used by researchers at MPI for Psycholinguistics and their collaborators.

***Is the resource language dependent or language independent?***

Language dependent: German.

### 6.9.8 Conclusion

***How interesting/important/high quality is the resource?***

Interesting for the description of spatial living space.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.10 MPI Locally-situated Narratives Corpus

## 6.10.1 Description header

***Main actor***

MPI: Gijs van Elswijk (gijsve@mpi.nl), Peter Wittenburg (pewi@mpi.nl)

***Verifying actor***

MPI: Sotaro Kita (kita@mpi.nl)

***Date of last modification of the description***

02/05/01

## 6.10.2 References

***Web site(s)***

Not available.

***Short description***

The resource contains locally situated narratives from Australia and Mexico.

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Not available.

## 6.10.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

More than two different humans have been recorded in the whole resource.

***How many humans are recorded at the same time?***

Either two (Australia) or three (Mexico) humans are visible in the same frame.

***What is their profile?***

The profile of the subjects is male, age 40 and above and a child (in the video from Mexico).

***Which human body parts are visible in the resource?***

The whole body of each subject is visible in the resource.

***Which modalities are annotated?***

Speech and gesture have been annotated.

***Which other modalities are available/visible in the resource but have not been annotated?***

Due to the whole body of each subject being visible in the resource several other modalities are available.

### 6.10.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 6.10.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The included file type is video; QuickTime with Cinepak compression. The files are organised in the MPI Gesture Database structure (Oracle), see DB-gesture note.

***How much data does the resource contain ?***

The resource contains 2 dialogues, about 850 Megabytes in total.

***Who created the resource and when?***

The resource was created by Stephen Levinson in 1998.

***How was the resource created?***

The resource was created using a Hi-8 camera for recording and MediaTagger for the annotations.

***What is the application area?***

The application area is scientific research.

***What was the original purpose of creating the resource?***

The original purpose of creating the resource was to do scientific research.

### 6.10.6 Accessibility

***How does one get access to the resource?***

By contacting Sotaro Kit, kita@mpi.nl and Stephen C. Levinson or Penelope Brown, penelope.brown@mpi.nl for the Tzeltal data.

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added in the form of speech and gesture annotations.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on hands-on experience with the resource.

### 6.10.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for collaborative research only.

***Who used the resource so far/who are the target users of the resource?***

The resource has been used by researchers at MPI for Psycholinguistics and their collaborators.

***Is the resource language dependent or language independent?***

Language dependent: Guugu Yimithirr from Australia, Tzeltal from Mexico

### 6.10.8 Conclusion

***How interesting/important/high quality is the resource?***

Interesting for research on the multilingual aspects of speech and gesture.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.11 MPI Narrative Elicited by an Animated Cartoon "Canary Row" Corpus 1

## 6.11.1 Description header

***Main actor***

MPI: Gijs van Elswijk (gijsve@mpi.nl), Peter Wittenburg (pewi@mpi.nl)

***Verifying actor***

MPI: Sotaro Kita (kita@mpi.nl)

***Date of last modification of the description***

02/05/01

## 6.11.2 References

***Web site(s)***

Not available.

***Short description***

The resource contains speech, gesture and sign in a narrative elicited by an animated cartoon (Dutch, Dutch Sign Language).

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Kita, Sotaro, van Gijn, Ingeborg, & van der Hulst, Harry: Movement phases in signs and co-speech gestures, and their transcription by human coders. In Ipke Wachsmuth and Martin Froelich (Eds.): Gesture and Sign Language in Human-Computer Interaction, International. Proceedings from Gesture Workshop Bielefeld, Germany, September 17-19, 1997, Lecture Notes in Artificial Intelligence, Volume 1371 (pp. 23-35), Berlin: Springer-Verlag.

van Gijn, Ingeborg, Kita, Sotaro, & van der Hulst, Harry: How phonetic is the symmetry condition in sign language?. In van Heuven, Vincent J., van der Hulst, Harry G., & van de Weije, Jeroen M. (Eds.): Phonetics and Phonology - Selected Papers of the Fourth HIL Phonology Conference. (to appear)

### 6.11.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

4 different humans have been recorded for Dutch Sign Language, and 6 for Dutch.

***How many humans are recorded at the same time?***

Only one human is visible in the same frame.

***What is their profile?***

The profile of the subjects are that they are women, aged 25-50.

***Which human body parts are visible in the resource?***

The whole body of each subject is visible in the resource.

***Which modalities are annotated?***

Gesture and signs have been annotated.

***Which other modalities are available/visible in the resource but have not been annotated?***

Due to the whole body of each subject being visible in the resource several other modalities are available.

### 6.11.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 6.11.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The file type included in the resource is video; QuickTime with Cinepak compression. The files are organised in the MPI Gesture Database structure (Oracle), see DB-gesture note.

***How much data does the resource contain ?***

The resource contains 2.5 Gigabyte.

***Who created the resource and when?***

The resource was created by Sotaro Kita, kita@mpi.nl, Ingeborg van Gijn and Harry van der Hulst in 1997 and 1998.

***How was the resource created?***

The resource was created using a Hi-8 camera for the recordings and MediaTagger for annotations.

***What is the application area?***

The application area is scientific research.

***What was the original purpose of creating the resource?***

The original purpose for creating the resource was to do scientific research.

### 6.11.6 Accessibility

***How does one get access to the resource?***

By contacting Sotaro Kita, kita@mpi.nl and Stephen C. Levinson

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added in the form of gesture and sign annotations.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on hands-on experience with the resource and the references listed above.

### 6.11.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for collaborative research only.

***Who used the resource so far/who are the target users of the resource?***

The resource has been used by researchers at MPI for Psycholinguistics and their collaborators.

***Is the resource language dependent or language independent?***

Language dependent: Dutch, Dutch Sign Language.

### 6.11.8 Conclusion

***How interesting/important/high quality is the resource?***

Interesting for research in spoken and sign languages.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.12 MPI Narrative Elicited by an Animated Cartoon "Canary Row" Corpus 2

## 6.12.1 Description header

***Main actor***

MPI: Gijs van Elswijk (gijsve@mpi.nl), Peter Wittenburg (pewi@mpi.nl)

***Verifying actor***

MPI: Sotaro Kita (kita@mpi.nl)

***Date of last modification of the description***

02/05/01

## 6.12.2 References

***Web site(s)***

Not available.

***Short description***

The resource contains Japanese, Turkish and American English data with speech and gesture in a narrative elicited by an animated cartoon.

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Ozyurek, Asli, & Kita, Sotaro: Expressing manner and path in English and Turkish: Differences in speech, gesture, and conceptualization. In Hahn, M., & Stoness, S. C. (Eds.). Proceedings of the Twenty First Annual Conference of the Cognitive Science Society, pp. 507-512, 1999.

## 6.12.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

More than two different humans have been recorded in the whole resource.

***How many humans are recorded at the same time?***

2 subjects are visible in the same frame.

***What is their profile?***

The subjects are aged 20-30. No other information available.

***Which human body parts are visible in the resource?***

The whole body of each subject is visible in the resource.

***Which modalities are annotated?***

None.

***Which other modalities are available/visible in the resource but have not been annotated?***

Due to the whole body of each subject being visible in the resource several modalities are available.

### 6.12.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 6.12.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The included file type is video; QuickTime with Cinepak compression The files are organised in the MPI Gesture Database structure (Oracle), see DB-gesture note.

***How much data does the resource contain ?***

The resource contains around 4 Gigabytes.

***Who created the resource and when?***

The resource was created by Sotaro Kita and Asli Ozyurek in 1995 and 1998.

***How was the resource created?***

The resource was created using a Hi-8 camera for the recordings of the Japanese and Turkish data and a not specified camera for the American English data.

***What is the application area?***

The application area is scientific research.

***What was the original purpose of creating the resource?***

The original purpose for creating the resource was to do scientific research.

## 6.12.6 Accessibility

***How does one get access to the resource?***

By contacting Sotaro Kita, kita@mpi.nl and Stephen C. Levinson

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

No information available.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on a hands-on experience with the resource and the reference listed above.

## 6.12.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for collaborative research only.

***Who used the resource so far/who are the target users of the resource?***

The resource has been used by researchers at the MPI for Psycholinguistics and their collaborators.

***Is the resource language dependent or language independent?***

Language dependent: Japanese, Turkish and American English.

## 6.12.8 Conclusion

***How interesting/important/high quality is the resource?***

Interesting for research on the multilingual aspects of speech and gesture.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.13 MPI Narrative Elicited by an Animated Cartoon "Maus" and "Canary Row" Corpus

## 6.13.1 Description header

***Main actor***

MPI: Gijs van Elswijk (gijsve@mpi.nl), Peter Wittenburg (pewi@mpi.nl)

***Verifying actor***

MPI: Sotaro Kita (kita@mpi.nl)

***Date of last modification of the description***

02/05/01

## 6.13.2 References

***Web site(s)***

Not available.

***Short description***

The resource contains speech and gesture data from a narrative elicited by an animated cartoon (Dutch).

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Not available.

## 6.13.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

More than two different humans have been recorded in the whole resource.

***How many humans are recorded at the same time?***

Only one subject is visible in the same frame.

***What is their profile?***

The profile of the subjects are that they are women, aged 20-40.

***Which human body parts are visible in the resource?***

The whole body of each subject is visible in the resource.

***Which modalities are annotated?***

Speech, gesture and eye movement have been annotated.

***Which other modalities are available/visible in the resource but have not been annotated?***

Due to the whole body of each subject being visible in the resource several other modalities are available.

### 6.13.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 6.13.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The file type included in the resource is video; MPEG1. The files are organised in the MPI Gesture Database structure (Oracle), see DB-gesture note.

***How much data does the resource contain ?***

The resource contains around 2.5 Gigabyte.

***Who created the resource and when?***

The resource was created by Sotaro Kita and Marianne Gullberg from 2000-2001.

***How was the resource created?***

The resource was created using a Hi-8 camera for recording and MediaTagger for annotation.

***What is the application area?***

The application area is scientific research.

***What was the original purpose of creating the resource?***

The original purpose of the creating the resource was to do scientific research.

### 6.13.6 Accessibility

***How does one get access to the resource?***

By contacting Sotaro Kita, kita@mpi.nl, Stephen C. Levinson, and Marianne Gullberg, marianne.gullberg@mpi.nl.

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added in the form of speech and gesture annotations and eye movement registrations.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on a hands-on experience with the resource.

### 6.13.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for collaborative research only.

***Who used the resource so far/who are the target users of the resource?***

The resource has been used by researchers at MPI for Psycholinguistics and their collaborators.

***Is the resource language dependent or language independent?***

Language dependent: Dutch.

### 6.13.8 Conclusion

***How interesting/important/high quality is the resource?***

Includes eye-movements.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.14 MPI Natural Conversation Corpus

## 6.14.1 Description header

***Main actor***

MPI: Gijs van Elswijk (gijsve@mpi.nl), Peter Wittenburg (pewi@mpi.nl)

***Verifying actor***

MPI: Sotaro Kita (kita@mpi.nl)

***Date of last modification of the description***

02/05/01

## 6.14.2 References

***Web site(s)***

Not available.

***Short description***

The resource contains natural conversations in Lao and Japanese.

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Kita, Sotaro: Japanese ideology of conversation and its structural manifestations: A study of aiduchi and head nods. In Verschueren, Jef (Ed.): Language and Ideology: Selected Papers from the 6th International Pragmatics Conference Vol. 1, pp. 262-269. International Pragmatics Association, 1999.

Wilkins, David P.: Spatial deixis in Arrernte speech and gesture: The analysis of a species of composite signal as used by a Central Australian Aboriginal group. In Poesio, André, M., & Rieser, H. (Eds.). Proceedings of the Workshop on Deixis, Demonstration, and Deictic Belief, pp. 30-42, Utrecht: ESSLLI XI, 1999.

Wilkins, David P.: Why pointing with the index finger is not a universal (in socio-cultural and semiotic terms). In Sotaro Kita (Ed.): Pointing: Where Language, Culture, and Cognition Meet. Mawhaw, NJ: Lawrence Erlbaum. (to appear).

### 6.14.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

More than two different humans have been recorded in the whole resource.

***How many humans are recorded at the same time?***

In the Lao data several subjects are visible in the same frame.
In the Japanese data 3 or 4 subjects are visible in the same frame.

***What is their profile?***

The profile of the subjects in the Lao data is that they are of all ages. No other information available.
The profile of the subjects in the Japanese data is that they are a family. Aged 20 and above. No other information available.

***Which human body parts are visible in the resource?***

The whole body of each subject is visible in the resource.

***Which modalities are annotated?***

Gesture in the Lao data have been annotated.

***Which other modalities are available/visible in the resource but have not been annotated?***

Due to the whole body of each subject being visible in the resource several other modalities are available.

### 6.14.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 6.14.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The included file type is video; QuickTime with Cinepak compression for the Japanese data and MPEG1 for the Lao data. The files are organised in MPI Gesture Database structure (Oracle), see DB-gesture note.

***How much data does the resource contain ?***

The resource contains around 64 Lao dialogues and one Japanese dialogue.

***Who created the resource and when?***

The resource was created by Nick Enfield and Sotaro Kita in 1998.

***How was the resource created?***

The resource was created using Hi-8 camera for recording and MediaTagger for annotations.

***What is the application area?***

The application area is scientific research.

***What was the original purpose of creating the resource?***

The original purpose of creating the resource was to do scientific research.

### 6.14.6 Accessibility

***How does one get access to the resource?***

By contacting Sotaro Kita, kita@mpi.nl and Stephen C. Levinson or Nick Enfield, nick.enfield@mpi.nl for the Lao data.

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added in the form of gesture annotation of the Lao data.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on a hands-on experience with the resource and the reference listed above.

### 6.14.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for collaborative research only.

***Who used the resource so far/who are the target users of the resource?***

The resource has been used by researchers at MPI for Psycholinguistics and their collaborators.

***Is the resource language dependent or language independent?***

Language dependent: Lao, Japanese.

### 6.14.8 Conclusion

***How interesting/important/high quality is the resource?***

Rich set of Lao dialogues including annotated gestures but other modalities as head nod are visible.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.15 MPI Naturalistic Route Description Corpus 1

## 6.15.1 Description header

***Main actor***

MPI: Gijs van Elswijk (gijsve@mpi.nl), Peter Wittenburg (pewi@mpi.nl)

***Verifying actor***

MPI: Sotaro Kita (kita@mpi.nl)

***Date of last modification of the description***

02/05/01

## 6.15.2 References

***Web site(s)***

Not available.

***Short description***

The resource contains dialogues on naturalistic route description: Direction-indicating gestures in Ghana.

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Kita, Sotaro and Essegbey, James: Left-hand taboo on direction-indicating gestures in Ghana: When and why people still use left-hand gestures. In the proceedings of the conference "Gesture: Meaning and Use." (to appear).

Kita, Sotaro and Essegbey, James: Pointing left in Ghana: How a taboo on left hand influences gesture practice. Gesture. (to appear).

## 6.15.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

More than two different humans have been recorded in the whole resource.

***How many humans are recorded at the same time?***

2 or 3 subjects are visible in the same frame.

***What is their profile?***

The age of the subjects range from about 10 to 40 years. No other information available.

***Which human body parts are visible in the resource?***

The whole body of each subject is visible in the resource.

***Which modalities are annotated?***

Speech have been annotated.

***Which other modalities are available/visible in the resource but have not been annotated?***

Due to the whole body of each subject being visible in the resource several other modalities are available.

### 6.15.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 6.15.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The file type included is video; QuickTime with Cinepak compression and the files are organised in MPI Gesture Database structure (Oracle), see DB-gesture note.

***How much data does the resource contain ?***

No information available.

***Who created the resource and when?***

The resource was created by James Essegbey and Sotaro Kita in 1998.

***How was the resource created?***

The resource was created using a Hi-8 camera for the recordings and Mediatagger for the annotation.

***What is the application area?***

The application area is scientific research.

***What was the original purpose of creating the resource?***

The original purpose of creating the resource was to do scientific research.

## 6.15.6 Accessibility

***How does one get access to the resource?***

By contacting Sotaro Kita, kita@mpi.nl and Stephen C. Levinson

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added in the form of speech annotations.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on hands-on experience with the resource and the references listed above.

## 6.15.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for collaborative research only.

***Who used the resource so far/who are the target users of the resource?***

The resource has been used by researchers at the MPI for Psycholinguistics and their collaborators.

***Is the resource language dependent or language independent?***

Language dependent: Ewe from Ghana.

## 6.15.8 Conclusion

***How interesting/important/high quality is the resource?***

Interesting for research on the multilingual aspects of speech and gesture, and for people studying route descriptions.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.16 MPI Naturalistic Route Description Corpus 2

## 6.16.1 Description header

***Main actor***

MPI: Gijs van Elswijk (gijsve@mpi.nl), Peter Wittenburg (pewi@mpi.nl)

***Verifying actor***

MPI: Sotaro Kita (kita@mpi.nl)

***Date of last modification of the description***

02/05/01

## 6.16.2 References

***Web site(s)***

Not available.

***Short description***

The resource contains dialogues on naturalistic route descriptions in Japanese: Expressing turns at an invisible location in route description.

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Kita, Sotaro: Expressing turns at an invisible location in route direction: The interplay of speech and body movement. In Hess-Luettich Ernest, B. W., Mueller, Juergen, E., & van Zoest, Aart (Eds.): Sign & Space (Raum & Zeichen), pp. 160-172. Tuebingen: Gunter Narr. 1998.

Kita, Sotaro: Interplay of gaze, hand, torso orientation and language in pointing. In Kita, Sotara (Ed.): Pointing: Where language, culture, and cognition meet. Mahwah, NJ: Lawrence Erlbaum. (to appear).

## 6.16.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

More than two different humans have been recorded in the whole resource.

***How many humans are recorded at the same time?***

2 different subjects are visible in the same frame.

***What is their profile?***

The subjects age from 20-40 and are men and women. No other information available.

***Which human body parts are visible in the resource?***

The whole body of each subject is visible in the resource.

***Which modalities are annotated?***

Speech and gesture have been annotated.

***Which other modalities are available/visible in the resource but have not been annotated?***

Due to the whole body of each subject being visible in the resource several other modalities are available in the resource.

### 6.16.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 6.16.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The file type included is video; QuickTime with Cinepak compression and the files are organised in the MPI Gesture Database structure (Oracle), see DB-gesture note.

***How much data does the resource contain ?***

The resource contains 19 dialogues.

***Who created the resource and when?***

The resource was created by Sotaro Kita in 1995-96.

***How was the resource created?***

The resource was created using a Hi-8 camera for recording and MediaTagger for annotations.

***What is the application area?***

The application area is scientific research.

***What was the original purpose of creating the resource?***

The original purpose of creating the resource was to do scientific research.

## 6.16.6 Accessibility

***How does one get access to the resource?***

By contacting Sotaro Kita, kita@mpi.nl and Stephen C. Levinson

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added in the form of speech and gesture annotations.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on a hands-on experience with the resource and the references listed above.

## 6.16.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for collaborative research only **.**

***Who used the resource so far/who are the target users of the resource?***

The resource has been used by researchers at MPI for Psycholinguistics and their collaborators.

***Is the resource language dependent or language independent?***

Language dependent: Japanese.

## 6.16.8 Conclusion

***How interesting/important/high quality is the resource?***

Interesting for research on the multilingual aspects of speech and gesture, and for people studying route descriptions.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.17 MPI Traditional Mythical Stories Corpus

## 6.17.1 Description header

***Main actor***

MPI: Gijs van Elswijk (gijsve@mpi.nl), Peter Wittenburg (pewi@mpi.nl)

***Verifying actor***

MPI: Sotaro Kita (kita@mpi.nl)

***Date of last modification of the description***

02/05/01

## 6.17.2 References

***Web site(s)***

Not available.

***Short description***

The resource contains dialogues with features in traditional mythical stories, Yucatec from Mexico and Mopan from Belize.

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Kita, Sotaro, Danziger, Eve, & Stolz, Cristel: Cultural specificity of spatial schemas, as manifested in spontaneous gestures. In Gattis, Merideth (Ed.). Spatial Schemas in Abstract Thought. pp. 115-146. Cambridge, MA: MIT Press. 2001

## 6.17.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

3 different humans have been recorded for Mopan, and 3 for Yucatec.

***How many humans are recorded at the same time?***

2 or 3 subjects are visible in the same frame.

***What is their profile?***

The subjects are aged from around 40-50. No other information available.

***Which human body parts are visible in the resource?***

The whole body of each subject is visible in the resource.

***Which modalities are annotated?***

Speech and gesture have been annotated.

***Which other modalities are available/visible in the resource but have not been annotated?***

Due to the whole body of each subject being visible in the resource several other modalities are available.

## 6.17.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

## 6.17.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The file type included is video; MPEG1 for Lao and QuickTime with Cinepak compression and the files are organised in the MPI Gesture Database structure (Oracle), see DB-gesture note.

***How much data does the resource contain ?***

The resource contains 6 dialogues.

***Who created the resource and when?***

The resource was created by Sotaro Kita, Eve Danziger and Cristel Stolz in 1996-1999.

***How was the resource created?***

The resource was created using MediaTagger for the annotations.

***What is the application area?***

The application area is scientific research.

***What was the original purpose of creating the resource?***

The original purpose of creating the resource was to do scientific research.

### 6.17.6 Accessibility

***How does one get access to the resource?***

By contacting Sotaro Kita, kita@mpi.nl and Stephen C. Levinson

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Value has been added in the form of speech and gesture annotations.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on a hands-on experience with the resource and the references listed above.

### 6.17.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for collaborative research only.

***Who used the resource so far/who are the target users of the resource?***

The resource has been used by researchers at MPI for Psycholinguistics and their collaborators.

***Is the resource language dependent or language independent?***

Language dependent: Yucatec from Mexico and Mopan from Belize.

### 6.17.8 Conclusion

***How interesting/important/high quality is the resource?***

Interesting for research on the multilingual and multimodal aspects of narrative.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.18 MPI Traditional Mythical Stories with Sand Drawings Corpus

## 6.18.1 Description header

***Main actor***

MPI: Gijs van Elswijk (gijsve@mpi.nl), Peter Wittenburg (pewi@mpi.nl)

***Verifying actor***

MPI: Sotaro Kita (kita@mpi.nl)

***Date of last modification of the description***

02/05/01

## 6.18.2 References

***Web site(s)***

Not available.

***Short description***

The resource contains Arrernte narratives in sand of traditional mythical stories.

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Wilkins, David P.: Alternative representations of space: Arrernte narratives in sand. In Biemans, M., & van de Weijer, J. (Eds.): Proceedings of the CLS Opening Academic Year 1997/1998, pp. 133-164. Center for Language Studies, The Netherlands. 1997.

## 6.18.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

More than two different humans have been recorded in the whole resource.

***How many humans are recorded at the same time?***

1 to 4 different subjects are visible in the same frame.

***What is their profile?***

The subjects are aged from 50 and above. No other information available.

***Which human body parts are visible in the resource?***

The whole body of each subject is visible in the resource.

***Which modalities are annotated?***

None.

***Which other modalities are available/visible in the resource but have not been annotated?***

Speech, gesture, and sand drawing have been annotated.

### 6.18.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 6.18.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

The file type included is video; QuickTime with Cinepak compression. The files are organised in the MPI Gesture Database structure (Oracle), see DB-gesture note.

***How much data does the resource contain ?***

The resource contains 4 dialogues.

***Who created the resource and when?***

The resource was created David Wilkins in 1996-1998.

***How was the resource created?***

The resource was created using a Hi-8 camera for recording and MediaTagger for annotations.

***What is the application area?***

The application area is scientific research.

***What was the original purpose of creating the resource?***

The original purpose for creating the resource was scientific research.

### 6.18.6 Accessibility

***How does one get access to the resource?***

By contacting Sotaro Kita, kita@mpi.nl and Stephen C. Levinson

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

No information available.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on hands-on experience with the resource and the reference listed above.

### 6.18.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used for collaborative research only.

***Who used the resource so far/who are the target users of the resource?***

The resource has been used by researchers at MPI for Psycholinguistics and their collaborators.

***Is the resource language dependent or language independent?***

Language dependent: Arrernte.

### 6.18.8 Conclusion

***How interesting/important/high quality is the resource?***

One original feature is the annotation on sand drawing. Interesting for research on the multilingual and multimodal aspects of narrative.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.19 National Autonomous University of Mexico, DIME multimodal corpus

## 6.19.1 Description header

***Main actor***

LIMSI: Jean-Claude MARTIN (martin@limsi.fr)

***Verifying actor***

UNAM: Luis Alberto PINEDA CORTÉS (luis@leibniz.iimas.unam.mx)

***Date of last modification of the description***

16th of April 2001 (authors of resource have completed the description)

## 6.19.2 References

***Web site(s)***

Publications, samples of video and transcription: http://cic2.iimas.unam.mx/multimod/dime/

***Short description***

A video corpus of speech and mouse-gesture interaction between subjects and simulated system in the field of kitchen design.

***Illustrative sample picture or video file***

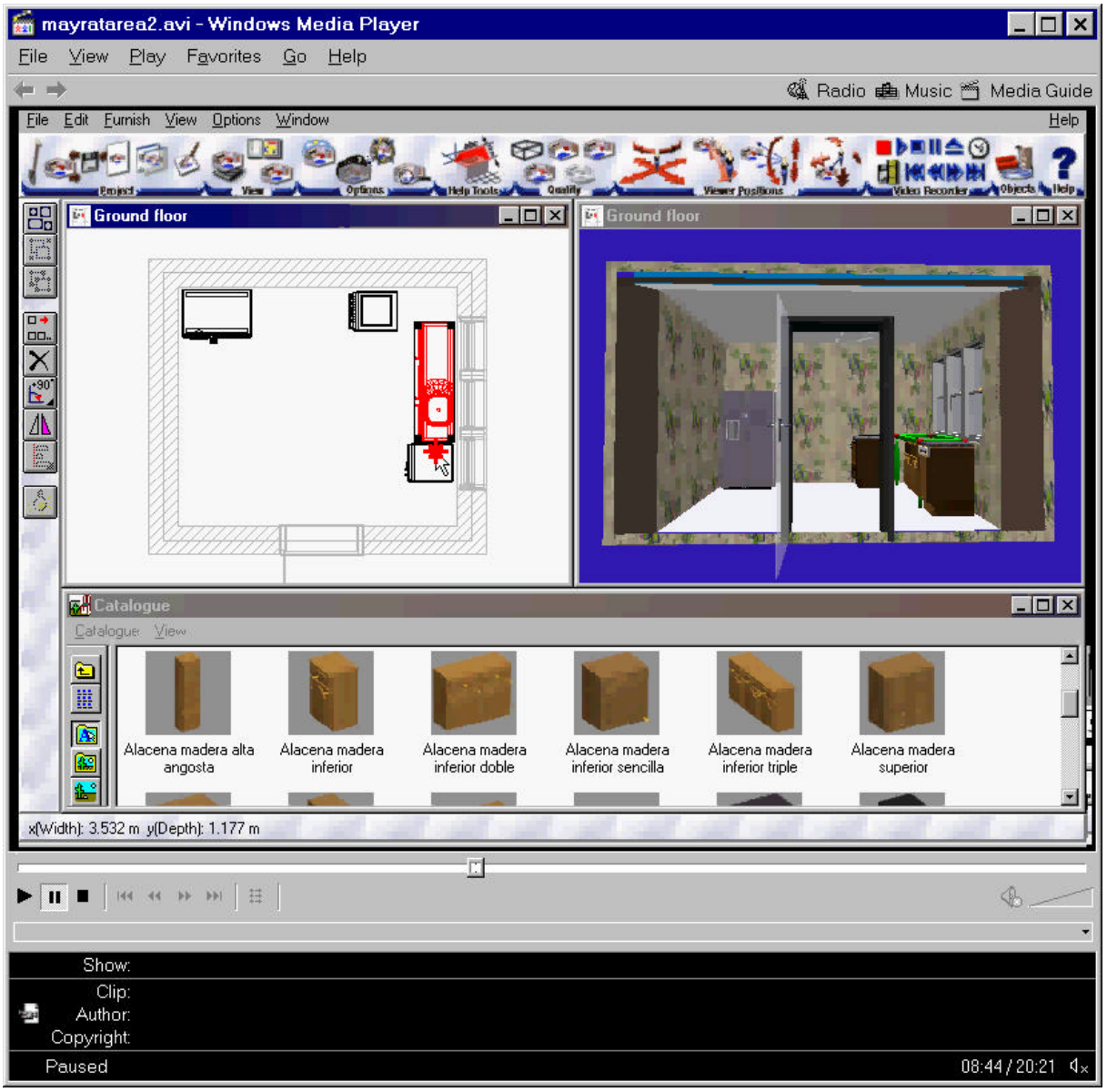


**Figure 6.19.1.** A video frame: one can hear the subject speaking while gesturing with the mouse on the graphical screen.

***References to additional information on the reviewed resource***

Luis Villaseñor, Antonio Massé, Luis A. Pineda:A Multimodal Dialogue Contribution Coding Scheme. 2000. Can be downloaded from:
http://www.mpi.nl/world/ISLE/documents/papers/villasenor_paper.pdf

Luis Villaseñor, Antonio Massé and Luis Pineda: Towards a Multimodal Dialogue Coding Scheme. Presented at CICLing-2000 Conference on Intelligent text processing and Computational Linguistics, February 13 to 19, 2000. México City, México. Talk and paper available at http://cic2.iimas.unam.mx/multimod/dime/index.html)

### 6.19.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

There were 15 experiments run with 15 different persons

***How many humans are recorded at the same time?***

Only one subject is visible in the same frame.

***What is their profile?***

Aged 30 on average, most of them were computer science related people

***Which human body parts are visible in the resource?***

None (the resource records speech, graphical actions with mouse gesture).

***Which modalities are annotated?***

Speech (orthographic transcriptions), speech acts (work in progress), referential expressions (work in progress).

***Which other modalities are available/visible in the resource but have not been annotated?***

None

### 6.19.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

A graphical screen (2D and a 3D views), loudspeaker (speech) and a mouse is visible in the resource.

### 6.19.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

Transcriptions, annotations (sgml); graphical environment video sessions (avi).

***How much data does the resource contain?***

The resource contains 31 dialogues (about 7 hours recorded) containing 185 utterances / 14 minutes on average per dialogue.

***Who created the resource and when?***

The work has been in development since October 1999 by the group of Multimodal Intelligent Systems of the Computer Science Department, IIMAS-UNAM

***How was the resource created?***

Wizard of Oz simulation (protocol detailed at http://cic2.iimas.unam.mx/multimod/dime/index.html)

The Laboratory consists of two attached rooms, the subject's room and the wizard's room. On the subjects' room there is a computer screen that shares the image with the computer screen on the wizard's side, two speakers where the subject can listen to the wizard, a microphone to talk to the system and a mouse to act over the screen and signal objects and regions. On the wizard's room there are the same objects for the same communicative purposes, besides there is a switch to inhibit mouse incoming signals from the user so this cannot interrupt when the system is working. There is also a multiplexer to share the same video signal between the two participants. There is another computer (CPU2) where the voice of the wizard is being recorded, while the voice of the subject and the graphical interaction of both system and user is recorded on CPU1. The graphical user interface is a commercial software [Alpha, 94] that provides us with the features specified on the Interaction Model. It lacked of a "start window" and a "system is busy window", we programmed these two on visual basic. To record the voices we used Wave Studio by Creative [Creative, 99] and to record the video of each session we used HyperCam by Hyperionics [Hyperionics, 99].

***What is the application area?***

Geometric design task (kitchen design).

***What was the original purpose of creating the resource?***

The main goal is to build and test an interactive multimodal Spanish spoken - graphics system to assist human-users in a geometric design task (kitchen design).

## 6.19.6 Accessibility

***How does one get access to the resource?***

By contacting luis@leibniz.iimas.unam.mx .

A dialogue example from the dime project can be found at http://cic2.iimas.unam.mx/multimod/dime.

***Is the resource available for free or how much does it cost?***

The resource files are freely distributed on request.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Yes. An extension to the DAMSL dialogue act markup scheme in order to consider multimodal events is proposed. The original Damsl tool, DAT, was modified to show the graphical environment in the annotation process. A snapshot of an annotation session can be found at: http://cic2.iimas.unam.mx/multimod/dime

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

No (some transcriptions are available on the web at http://cic2.iimas.unam.mx/multimod/dime/doctos/transcripciones/ given that the annotation work is still in progress)

## 6.19.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The corpus is been used to study multimodal referent resolution processes. It is also been used to study the multimodal discourse structures.

***Who used the resource so far/who are the target users of the resource?***

The corpus has been collected mainly for our project purposes. However it can also be used for the international computational linguistics community.

***Is the resource language dependent or language independent?***

The resource is language dependent: Spanish.

## 6.19.8 Conclusion

***How interesting/important/high quality is the resource?***

The Wizard of Oz experiments were specially designed to count with an intensive multimodal interaction in which speakers can freely refer to objects in a graphical environment. The importance of the resource lies on this rich multimodal interaction.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.20 National Center for Sign Language and Gesture Resources

## 6.20.1 Description header

***Main actor***

IMS: Steve Berman (steve@ims.uni-stuttgart.de)

***Verifying actor***

LIMSI-CNRS : Jean-Claude MARTIN (martin@limsi.fr)

***Date of last modification of the description***

7th of August 2001 References

***Web site(s)***

The National Center for Sign Language and Gesture Resources web site: http://www.bu.edu/asllrp/cslgr/

***Short description***

The National Center for Sign Language and Gesture Resources (NCSLGR) is an NSF-supported collaborative project between Boston University (C. Neidle and S. Sclaroff, principal investigators) and the University of Pennsylvania (D. Metaxas, N. Badler and M. Liberman, principal investigators). Current funding began October 1, 1998 and runs until September 30, 2002. Organizationally, the NCSLGR involves researchers from Boston University's Image and Vision Computing Group http://www.cs.bu.edu/groups/ivc/ and American Sign Language Linguistic Research Project (ASLLRP) http://www.bu.edu/asllrp/. The ASLLRP also includes projects investigating the grammar of ASL and developing tools for computer-aided research on sign language and visual-gesture language data in general (SignStream is the principal tool produced by this research).

The principal activities of the NCSLGR include the following (from http://www.bu.edu/asllrp/cslgr/):

- A facility for collection of video-based language data has been established, equipped with synchronized digital cameras to capture multiple views of the subject.
- A substantial corpus of American Sign Language [digital] video data is being be collected from native signers and made available in both compressed and uncompressed forms.
- Significant portions of the collected data are being linguistically annotated. The database of linguistic annotations will be made publicly available, along with the applications needed to access the database.
- Video data will be analysed using computer-based algorithms, with analysis and software made publicly available.

In addition to the NCSLGR database, a number of digital videos are available under the auspices of the ASLLRP. Digital video examples of ASL sentences discussed in ASLLRP publications are available on the Web and on CD-ROM. In addition, there is a SignStream data repository; additional

SignStream databases and associated digital video files are distributed on CD-ROM. Where appropriate, reference to these will be made below. But most of the information relevant to WP8 that I have been able to find specifically concerns the NCSLGR database, so that is the focus of this report. Most of the information cited here is from the NCSLGR website. (While there are links to websites at the University of Pennsylvania, which is participating in the NCSLGR, I have found no information there about the NCSLGR.)

***Illustrative sample picture or video file***


**Figure 6.20.1.** An example from the resource.

***References to additional information on the reviewed resource***

The website does not mention any papers, conference presentations or technical reports relating specifically to the creation of the NCSLGR database (there are a number of papers and presentations about the SignStream tool, however; these are listed in the ASLLRP website). The only project-external source of information is the following article from a Boston University newspaper (a link to this article appears in the SignStream website):

"New facility will aid in understanding sign language and human movement," by Eric McHenry. BU Bridge, Week of 21 January 2000,Vol. III, No. 20, pp. 1, 4. (Online at http://www.bu.edu/bridge/archive/2000/01-21/features3.html.)

The ASLLRP website lists a number of linguistic publications that make use of data from the NCSLGR and ASLLRP databases:

Neidle, C. and D. MacLaughlin (in press) The Distribution of Functional Projections in ASL: Evidence from Overt Expressions of Syntactic Features. In G. Cinque, ed., *Functional Structure in the DP and IP: The Cartography of Syntactic Structures, Vol. 1*, Oxford University Press.

Neidle, C., J. Kegl, D. MacLaughlin, B. Bahan, and R.G. Lee (2000) *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure*. Cambridge, MA: The MIT Press.

MacLaughlin, D., C. Neidle, B. Bahan, and R. G. Lee (2000) Morphological Inflections and Syntactic Representations of Person and Number in ASL. In M. Blondel and L. Tuller (eds), *Recherches linguistiques de Vincennes 29 : Langage et surdité*, 73-100.

Bahan, B., J. Kegl, R.G. Lee, D. MacLaughlin, and C. Neidle (2000) The Licensing of Null Arguments in American Sign Language. *Linguistic Inquiry 31:*1, 1-27.

Neidle, C., D. MacLaughlin, R.G. Lee, B. Bahan, and J. Kegl (1998) The Rightward Analysis of Wh-movement in ASL: A Reply to Petronio and Lillo-Martin 1997. *Language 74:*4, 819-831.

Neidle, C., B. Bahan, D. MacLaughlin, R.G. Lee, and J. Kegl (1998) Realizations of Syntactic Agreement in American Sign Language: Similarities between the Clause and the Noun Phrase. *Studia Linguistica 52*:3, 191-226.

Neidle, C., D. MacLaughlin, R.G. Lee, B. Bahan, and J. Kegl (1998). Wh-Questions in ASL: A Case for Rightward Movement. ASLLRP Report No. 6.

Neidle, C., J. Kegl, B. Bahan, D. Aarons, and D. MacLaughlin (1997) Rightward Wh-Movement in American Sign Language. In D. Beerman [sic], D. LeBlanc, and H. van Riemsdijk (eds), *Rightward Movement*. Amsterdam: John Benjamins, 247-278.

Neidle, C., D. MacLaughlin, J. Kegl, and B. Bahan (1996). Non-Manual Correlates of Syntactic Agreement in American Sign Language. ASLLRP Report No. 2.

Also listed on the ASLLRP site are publications that describe the SignStream and NCSLGR projects:

Neidle, C. (in press), SignStream™: A Database Tool for Research on Visual-Gestural Language. To appear in *Databases, Transcription and Tagging Tools,* a special edition of the *Journal of Sign Language and Linguistics* edited by P. Boyes Braem, T. Hanke, B. Bergman, and E. Pizzuto.

Neidle, C., S. Sclaroff, and V. Athitsos (in press), SignStream™: A Tool for Linguistic and Computer Vision Research on Visual-Gestural Language Data. *Behavior Research Methods, Instruments, and Computers.*

The Image and Video Computing Group web site lists publications of computer vision research that makes use of data collected in the NCSLGR:

La Cascia, M., and Sclaroff, S., Fast, Reliable Head Tracking under Varying Illumination, Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (CVPR), June, 1999.

La Cascia, M., Sclaroff, S., and Athitsos, V., Fast, Reliable Head Tracking under Varying Illumination: An Approach Based on Robust Registration of Texture-Mapped 3D Models, IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), 22(4), April, 2000.

Rosales, R., Athitsos, V., Sigal, L., and Sclaroff, S., 3D Hand Pose Reconstruction Using Specialized Mappings, International Conference on Computer Vision (ICCV) , July, 2001 (to appear).

Publications by the group at the University of Pennsylvania are listed at http://www.cis.upenn.edu/~cvogler/research/research.html#publications and include the following:

C. Vogler and D. Metaxas (1999). Parallel hidden Markov models for American Sign Language recognition. Proceedings of the IEEE International Conference on Computer Vision, pages 116-122, Kerkyra, Greece, 1999

C. Vogler, H. Sun and D. Metaxas (2000). A Framework for Motion Recognition with Applications to American Sign Language and Gait Recognition. Proceedings of the Workshop on Human Motion, Austin, TX.

C. Vogler and D. Metaxas (2001). A framework for recognizing the simultaneous aspects of American Sign Language. Computer Vision and Image Understanding 81, pp. 358-384.

## 6.20.2 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

No information available.

***How many humans are recorded at the same time?***

Mostly one:

- sets of sentences elicited to illustrate different possible sentence structures in ASL
- sets of sentences aimed specifically at computer vision researchers, containing a fixed vocabulary with signs occurring in many different contexts and word orders
- several short stories told in ASL

And one dialogue between two native signers (with two views of each signer: a front view and a close-up of the face)

***What is their profile?***

No information available.

***Which human body parts are visible in the resource?***

The upper body of each subject is visible in the resource.

***Which modalities are annotated?***

Since the video files contain no audio signal, only visual modalities are available, in particular, signing, gesture (other than sign language), and facial expression.

For those video sequences that are transcribed with SignStream (see also below), all of the above modalities can in principle be annotated; it is a stated goal of the NCSLGR to make such transcriptions available together with the video files (see above). Currently, about 30 of the video sequences are available from the NCSLGR website along with SignStream export data files and the SignStream database file itself (see below). In addition, between 450 and 500 utterances (about half of which were collected specifically with computer vision research in mind, as they make use of a relatively limited set of vocabulary in a variety of different syntactic contexts) are currently being transcribed and these video files and annotations will be released in the next few months. Other data that have already been collected include several brief stories and a 22-minute dialogue between two native ASL signers; those data are next in line for transcription and SignStream annotations will be released as soon as they are complete. The compressed versions of the video data just mentioned will be made available over the Internet from the NCSLGR Web site and on CD-ROM (through the LDC). Uncompressed versions in a variety of formats will also be made available (some in the same ways just listed, and others available on tape upon request). The SignStream data repository (http://www.bu.edu/asllrp/SignStream/burepos.html) contains further partially transcribed video sequences; some of these include a signed dialogue with multiple participants. In addition to two small sample databases that accompany the SignStream distribution (and are available as well via the data repository), there are 230 additional utterances for which SignStream annotations are provided on the SignStream version 2 CD-ROM, which also contains annotations for 32 sentences (3 views) collected in the NCSLGR. Additional SignStream databases for excerpts of ASL stories are available on CD-ROM (http://www.bu.edu/asllrp/cd.html).

***Which other modalities are available/visible in the resource but have not been annotated?***

From the examples downloaded on the web site, the annotation seem detailed including head position, head movement, gaze brows, gaze aperture …

## 6.20.3 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

No information available.

## 6.20.4 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

According to the NCSLGR website, "[t]he video data are made available in both uncompressed [AVI] and compressed [Quicktime, .mov] formats." The non-NCSLGR video sequences from the ASLLRP and the SignStream data repository are available only in the Quicktime format (it is noted that the video sequences recorded prior to the start of the NCSLGR are of lower quality than those made specifically for the NCSLGR).

***How much data does the resource contain ?***

The NCSLGR website does not provide a summary of statistical information about the video files, but by my calculations, at the URL listed below (under the Accessibility heading) there are 804 AVI (uncompressed) video files, 201 in each of four views, totalling just under 18.5 GB, with an average file size of approximately 23 MB. Of the Quicktime (compressed) video files there are only 799, since one of the directories contains 196 instead of 201 files; their total byte count is just under 660 MB, the average file size being approximately 825 KB. Note, again, that these are video files without annotation (but with some textual enhancement such as time codes and frame counts; see also below). The 30 available SignStream export data files are each smaller than 1 KB. I have not calculated the total duration of the video sequences, but if those I have examined (two seconds long each) are typical, then the total length of the sequences in each view would be about half an hour. Over 3.5 GB of additional QuickTime files (and corresponding AVI files totalling over 87 GB) are being prepared for distribution in the near future. This includes about 300 utterance-length video segments, plus 4 brief stories: 3 views each. This also includes a 22-minute dialogue between two native signers, with one front view and one close-up of the face for each. (All of the data will be distributed in QuickTime and AVI formats.) The SignStream software CD-ROM comes with two ASL datasets amounting to 230 utterances (this corresponds to QuickTime video files totalling 416 MB). There is also one ASL dataset for 32 sentences (3 views for each) collected in the NCSLGR (approx. 112 MB).The SignStream Databases vol. 1 CD-ROM contains SignStream transcriptions for excerpts of 4 ASL stories (total 430 MB). The SignStream data repository also contains SignStream annotations for 104 MB of QuickTime video of Nicaraguan Sign Language.

***Who created the resource and when?***

The digital video recordings of the NCSLGR database have been produced in the NCSLGR data collection facility described above; according to the website: "Data collection began in December 1999." The other video sequences available from the ASLLRP have been collected since 1995; these predate the creation of the NCSLGR data and are often of lower quality. The SignStream data repository includes additional videos from various sources: one source is DawnSign Press (also the source of the video for the SignStream Databases vol. 1 CD-ROM); another is from a project on Nicaraguan Sign Language (information and links are at the repository website http://www.bu.edu/asllrp/SignStream/burepos.html).

***How was the resource created?***

The data available from the NCSLGR are digital video recordings, without audio but with some textual enhancement (details given below). Here are three short descriptions of the data quoted from the NCSLGR website:

- A substantial corpus of American Sign Language (ASL) video data from native signers is being collected and made available.
- The signing is captured simultaneously from four different cameras, at a frame rate of 60 frames per second and at an image resolution of 648x484 (width x height). The downsampled versions are at 30 frames per second, and they have half the width and half the height of the original files, i.e. they are 312x242.
- Currently the four cameras are configured in the following way: Two cameras make a stereo pair, facing towards the signer, and covering the upper half of the signer's body. One camera, the colour one, faces towards the signer and zooms in on the head of the signer. One camera is placed on the side of the viewer, and covers the upper half of the signer's body.

The following technical description of the recording set-up is quoted from the NCSLGR website: The signing was captured simultaneously from four different cameras, at a frame rate of 60 frames per second and at an image resolution of 648x484 (width x height). At this moment we are not making available the original files, but only downsampled versions of them. The downsampled versions are at 30 frames per second, and they have half the width and half the height of the original files, i.e. they are 312x242. The hardware we use consists of: Four PCs, custom built by Vision1. Each PC has a 500MHz Pentium processor, 256MB memory and 64GB of hard drive storage. Four Kodak ES310 cameras. Each camera is connected to one PC. Three of the cameras give grey scale output, and one camera gives colour output. A sync source, that is used to synchronize the cameras together. The sync source is connected to each PC and to each camera. It can be set so that capturing is done in 30, 60, or 85 frames per second. An Ethernet switch, that allows the four PCs to communicate with each other. For frame grabbing, the PCs use the Bitflow RoadRunner software. For synchronized vide capturing by all PCs, we use IO Industries' VideoSavant software. One of the machines is designated as the "master" machine, and the other three are designated as "slaves". In order to capture a video sequence, we have to start the appropriate program on the "master" machine, and the appropriate client programs on the "slave" machines. Then, using the master machine we can specify how many frames we want to capture, and start the recording. The captured frames are stored on the hard drives in real time. With 64GB of hard drive storage available, we can record continuously for 60 minutes, at 60 frames per second, in all four machines simultaneously, at an image resolution of 648x484 (width x height).

***What is the application area?***

No information available.

***What was the original purpose of creating the resource?***

According to the NCSLGR website: "The goal of this project is to make available several different types of experimental resources and analysed data to facilitate linguistic and computational research on signed languages and the gesture components of spoken languages. This project will make available sophisticated facilities for data collection, a standardization of protocol for such collection, and large amounts of language data." Similarly, according to the SignStream website (http://www.bu.edu/asllrp/SignStream/index.html): "One goal of the SignStream project is to develop a large database of coded American Sign Language utterances" and "SignStream users are invited to contribute coded data to the Internet repository" along with the source video file. (Though, as noted, the repository also includes non-ASL data.)

## 6.20.5 Accessibility

***How does one get access to the resource?***

The NCSLGR website has links to ftp URLs where the uncompressed AVI and compressed Quicktime video files, in all four synchronized views, are downloadable. For information about access to files on CD-ROM and via FTP and WWW, see http://www.bu.edu/asllrp/cslgr/; significant quantities of new data will be released soon 32 of those utterances from the NCSLGR database (Quicktime format only), along with SignStream annotations of those utterances, are contained in the CD-ROM that comes with SignStream version 2.0, which is available from the ASLLRP. In addition, QuickTime files for 116 of those utterances (4 views each, with glosses but without SignStream annotations) are distribution on the ASLLRP Electronic Publications 2.0 CD-ROM. In the future, data will also be distributed on CD-ROM through the Linguistic Data Consortium. The video recordings in the SignStream data repository are downloadable from the repository website; most of these resources, as well as additional video files with SignStream annotations, are available on CD-ROMs, which, as noted above, also contain additional video recordings from the ASLLRP. See http://www.bu.edu/asllrp/cd/

"At this moment we are not making available the original files, but only downsampled versions of them."

***Is the resource available for free or how much does it cost?***

The video files accessible from the NCSLGR and ASLLRP websites are downloadable for free; the SignStream CD-ROM (which includes the SignStream transcription program) costs $25 per site, or $10 for an additional copy or for students. The additional video recordings available from the ASLLRP on two CD-ROMs (one of which includes corresponding SignStream database files as well) cost $10 each.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

SignStream transcriptions of these data (in both SignStream and text export format) are also being made public as the annotations are completed.

As noted above, the video recordings are enhanced with text information. The following technical description of the video frame and text stream formats is from the NCSLGR website:

Frame Format

[...E]ach frame contains two parts: The top part is the original image data, captured from one of the cameras. The bottom part (70 pixel rows) contains information about the frame, as follows:

First text row, from left to right:

- Date on which the frame was recorded.
- Time on which the frame was recorded.

Second text row, from left to right:

- Number of this frame in the sequence. Note that the first frame is number 1, and the frame numbers increase by 2 and not by 1. The reason is that, as mentioned earlier, these files are downsampled versions original sequences which were recorded at 60 frames per second.

- Number of milliseconds from beginning of recording. This number will probably not be of much use to anyone outside our lab. Note that the first frame of the sequence is not necessarily recorded at time 0.

Third row:

- Frame ID of this frame. This frame ID is guaranteed to be unique among all frames ever made available in this site. Two frames share the same frame ID if and only if they correspond to multiple viewpoints captured at the same instant. So, obviously, in each of the four AVI files that correspond to the four viewpoints of a sequence, the frame ID numbers are identical and in one-to-one correspondence.

Bottom 14 pixel rows

- At the bottom 14 pixel rows of each frame you will notice some black and white squares. Each square has a size of 7x7 pixels. This is a binary encoding of the frame ID. We use that encoding to make it easy to automatically recover the frame ID of a frame using trivial computer vision techniques. The squares are lined up in two rows of 32 squares each. Each row starts at the left edge of the image and extends 224 pixels to the right (enough for 32 squares). A white square stands for the digit 1 and a black square stands for the digit 0. The least significant digit is at the bottom right square. Digit significance increases from right to left and from bottom to top. So, for example, the 32nd least significant digit is represented by the leftmost square at the bottom row of squares.

In the AVI files, which are all uncompressed, black pixels have RGB values set to 0, and white pixels have RGB values set to 255. In the Quicktime files the values may have been modified. However, the average intensity value in each square should make it obvious whether the square is supposed to be a "white" or a "black" square.

Text Stream Format

Each AVI file, in addition to the video stream, which is visible if you play the file using any AVI player, also contains a text stream. This stream should not be confused with the bottom 70 pixel rows of the frames in the video stream, although it contains the same information as those 70 pixel rows.

In the text stream, there is one frame for each frame in the video stream. So, the first text frame contains information about the first video frame and so on. Each frame contains 74 bytes. The information in those bytes is in the following format:

- Bytes 0-19: The date and time the video frame was recorded, in text.
- Bytes 20-34: The 32 most significant digits of the number of milliseconds since the beginning of the recording, in text. Note that the first frame was not necessarily recorded at time 0.
- Bytes 35-49: The 32 least significant digits of the number of milliseconds since the beginning of the recording, in text.
- Bytes 50-53: A checksum of the data in the top 242 rows of the video frame. 4 bytes, in binary.
- Bytes 54-57: The time at which the frame was recorded, in binary, encoded as the number of seconds since the first second of the year 1970 (as returned by the standard C function "time").
- Bytes 58-61: The 32 most significant digits of the number of milliseconds since the beginning of the recording, in binary.
- Bytes 62-65: The 32 least significant digits of the number of milliseconds since the beginning of the recording, in binary.

- Bytes 66-73: The frame id of the video frame, in binary.

As you may have noted, there is no information in the text frame that is not visible in the bottom 70 rows of the video frame, except for the checksum. Text streams are added to the AVI files just to make it easier to access that information. People who download the Quicktime versions of the sequences have to recover all this information by decoding the bottom part of the frames. For the information in the text frame that is in binary, it is stored in PC byte order. To verify you are using the right byte order, simply check your values against what you see in the video frame.

[End of quotation]

Aside from this information coded onto the video, for some of the video files, the website contains links to their SignStream export data formats, which are tab separated frame-aligned text annotations of the video sequences. As noted above, the SignStream tool itself is available from the ASLLRP. A detailed description and review of this tool is available in ISLE deliverable D11.1. Specific details of the annotation scheme would seem appropriate for inclusion in D9.1.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

I downloaded and examined all four views of one of the AVI video files as well as the corresponding SignStream export data file. I have not had the opportunity to examine the SignStream tool itself, though I have browsed through the user manual (which is downloadable from the website).

## 6.20.6 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The primary purpose of the both the NCSLGR and ASLLRP/SignStream databases is to support research on sign language (specifically ASL, though at least the SignStream data repository is open to other sign language data). To date the material has been used primarily in linguistic investigations of ASL grammar (in particular syntactic and morphological properties; see the literature references below). The data and annotations that can be coded with SignStream also facilitate research on gesture more generally, and indeed on all visual modalities (including in particular facial expression). Moreover, the current release of SignStream also supports audio signals (though none of the data available within the NCSLGR and ASLLRP/SignStream databases contain audio signals). As noted above, according to the website: "The video data will also be analysed by various computer algorithms. The SignStream annotations of the data will be invaluable in evaluating [and training] such algorithms." The aim of this analysis is to "lay the foundation for automatic sign language recognition devices of the sort that already exist for oral speech" (cited from http://www.bu.edu/bridge/archive/2000/01-21/features3.html). The annotated corpora resulting form the NCSLGR initiative are also being collected for use by computer vision researchers, for use in development and validation of computer vision algorithms for ASL recognition. The detailed linguistic annotations can serve as "ground truth" for validation of new computer vision algorithms.

***Who used the resource so far/who are the target users of the resource?***

As noted, the NCSLGR and ASLLRP/SignStream have been used principally by linguists conducting research on the grammar of ASL (see references below). Target users include additionally computational linguists researching multimodal communication and computer scientists researching automatic recognition of hand motion and facial expression. The video data is also being used in machine vision research at Boston University and University of Pennsylvania. The particular emphasis has been on machine vision methods for automatic annotation and recognition of ASL communication,

as well as methods for indexing and retrieval of ASL video. The video data sets and their annotations are used in the development and quantitative evaluation of new machine vision techniques (see publications listed later in the report). Computer vision methods for automatic estimation of linguistically significant head and facial motion, as well as hand motion, orientation, and shape are being developed at BU.

The group at the University of Pennsylvania has also been working on head and face motion, as well as various aspects of ASL recognition (see, e.g., http://www.cis.upenn.edu/~cvogler/research/research.html). They are currently using the data for research into recognizing the facial expressions relevant to American Sign Language, as these form the bulk of the grammar of this language. To this end, they are extending their previous work on 3D deformable model-based face tracking to handle the expressions specific to sign language, such as raised eyebrows, puffed cheeks, and so on. In addition, they are extending this work to use statistical methods for cue integration to make the tracking more robust. The output of this method is the 3D parameterisation of the face model. Together with the topology of the model, this information can be used to recover the 3D coordinates of any point on the model. These coordinates, particularly the ones of prominent points on the face, such as the mouth, and the eyebrows, can be distilled into a feature vector suitable for hidden Markov model-based facial expression recognition.

***Is the resource language dependent or language independent?***

All of the video files made within the NCSLGR are of native signers of American Sign Language. The SignStream data repository, which is also available from the ASLLRP, includes two video files of Nicaraguan Sign Language. All other video recording available from the ASLLRP appear to be of ASL.

## 6.20.7 Conclusion

***How interesting/important/high quality is the resource?***

The NCSLGR database seems to be a quite interesting and useful resource for research on sign language in particular, and visual modalities (i.e., including gesture and facial expression) in general. The high quality of the digital video and its existence in four synchronized views are especially attractive in this regard. Its usefulness for these purposes depends on the availability of a suitable annotation tool, a requirement that SignStream is designed to meet (in fact, the development was just the reverse: the NCSLGR database has been established largely to exploit the capabilities of SignStream). Since the uncompressed video files consume substantial storage capacity, it is useful that the data are also made available in compressed (QuickTime) format. Likewise, the SignStream transcription tool is currently implemented only on Macintosh (although there is now a Java implementation under development); for the time being, the SignStream export data text files make the essential annotation data available platform independently. In general, the combination of the NCSLGR database and the SignStream tool seem, aside from the issue of portability, to fulfil a number of the ISLE desiderata.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.21 RWC Multimodal database of gestures and speech

## 6.21.1 Description header

***Main actor***

LIMSI: Jean-Claude MARTIN (martin@limsi.fr)

***Verifying actor***

LIMSI: Jean-Claude MARTIN (martin@limsi.fr)

***Date of last modification of the description***

4th of August 2001.

## 6.21.2 References

***Web site(s)***

The RWC web site: http://www.rwcp.or.jp/wswg/rwcdb/mm/index-english.html

***Short description***

The database consists of image data of human gestures and corresponding speech data for the research on multimodal interaction systems. The purpose of this database is to provide an underlying foundation for research and development of multimodal interactive systems. Our primary concern in selecting utterances and gestures for inclusion in the database was to ascertain the kinds of expressions and gestures that artificial systems could produce and recognize. Total 25 kinds of gestures and speech were repeated four times for the recording of each subject. The speech and gestures for a total of 48 subjects were recorded, converted into files and in the first version, the files for 12 subjects were recorded on CD-ROMs.

***Illustrative sample picture or video file***


**Figure 6.21.1.** An example taken from the web site.

***References to additional information on the reviewed resource***

Satoru Hayamizu, Shigeki Nagaya, Keiko Watanuki, Masayuki Nakazawa, Shuichi Nobe and Takashi Yoshimura: A Multimodal Database of Gestures and Speech. Proceedings of ESCA Eurospeech99, Budapest, Hungary, ISSN 1018-4074, Volume 5, Page 2247-2250. 1999.

## 6.21.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

48 subjects have been recorded in the whole resource.

***How many humans are recorded at the same time?***

2 humans are visible in the same frame.

***What is their profile?***

The project recruited people in their 20s and 40s as subjects, and tried to balance the number of women and men in each age category. A total of 48 subjects was recruited; grouped by age and gender.

***Which human body parts are visible in the resource?***

Four part split screen : subject face, subject from the waist up, partner from the waist up, monitor screen).

***Which modalities are annotated?***

Both image and speech data were tagged. Start and end position of each utterance were manually attached. Shape, place and orientation of each hand were described in the tag.

***Which other modalities are available/visible in the resource but have not been annotated?***

Face and upper body posture are also available in the resource.

## 6.21.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None

## 6.21.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

Time synchronized continuous video and audio files. In addition MPEG-1 format video files and AIFF format speech files are included for reference.

The video recording equipment was used to record a four-part split screen (showing the subject's face, subject from the waist up, partner from the waist up, and monitor screen) and a solitary image of the subject from the waist up.

Stereo sound was recorded with the subjects voice over the left channel and the partner's voice over the right channel at the same time the images were video recorded. DAT recordings were also made for reference purposes.

Files were created based on 3-to-4-second clips of continuous human action using SGI Indy.

***How much data does the resource contain ?***

The speech and gestures for a total of 48 subjects were recorded, converted into files, and saved on CD-ROMs (in the first version, the files for 12 subjects were recorded on CD-ROMs). Two CD-ROMs are used per subject. 25 kinds of gestures were repeated four times for the recording. The total of 25 kinds of gestures were video recorded including 8 gestures conveying agreement or disagreement, and 17 types of gestures indicating direction and relative size. Previously recorded gestures and speech were shown to the subjects; then they were instructed to imitate or re-enact both the spoken utterances and the observed gestures.

Gestures and utterances conveying agreement/disagreement (1-1 through 1-8 convey the following gestures: 1-1 conveys ordinary agreement; 1-2, 1-3, and 1-4 express strong agreement; 1-5 conveys ordinary disagreement; and 1-6, 1-7, and 1-8 indicate strong disagreement.)

1-1: One medium nod up and down at normal speed while saying hai (yes) one time.

1-2: Two medium nods at normal speed while saying hai(yes) one time.

1-3: One large nod at normal speed while saying hai(yes) one time.

1-4 One medium nod at vigorous speed while saying hai(yes) one time.

1-5: One medium shaking of the head back and forth at normal speed while saying iie(no) one time.

1-6: One large shaking of the head at normal speed while saying iie(no) one time.

1-7: One medium shaking of the head at rapid speed while saying iie(no) one time.

1-8: One medium shaking of the head and open hand back and forth at normal speed while saying iie(no) one time.

Gestures and utterances conveying direction and size

2-1: Convey upward motion while saying ue, ue(up, up).

2-2: Convey downward motion while saying shita, shita(down, down).

2-3: Convey motion to the right while saying migi, migi(right, right).

2-4: Convey motion to the left while saying hidari, hidari(left, left).

2-5: Convey inward motion while saying temae(closer).

2-6: Convey outward motion while saying muko(over there).

2-7: Indicate the size and shape of a briefcase while saying kona okisa no(one that's this size).

2-8: Convey emphasis while saying kore ga jyuyo nan desu yo(this is important).

2-9: Indicate length by holding two hands apart while saying kono gurai no okisa no sakana(a fish this big).

2-10: Indicate self by pointing at one's own chest while saying watashi (me).

2-11: Point downward toward the left with the right index finger while saying hidari(left).

2-12: Point downward toward the right with the right index finger while saying migi(right).

2-13: Trace an imaginary circle to the right around a perpendicular axis while saying migi wawari(clockwise).

2-14: Trace an imaginary circle to the left around a perpendicular axis while saying hidari wawari(counter clockwise).

2-15: Indicate stop by holding up one's hand palm outward while saying sutoppu(stop).

2-16: Draw the palms apart indicating enlargement or expansion while saying kakudai(expansion).

2-17: Bring the palms together indicating contraction while saying shukusho(contract).

***Who created the resource and when?***

The RWC program in 1996.

The authors of the 1999 paper are:

- Electrotechnical Laboratory
- Central Research Laboratory, Hitachi
- RWC Multimodal Sharp Laboratory

- Aoyama Gakuin University

***How was the resource created?***

Three recording methods were tested in preliminary experiments carried out in 1995 :

- Subjects freely improvised gestures and speech as indicated by a written script (large individual variations).
- Subjects first observe gestures and listen to speech that is pre-recorded, then mimic those gestures and speech (little variation).
- Subjects express themselves through gestures and speech as naturally as possible in interaction with another person.

Here we gave priority to the second approach where subjects mimic pre-recorded gestures because the acquisition cost was relatively low with respect to the amount of data collected and because this method produces little variation among the data.

The speech and gestures included in the database were first pre-recorded. Subjects then observed the gestures and speech, and were asked to mimic the gestures and speech for inclusion in the database. These were mirror-inverted imitations. Adopting this approach we could efficiently collect fairly consistent data that is well suited for recognition experiments.

- Image Recording Equipment Set-up

  The video recording equipment was used to record a four-part split screen (showing the subject's face, subject from the waist up, partner from the waist up, and monitor screen) and a solitary image of the subject from the waist up. (Just waist-up images of subjects were recorded on the CD-ROMs.)

  In positioning the camera and shooting angle, the main concerns were to have the subjects' eyes aligned as much as possible and to have the gestures fit within the confines of the screen.

  Shots of subjects from the waist up were filmed with a video camera were then recorded using a betacam. Output from the four-part split screen were recorded by separate betacams using a multiviewer. In addition, refresh signals were input to synchronize the various betacams.

  A video recording lighting set-up was used. Two 1,000-watt lights were used to illuminate for the subjects, and three 500-watt lights were used for background lighting. The subjects were not exposed to direct lighting. Indirect lighting reflected off white reflective surfaces (i.e., reflective paper) on the ceiling and floor was used. White reflective paper was also placed on the desks in front of the subjects to minimize facial shadows.

  A uniform green was used as the background colour for both subjects and partners to facilitate image recognition.

  In addition, markers were applied to indicate the approximate positions of the arms. For purposes of colour separation, red markers were used for the arms, green for the elbows, and white for the shoulders. Yellow shirts were worn. These arrangements made it possible to measure position through image processing without the use of elaborate mechanical equipment.

- Sound Recording Equipment Set-up

  Stereo sound was recorded with the subjects voice over the left channel and the partner's voice over the right channel at the same time the images were video recorded. DAT recordings were also made for reference purposes. A head-set type microphone was used in 1995 for the preliminary trial, but a lapel microphone was adopted for the actual recording to facilitate image processing. Pre-recorded speech was used instead of the speech of the partners. For monitoring purposes, a third party could listen in on the dialog through headphones. The recording state had to be left on in order for all of the speech to be picked up by the betacam. The recording state was also left on for the DAT to transmit both voices.

***What is the application area?***

No information available.

***What was the original purpose of creating the resource?***

The purpose of this database is to provide an underlying foundation for research that will lead to the development and deployment of multimodal interactive systems. More particularly, our objective is to build a speech and video database that can be shared among different research groups pursuing similar work that will promote research and development of multimodal interactive systems integrating speech and video data.

## 6.21.6 Accessibility

***How does one get access to the resource?***

See http://www.rwcp.or.jp/wswg/rwcdb/mm/riyou-e.htm

A major prerequisite is that the database is only to be used for research purposes. To obtain a copy of the database, please submit the prescribed contract agreement to the Real World Computing Partnership.

***Is the resource available for free or how much does it cost?***

Shipping and handling charge of 5,000 yen (includes 5% consumption tax)

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

No information available.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on the web site and reference listed above.

## 6.21.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource can be used for specification of a multimodal system.

***Who used the resource so far/who are the target users of the resource?***

The resource has been used by developers of multimodal systems.

***Is the resource language dependent or language independent?***

Language dependent (Japanese).

### 6.21.8 Conclusion

***How interesting/important/high quality is the resource?***

Interesting resource for builders of multimodal systems using image processing for gesture recognition.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.22 University of Chicago Origami Multimodal corpus

## 6.22.1 Description header

***Main actor***

LIMSI-CNRS: Jean-Claude MARTIN (martin@limsi.fr)

***Verifying actor***

IMS: Ulrich Heid (uli@IMS.Uni-Stuttgart.DE)

***Date of last modification of the description***

7th of August 2001 (author has been contacted on that date)

## 6.22.2 References

***Web site(s)***

The web site of the department of psychology at University of Chicago: http://psychology.uchicago.edu/research/research.htm

***Short description***

The resource is a video corpus of multimodal interactions between two subjects on a origami task.

***Illustrative sample picture or video file***

Not available.

***References to additional information on the reviewed resource***

Furuyama, N. (2000) Gestural interaction between the instructor and the learner in origami instruction. In "Language and gesture", McNeill D. (Ed), Language, culture and cognition, Cambridge University Press. P 99-117.

## 6.22.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

18 different humans have been recorded in the whole resource.

***How many humans are recorded at the same time?***

2 different subjects are visible in the same frame.

***What is their profile?***

Students. Average age was 20. Ten were male and eight female.

***Which human body parts are visible in the resource?***

Whole body of both subjects.

***Which modalities are annotated?***

Gestures were coded according to whether or not they were synchronised with speech. Then the same gestures were coded according to whether or not they were collaborative

***Which other modalities are available/visible in the resource but have not been annotated?***

Speech and body movements are available as modalities.

### 6.22.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

### 6.22.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

Not mentioned in the paper.

***How much data does the resource contain ?***

The resource contains 9 dialogues.

***Who created the resource and when?***

The resource was created by Furuyama, N. No date has been specified.

***How was the resource created?***

Subjects were shown videotaped instructions and a sheet of paper with a drawn version of the instructions, both of which showed how to make an origami balloon in a detailed step-by-step fashion. The instructors were permitted to ask the experimenter for help. The experiment took place in three steps. (1) a learning session for the instructor. (2) the instruction session during which the instructor explained the learner how to make the balloon. (3) the origami session where the learner attempted to make the balloon.

***What is the application area?***

The application area is origami.

***What was the original purpose of creating the resource?***

The original purpose was to study origami.

## 6.22.6 Accessibility

***How does one get access to the resource?***

No information available. (An attempt to contact the author at dmcn@ccp.uchicago.edu has been made without success.)

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Gestures were annotated on two dimensions: 1) collaborative/non collaborative, 2) whether the focal side of the instructor's gesture faced the learner or not.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

The review is based on web information and the references listed above.

## 6.22.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource has been used to study learner gestures (with and without speech, collaborative gestures), learner gestures in relation to instructor gestures)

***Who used the resource so far/who are the target users of the resource?***

The author of the above reference is the only known user of the resource.

***Is the resource language dependent or language independent?***

The resource is language dependent: American English

## 6.22.8 Conclusion

***How interesting/important/high quality is the resource?***

An original resource showing how one teaches/learns task related gestures.

***What do the authors regret (if anything) not to have done while building the resource?***

No information available.

# 6.23 VISLab Cross-Modal Analysis of Signal and Sense Data and Computational Resources for Gesture, Speech and Gaze Research

## 6.23.1 Description header

***Main actor***

IMS: Ulrich Heid (uli@IMS.Uni-Stuttgart.DE)

***Verifying actor***

LIMSI-CNRS: Jean-Claude MARTIN (martin@limsi.fr)

***Date of last modification of the description***

8th of August 2001 (authors have been contacted by Ulrich Heid but no answer was given)

## 6.23.2 References

***Web site(s)***

The KDI website: Cross-modal Analysis Signal and Sense Data and Computational Resources for Gesture, Speech and Gaze Research: http://vislab.cs.wright.edu/KDI/

***Short description***

Video Database with multistream time-based features/annotation. Project: Cross-modal Analysis of Signal and Sense: Data and Computational Resources for Gesture, Speech and Gaze Research.

The most frequently quoted video is one where one female describes her living space to an interlocutor.

***Illustrative sample picture or video file***

**Figure 6.23.1.** A data example taken from the above website.

***References to additional information on the reviewed resource***

David McNeill, Francis Quek, Karl-Erik McCullough, Susan Duncan, Nobuhiro Furuyama, Robert Bryll, Xin-Feng Ma, and Rashid Ansari: Catchment, Prosody, and Discourse. To appear in: Gesture. http://vislab.cs.wright.edu/Publications/McNetal01.html

## 6.23.3 Recorded human behaviour

***How many different humans have been recorded in the whole resource?***

No information available.

***How many humans are recorded at the same time?***

Two subjects are visible in each frame. (one subject who describes her living space and one interlocutor)

***What is their profile?***

The subjects are female. No other information available.

***Which human body parts are visible in the resource?***

The data collection contains video, speech, and analyses of the following types (Our list is based on the draft of [Quek et al. 2000: 4], version of 23-3-2000 which is available on the following Web site(s): URL: http://vislab.cs.wright.edu/KDI/; all citations from that paper make reference to this specific draft version.):

Speech:

F0 envelopes;

Prosodic annotation (not clear in detail which kind of annotation is selected);

Text transcripts of the spoken utterances (typically as ASCII text, see below);

Gesture:

three-dimensional traces of hand motion (separate for left and right hand);

three-dimensional traces of head motion;

Gaze: three-dimensional traces of gaze orientation.

This is the program of resource data types described in [Quek et al. 2000]; as the author did not have access to an example online, and as the cited article is not centrally about the state of advancement of the existing resources, we can at the current point not give any indications as to the availability of all these in larger amounts of data.

However, as videos have been annotated already with the tool described in [Quek et al. 2000], we expect such resources to be available (See below for more details: there is a 100 MB demo video available, which we did not have an opportunity to access; there are also two more videos of 39 and 156 seconds duration respectively).

***Which modalities are annotated?***

In their final, maximally annotated form, videos will contain annotations of the following types:

Speech:

F0 values (curves);

Speech amplitude (RMS);

Prosodic annotation (not clear in detail which kind of annotation is selected, no examples are provided);

Text transcripts of the spoken utterances;

Gesture:

three-dimensional traces of hand motion (left and right hand): velocities and positions;

three-dimensional traces of head motion: position;

Gaze: three-dimensional traces of head gaze direction: turn, nod and roll angles (cf. [Quek et al. 2000: 22]);

Psycholinguistic annotations introduced by analysts (e.g. concerning speech acts, hesitations, etc. These are interpretative and there may be different alternative such annotations for one phenomenon.

***Which other modalities are available/visible in the resource but have not been annotated?***

Facial expression and head movement.

## 6.23.4 Recorded computer behaviour

***Which interactive media are visible/audible in the resource and are used by the humans?***

None.

## 6.23.5 Recording

***What are the file types included in the resource? Are they organised in a database structure?***

As far as we can see from the documentation in [Quek et al. 2000: 8, 21], the resource comes as a set of files with time-aligned data.

Video: MJPEG, MPEG, QuickTime;

Textual transcripts: ASCII text, possibly structured by means of laboratory- or project-specific conventions;

Annotations and interpretations: ASCII text, possibly structured by means of laboratory- or project-specific conventions.

All data are synchronized, with the time stamps derived from the video being the reference time.

The format used is in line with psycholinguistic traditions. The protocol data can be stored in a database and thereafter queried selectively. We think that, provided an adequate standard for the annotation of such data were available (e.g. embodied in xml format), a conversion of the existing data into such a standard format would be possible. No reference is made, however, to any ongoing standardization. XML has not been documented to have been used in the project so far.

***How much data does the resource contain ?***

Sue-c: 39 seconds (1 camera)

Wombat-2: 156 seconds (left, right and close up camera)

In addition: 100 MB video (for password owners only)

***Who created the resource and when?***

Vislab, before spring 2000.

***How was the resource created?***

Video is recorded by means of three cameras: left and right, calibrated stereo, as well as one close-up on the head (cf. [Quek et al. 2000: 22]).

Speech is recorded along with the video signal and synchronized by means of time stamping.

Text transcripts are produced (as of the time when the paper [Quek et al. 2000] was written) purely manually (no mention of a transcription (support) tool in the paper); synchronization of text transcripts with the audio and video material is by time stamps on the onsets of relevant syllables or words (the granularity of the transcription may vary).

Gesture and gaze data (three-dimensional position data for hands and head, see above) are extracted from the video data automatically (Details can be found in further specialized papers by Quek and co-workers.).

The tool described in [Quek et al. 2000] includes a ``transcript generator'' (cf. op. cit., p. 23s.) which allows the user to produce (quasi-) textual protocol files containing the text transcript of the utterance, time stamps, gesture annotations, possibly interpretative psycholinguistic annotations, as well as references to the video frames.

***What is the application area?***

Description of living space.

***What was the original purpose of creating the resource?***

The tool described in [Quek et al. 2000] (and with it the resource) are designed to support psycholinguistic research on the interaction of speech, gesture and gaze in conversation.

On the Web site(s) of the KDI project (at VisLab, URL: http://vislab.cs.wright.edu/KDI/), the following motivation and description of the purpose of the work is given:

*investigation of cross-modal super-segmental cues for natural language processing: fusion of visual and spoken information that supports accurate speech understanding. Outcome aimed at: improved accuracy of various speech processing tasks by incorporating non-verbal inputs.*

*multimodal cues to detect speech disfluencies*

*the analysis of cross-modal deficits in Parkinson's Disease patients with a view toward a better understanding of the disease and possible diagnostic advancements.*

Technological goals:

*the processing of audio/video data to extract salient communicative entities;*

*parallelization of these algorithms;*

*the representation and maintenance of such entities in a multimedia database (particularly relevant for ISLE);*

*providing access to these entities in an intelligible fashion to a wide audience via network and physical media.*

### 6.23.6 Accessibility

***How does one get access to the resource?***

No information available.

***Is the resource available for free or how much does it cost?***

No information available.

***Has value been added to the original resource in terms of, e.g., transcriptions, annotations and/or tools, which are now available along with the original resource or otherwise available?***

Yes.

***Did the reviewer have access to the resource to write his/her contribution to 8.1?***

No. The review is based on the reference listed above and web information.

### 6.23.7 Usage

***Which purpose(s) can the resource be used for/has the resource been used for?***

The resource can be used for understanding the relation between speech and gesture.

***Who used the resource so far/who are the target users of the resource?***

So far the only known users are the authors of the reference listed above.

***Is the resource language dependent or language independent?***

Language dependent (American English)

### 6.23.8 Conclusion

***How interesting/important/high quality is the resource?***

Rich resource with detailed annotations including 3D features of gesture and gaze orientation.

***What do the authors regret (if anything) not to have done while building the resource?***

The authors state in a paper that more analysis could be done with 3D data coming from multiple cameras.

# 7 Lesser Known/Used Gesture Data Resources

## 7.1 ATR sign language gesture corpora

### 7.1.1 Description header

***Main actor***

IMS : Ulrich Heid (uli@IMS.Uni-Stuttgart.DE)
LIMSI: Jean-Claude MARTIN (martin@limsi.fr)

***Date of last modification of the description***

August 4th, 2001 (nakamura@slt.atr.co.jp has been contacted by email on august 8th)

### 7.1.2 References

Nakamura, S. et al: Multimodal Corpora for Human-Machine Interaction Research, Proc. of ICSLP, Volume IV, pp. 25-28. 2000

***Web site***

Contact nakamura@slt.atr.co.jp

### 7.1.3 Description

The Resource has been created by a consortium of ATR, Sharp and Tsukuba Electrotechnical Laboratories. An exact date is not indicated. The level of development described here corresponds to the status of the project at the point of publication of (Nakamura et al. 2000).

The resource was created under special video recording conditions (lighting, background colour). The purpose of the creation of the resource is the creation of an inventory of the most important words of Japanese sign language, as a basis for the development and the evaluation of gesture recognition systems.

- Available modalities:
  - Raw data for gesture:
    - digitised movie of Japanese sign language (JSL)
    - video of persons from the waist upwards.

- Annotated modalities:
  - Manually tagged for each word of Japanese sign language produced in the samples.

Set-ups

| No. | Cameras | Persons | Status of p. | Words | Repet. | Sentences/words |
|---|---|---|---|---|---|---|
| 1 | 1 cam | 2 pers. | know JSL | 300 | 1 | 301 s with 2 or 3 w |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2 | 2 cams | 4 pers. | experts in JSL | 300 | 2 | none |
| 3 | 2 cams | 4 pers. | experts in JSL | none | 2 | 64 s with 3-5 w |

Format of resource: very large resource, on DVD-RAMs.

Use of the database is limited to research and only possible upon receipt of an application.

Added value is provided by the fact that there are strictly regulated recording and digitising conditions, and that the participants have been asked to repeat the words and sentences.

The data are for Japanese Sign Language.

# 7.2 IRISA Georal Multimodal Corpus

## 7.2.1 Description header

***Main actor***

LIMSI-CNRS : Jean-Claude MARTIN (martin@limsi.fr)

***Date of last modification of the description***

27th of June 2001

## 7.2.2 References

***Web site***

The IRISA website: http://www.irisa.fr/cordial/ficheprojet-eng.htm

***References to additional information on the reviewed resource***

Guyomard, M., Le Meur D., Poignonnec and S., Siroux, J.: Experimental work for the dual usage of voice and touch screen for a cartographic application. Proceedings of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems, Vigsø, Denmark. 30 may - 2 June 1995, pp 153-156.

Siroux et al.: Modeling and processing of the oral and tactile activities. Proceedings of the International Conference on Cooperative Multimodal Communication(CMC/95), Eindhoven, II : 287-295. 1995.

Siroux, M. Guyomard, Y. Jolly, F. Multon, and C. Rémondeau: Speech and Tactile-Based Géoral System. In proceedings of 4th European Conference on Speech Communication and Technology, pages 1943--1946, Madrid, Espagne, 18-21 septembre 1995. http://www.irisa.fr/EXTERNE/projet/dialogue/ra95/node10.html - SECTION00022000000000000000

Siroux, J., Guyomard, M., Multon, F. and Remondeau, C.: Oral and gestural activities of the users in the GEORAL system. Proceedings of the First International Workshop on Intelligence and Multimodality in Multimedia Interfaces: Research and Application, 1995. Can be downloaded from: http://www.cogsci.ed.ac.uk/~john/IMMI/. Can be downloaded from:

## 7.2.3 Description

Guyomard, M., Le Meur D., Poignonnec, S., Siroux, J. (1995). Experimental work for the dual usage of voice and touch screen for a cartographic application. Proceedings of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems, Vigsø, Denmark. 30 may - 2 June 1995 pp 153-156

A speech-only interface to a cartographic map of Brittany was developed before but the recognition errors were found detrimental to the quality of the interaction. So a WOZ was made to study how people use speech and gestures on a tactile screen to interact with a graphical tourist map. The authors wanted to verify two hypothesis : H1) the presence of a touch screen modifies the linguistic behaviour of the user, and H2) when faced with difficulties with the oral communication, the user refers to the touch screen. The obtained results were not systematically quantified. They have changed the proper nouns to avoid a comprehension bias by subjects familiar with the region.

*Application*: tourist map

*Degree of simulation*: system fully simulated

*Input modalities*: speech, 2D gestures on tactile screen

*Output modalities*: 2D graphics, speech output

*Number of users*: 20

*Number of sessions by user*: not specified in the paper (one ?)

*User profile*: ranging from familiar with traditional man-machine interaction to the beginner in this domain

*Preliminary explanations* : the system is presented to the user providing an application closed to that simulated. The subject was then invited to familiarize with the touch screen using a drawing software.

*Scenario*: a scenario was given.

*Others*: several errors were produced by the accomplice

*Debriefing*: a free discussion followed by a semi-directed one eventually carried on

*Conclusions*:

Hypothesis H1 was verified (the presence of a touch screen modifies the linguistic behaviour of the user).

Hypothesis H2 was partly verified (when faced with difficulties with the oral communication, the user refers to the touch screen) : there is a threshold under which the subjects do not realize the existence of the tactile screen ; but once they have realized it, use of the touch screen is encompassed. The subjects who did not use much the tactile screen said during the de-briefing that 1) they were concentrated on the content of the questions not on the modalities, 2) the good capacities of the speech recognition system (simulated) did not exert the use of the tactile screen, 3) the touch on the tactile screen was not «normal» (during the IMMI workshop, J. Siroux said that the subjects might have been afraid to dirty the screen when touching it with their finger).

Globally the users thought that speech was sufficient and simpler. Yet most of them preferred combining speech and gestures.

observed combinations:

- are there any campsites here + pointing gesture
- give me the distance between Lansing and Morestel + two pointing gestures
- (small frequency) what are the campsites at + pointing gesture

observed gestures:

- pointing
- areas (either curved and completely joined up, or with angles and not joined)
- lines
- contours (follows a line on the map : road, coast)

gesture interpretation:

- most of the time the gesture was compatible with the speech (here => pointing, this area => drawing a circle)
- some exceptions: «this region» + pointing gesture (instead of drawing around a region) => speech modifying gesture interpretation
- some tactile events use element of the map (« right of this line » + gesture drawing a line) => the relevant zone was between the drawn line and the coastline displayed on the screen

Siroux, J., Guyomard, M., Multon, F. & Remondeau, C. (1995). Oral and gestural activities of the users in the GEORAL system. Proceedings of the First International Workshop on Intelligence and

Multimodality in Multimedia Interfaces: Research and Applications http://www.cogsci.ed.ac.uk/~john/IMMI/

Partly the same as the previous paper. The results of the analysis of the corpus have been used to improve the system which was then evaluated.

*Application*: tourist map

*Degree of simulation*: a WOZ user study (system simulated) followed by the system realization and a primary evaluation with some users

*Input modalities*: speech, 2D gestures on tactile screen

*Output modalities*: 2D graphics, speech output

*Number of users*: 20 (study) and 12 (system evaluation)

*Conclusions*:

- usage of tactile function not spontaneous (points, drawing a region)
- the system ignores gesture when not appearing during spoken utterance
- primary evaluation: the multimodal version increases the recognition of communication acts of 7% compared to the speech-only version of the system
- despite advice and low recognition of town names, the rate at which the tactile function is used remains low

# 7.3 LORIA Multimodal Dialogues Corpus

## 7.3.1 Description header

***Main actor***

LIMSI-CNRS : Jean-Claude MARTIN (martin@limsi.fr)

***Date of last modification of the description***

3rd of April 2001 (authors have been contacted to check the description but have not replied)

## 7.3.2 References

***Web site(s)***

(In French)
http://www.loria.fr/~romary/Documents/index.html

***Short description***

The communication part was simulated using a Wizard of Oz setting. A wizard is in one room and the subject in another room. They communicate via a microphone. Graphical objects are displayed on a screen. Subjects are instructed to build a surface using a geological tool (GOCAD). Speech and gesture has been annotated as well as the graphical screen. The resource has been recorded in 1993.

The document at http://www.loria.fr/~romary/Documents/index.html describes a coding scheme and some examples of annotations.

***How does one get access to the resource?***

By contacting romary@loria.fr.

# 7.4 University of California Video Series on Nonverbal Communication

## 7.4.1 Description header

***Main actor***

LIMSI-CNRS : Jean-Claude MARTIN (martin@limsi.fr)

***Date of last modification of the description***

27th of June 2001

## 7.4.2 References

***Web site***

A web site on exploring nonverbal communication: http://zzyx.ucsc.edu/~archer/

***Illustrative sample picture or video file***

**Figure 7.4.1.** Examples taken from the web site.

## 7.4.3 Description

NONVERBAL COMMUNICATION includes facial expressions, tones of voice, gestures, eye contact, spatial arrangements, patterns of touch, expressive movement, cultural differences, and other "nonverbal" acts.

Several videos can be ordered from the web site :

- A WORLD OF DIFFERENCES: Understanding Cross-Cultural Communication
- THE HUMAN BODY: Appearance, Shape and Self-Image
- THE HUMAN FACE: Emotions, Identities and Masks

- THE HUMAN VOICE: Exploring Vocal Paralanguage
- A WORLD OF GESTURES: Culture and Nonverbal Communication
- THE INTERPERSONAL PERCEPTION TASK (IPT)
- THE INTERPERSONAL PERCEPTION TASK-15 (IPT-15)

A WORLD OF DIFFERENCES (30 minutes) explores 14 different ways--verbal and nonverbal--that two people from different cultures can fail to understand each other. Some of these differences reflect language and translation problems. But many others involve subtle differences in etiquette, gestures, values, norms, rituals, expectations, and other important cross-cultural variations.

THE HUMAN FACE: EMOTIONS, IDENTITIES AND MASKS (30 minutes) explores the power of the face. Twelve different facial properties are examined, and the importance of this extraordinary human instrument is demonstrated for each.

THE HUMAN VOICE, explores the power and importance of this uniquely human instrument. When we speak, we use words, but we also "perform" these words using the range and subtlety of our voices. Spoken language therefore contains two distinct types of communication: (1) "text" (the words themselves) and (2) "vocal paralanguage", the thousands of ways in which any given words can be said. Text is whatever can be typed on a page. Vocal paralanguage is everything else--intonation, pitch, regional accent, sarcasm, hesitations, truthfulness, emotion, etc.

In A WORLD OF GESTURES, we see people from dozens of nations performing gestures that are powerful, poignant, subtle, and sometimes outrageous. Different types of gestures are shown, including those for beauty, sexual behaviour, suicide, aggression, and love. Since gestures often involve powerful emotions, many of these sequences are provocative, humorous, and entertaining. The meaning and function of gestures are also explored. For example, why is it that while some cultures have a huge number of obscene gestures, other cultures have not a single one? How old are some gestures? How are new gestures created in a society? A WORLD OF GESTURES also explores the origin of gestures and examines how "fluency" in gestures is acquired as children develop. Famous instances of "gesture controversy" are described--for example, when the Prime Minister of England unwittingly gave an obscene gesture to large crowds of enthusiastic admirers.

The INTERPERSONAL PERCEPTION TASK (IPT) is a videotape about nonverbal communication and social perception. Unlike most videotapes, the IPT gives viewers an opportunity for active participation. Viewers are asked to guess or "decode" something about each of the IPT scenes. Viewers see 30 brief scenes, each 30 to 60 seconds long. After each scene, there is an opportunity to answer a question. In one scene, the viewer sees a woman talking on the telephone. Immediately after this scene, the IPT video asks the viewer whether the woman is talking to (a) her mother, (b) a close female friend, or (c) her boyfriend. In another scene, the viewer sees two men who have just played basketball; the viewer is asked to decide which man won the basketball game. In a third scene, the viewer sees a woman giving two descriptions of her childhood; the viewer is asked to decide which description is a lie. The 30 IPT scenes depict five common types of social judgments-- intimacy, competition, deception, kinship, and status. After each scene, the viewer has a chance to "decode" something important about what he or she has just seen. The viewer can try to determine the correct answer by "reading" nonverbal behaviour--perhaps a facial expression, tone of voice, gesture, touch, glance, or hesitation. The 30 IPT scenes contain a full range of spontaneous nonverbal behaviours in context. For each scene, there is an OBJECTIVELY correct answer.

# 7.5 University of Venice Multimodal Transcription of a Television Advertisement

## 7.5.1 Description header

***Main actor***

LIMSI-CNRS : Jean-Claude MARTIN (martin@limsi.fr)

***Date of last modification of the description***

26 June 2001

## 7.5.2 References

***Web site***

A paper on: Towards multimodal corpora: http://www.eng.helsinki.fi/doe/ESSE5-2000/Baldry-Thibault.htm

***Illustrative sample picture or video file***

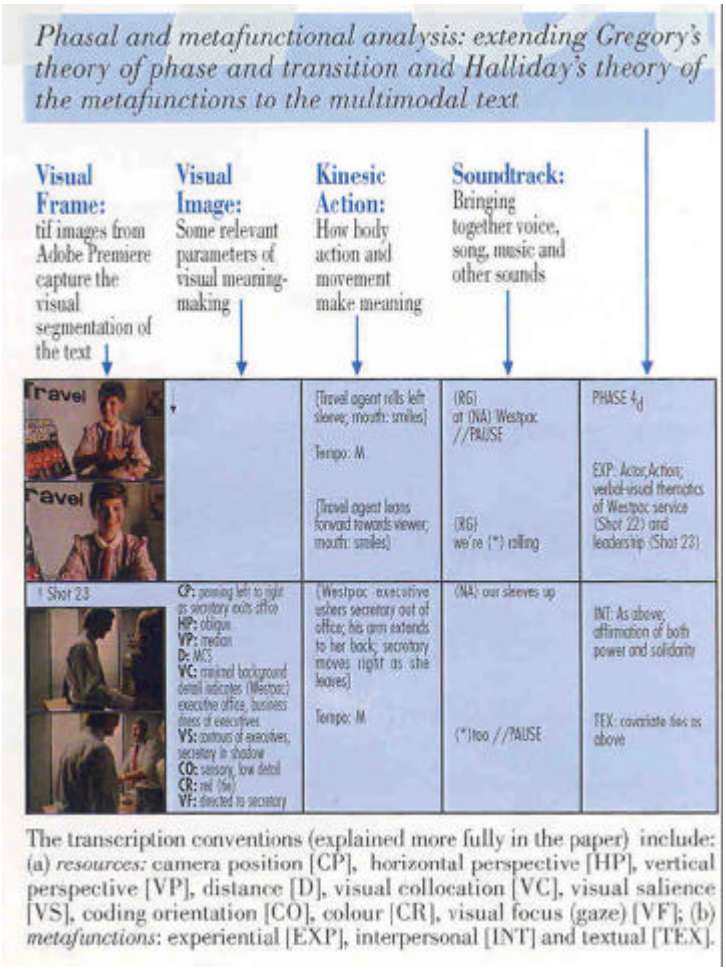


**Figure 7.5.1.** A sample multimodal transcription: The sample multimodal transcription shown here, is taken from the flyer for Multimodality and Multimediality in the Distance Learning Age and is derived from Paul Thibault's paper Multimodal Transcription of a Television Advertisement: Theory and Practice. It shows that unlike the majority of transcription procedures which are overwhelmingly biased to language, the multimodal transcription treats all modalities, and not just language, as systematically semiotic. Some of the transcription conventions used to describe a text's visual images, kinesic actions and soundtrack are also shown. As well as description of the individual resources, the transcription includes a phasal and metafunctional analysis based on Halliday's notion of metafunctions and Gregory's notion of phase and transition, both of which were originally confined to the linguistic semiotic. The sample transcription illustrates the application across semiotic modalities of (i) the experiential metafunction [EXP], which is concerned with interpreting phenomena in terms of configurations of a process, the participant(s) in the process and the circumstances that accompany the process, (ii) the interpersonal metafunction [INT] concerned with the social relations between the participants in the interaction and the ways in which they modally and evaluatively orient both to each other and to the experiential meanings, (iii) the textual metafunction [TEX] concerned with those resources which enable the text user to keep track of the text as a coherent unit of meaning and to relate the text to its context of situation. The transcription also shows how phase and transition analysis is essential in indicating a text's peak-and-trough co-patterning of resources. This can help the analyst to account for the ways in which variations in the kind or degree of selections in any given semiotic mo dality may impact upon the overall wave-like patterning of resources in a text.

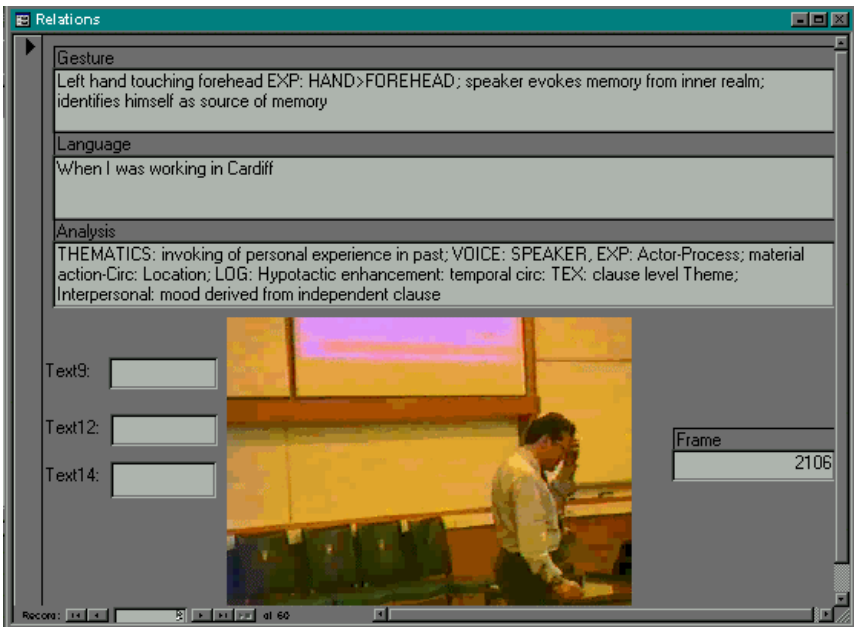


**Figure 7.5.2.** A screendump of a multimodal ressource annotated using the Multimodal Corpora Authoring System.

## 7.5.3 Description

The main interest of the authors lies in understanding the properties and functions of dynamic genres - classroom encounters, lectures, documentaries, TV ads and training films - whose meaning is dependent on a highly complex integration of a rather longer list of resources: verbal and written discourse, gesture, gaze, colour, voice quality.

The corpus they have in mind is designed to allow a researcher to obtain information about the use of a particular instance of a meaning-making resource in the context of other resources.

The work we have described above is part of the ACADI Project of the University of Pavia, co-ordinated by Maria Pavesi, which is concerned with the contrastive analysis of scientific texts within the context of on-line corpora, distance learning and hypermedia, which in turn is part of the CITATAL project. The latter project, which is partly financed by MURST, the Italian Ministry for Higher Education and Research, involves the Universities of Padua, Pavia, Pisa and Trieste and is co-ordinated by Carol Taylor Torsello (Univ. of Padua). See the "Moonlets" site (http://moon.unipv.it) for further information on multimodality & multimediality and related educational applications.

Address for correspondence: Baldry@moon.unipv.it

Thibault, Paul (In Press A) Multimodal Transcription of a Television Advertisement: Theory and Practice in Anthony Baldry (ed.) Multimodality and multimediality in the distance learning age, Campobasso: Editrice Lampo

# 8 Market and User Needs

## 8.1 Introduction

The study reported in this chapter was carried out by ELDA. It aims at determining market needs for NIMM (Natural interaction and multimodal) data resources. The main objective of the market study is to collect strategic information on current and future needs for NIMM data resources.

The survey provides a targeted approach to gathering information on market needs. It was directed to over 150 contacts identified as specialists in the domain of human language technologies, some of them being identified more particularly as specialists of NIMM resources.

The survey was sent out in two phases: 14 responses were compiled in October 2001 and 11 other responses were compiled at the end of December 2001, for a total of 25 respondents (ca. 16% feedback according to the original set of contacts). 60% of the respondents are academic and 40% are commercial.

The survey confirmed the point that there is a significant interest in NIMM data resources, although these resources seem to be currently under-estimated in terms of budget or funding. It also raises the idea that within 2-5 years time, these resources would increase in terms of availability and common usage. The survey results will be used in the ISLE project as a basis for the preparation of guidelines for the creation of NIMM data resources.

This chapter presents the methodology followed to carry out the survey and aims at gathering useful facts on present and future needs for NIMM resources among major players in the market. Questionnaire and corresponding figures are given in Appendices 2 and 3.

## 8.2 Methodology

### 8.2.1 Questionnaire

The study was targeted towards major players in the NIMM data resources market.

The primary objectives of the study were as follows:

- First, to define/identify the current and future market structure (profile of key players, key applications, trends).
- Second, to obtain a clear picture of users' needs and expectations, in order to be able to plan future activities and developments.

The key areas of concern were the following:

1. Defining the key players in the area: user or provider, their point of view with regards to a potential market.

2. Defining multimodal LRs: type of LRs produced or needed, aim of LRs, application areas, languages of interest.

3. Identifying the production and validation procedures for multimodal LRs, as well as best practices and standards.

4. Identifying the distribution objectives.

### 8.2.2 Survey

The questionnaire was distributed through 4 different contact lists:

1. The ISLE short list which consisted of 38 people.
2. The ISLE partners (8 contacts).
3. ELRA members and associates (112 contacts).
4. Contacts proposed during the survey (through a specific question in the questionnaire: *Would you recommend the questionnaire to be mailed to someone from another organisation?*): 2 new contacts were proposed.

The questionnaire was first emailed to the ISLE partners for comments. It was then sent out to the first three contact lists mentioned above on 30 July 2001. The questionnaire was also sent to the contacts proposed during the survey (1 out of 2 was already contacted through one of the other lists). The first set of replies was gathered in October 2001. At that time, 14 responses were compiled. In order to increase the number of responses, we decided to send the questionnaire a second time by directing it only to the non respondents. At the end of December 2001, 11 new responses were compiled, for a total of 25 respondents (out of more than 150 contacts, i.e. ca. 16% of responses).

# 8.3 Results

An analysis of all 25 responses to the Market Needs for NIMM data resources survey is provided in Appendix 3.

Figure 8.3.1 shows the distribution over academic and commercial respondents according to their country of origin.

| **Country** | **Academic** | **Commercial** | **Total** |
|---|---|---|---|
| Belgium | 1 | 0 | 1 |
| France | 6 | 1 | 7 |
| Germany | 2 | 5 | 7 |
| Greece | 1 | 0 | 1 |
| Italy | 1 | 0 | 1 |
| Spain | 1 | 1 | 2 |
| The Netherlands | 1 | 0 | 1 |
| United Kingdom | 1 | 0 | 1 |
| USA | 1 | 3 | 4 |
| **Total** | 15 | 10 | 25 |

**Figure 8.3.1.** Distribution of academic and commercial respondents across countries.

The key figures from the study are explained below.

## 8.3.1 Multimodal Language Resources (LR) Market

This first question aims to define the players with regard to their market group (user and/or provider) and to estimate the exchange possibilities of LRs produced within this market.

Although the respondents seem to be divided equally into the user and the provider group, a greater number of them (76%) said to be using LRs which are produced internally. 56% use LRs produced by specific contracted vendors.

## 8.3.2 Data

The type of resources that are currently acquired or offered by the respondents are presented in the histogram in Figure 8.3.2.

Other types of data linked to NIMM LRs were highlighted by respondents, including: annotations (speech, gesture coding, transcriptions), interface surfaces, infrared video of gestures, coordinates of pointing gestures on the workspace, magnetometer and 3-d optical, written language.

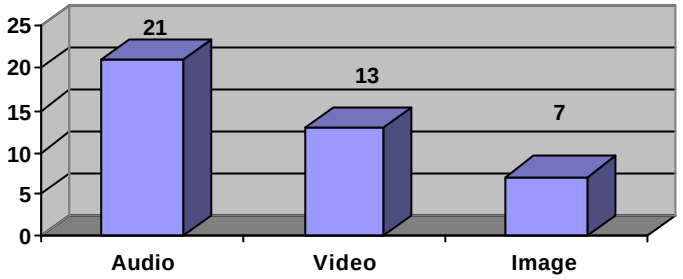


**Figure 8.3.2.** Type of resources currently acquired or offered by respondents.

## 8.3.3 Applications

In order to check the needs for specific NIMM LRs, a list of possible applications for LRs was proposed to the respondents. This enabled to highlight the main trends in terms of applications.

The answers for each application are given below:

***Authentication***

In descending order: Speech verification (8), Face verification (6), User authentication (5). Other authentication-oriented applications were proposed, including: finger print and signature, biometric authentication (speech, signature).

***Recognition***

Speech recognition (14), Face recognition (7), Person recognition (3), Expression recognition (3). Other applications proposed include: mimic, music and other sounds, gesture recognition, gestures on a touch screen.

***Analysis***

Speech/lips correlation (7), Body movements tracking (lips, hands, head, arms, legs, etc.) (6). Other applications proposed include: Cooperation between gesture and speech, acoustic, video, 3-d optical, midsagital magnetometry, written language analysis.

***Synthesis***

Multimedia development (6), Talking heads (5), Humanoid agents (5), Avatars (2). Other applications proposed include: text generation.

***Control***

Voice control (7), Speech assisted video (1).

***Other***

Information retrieval (14).

Other applications proposed include: multimodal command languages (speech + gestures), research into cross-modality issues, multimodal dialogue (speech + gesture), linguistic research, information extraction, text summarisation.

To sum up, the six main applications of current interest are: Information retrieval, Speech recognition, Speech verification, Face recognition, Speech/lips correlation, Voice control.

### 8.3.4 Application Areas

Question 4 is a complement to the previous question regarding the trends in terms of applications. This question enables to highlight more specifically the market areas. The application area mentioned the most by the respondents is Research (21). Then follow Information Systems (e.g. banking, tourism, telecommunication) (14), Web Applications (10), Education/Training (9), Entertainment (6). Other areas were proposed by the respondents, including: security, control of consumer devices, media archiving for content providers.

### 8.3.5 Languages

To have a good idea of the NIMM LRs needed, it is necessary to know the needs in terms of languages. 72% of the respondents show a trend towards obtaining/producing language-dependent resources, although a good number of them show interest in language-independent resources as well (52%). The main languages of interest are: English UK and US(14), French (10), German (7), European and Latin American Spanish (7) and Italian (5).

### 8.3.6 For Users Only

The previous questions aimed to check the present trends. In order to have an idea of future trends on a short-term basis, we added an open question about resources or applications which users believe to need within the next 2-5 years.

Unfortunately, only 4 of the 25 respondents answered this question, which is too few to extract any significant conclusions. The 4 answers can be found in Appendix 3.

### 8.3.7 For Providers Only

This section is dedicated to the production and validation of NIMM LRs.

The majority (60%) of the producers of NIMM LRs choose to follow internal specifications, whereas 20% choose to follow external specifications. Four external specifications were mentioned: ISLE guidelines, Eagles, Speecon and Orientel. Of course, none of these are "really" well-established specifications for NIMM LRs.

Nearly half of the respondents said to follow specific standards. Mentioned standards are: BAS, Eagles, TEI, Speecon, Orientel, XML/Unicode, ISLE, MPEG-7, ESPS.

In terms of validation policy, most of the respondents answered to carry out validation internally. Only one respondent said to carry out external validation. Only 24% of the respondents said to produce validation reports, whereas 40 % confirmed that they did not produce validation reports.

### 8.3.8 Evaluation Frequency of LR Needs

The frequency of re-evaluation within the company was also included in the questionnaire through the question *"How often do you re-evaluate your LR needs and seek available databases?"*. The following answers were obtained (in descending order): Once per quarter (5), Not applicable (5), Once per year (4), Once per semester (3), Monthly (2), Once every 1-2 years (1).

### 8.3.9 Distribution

A specific question concerning the distribution of NIMM LRs was added in order to evaluate potential active providers in a potential market. 12 respondents in 20 said to be willing to make their resources available to others. The main reason given by the 8 other respondents for not distributing to others is a legal one, whereas strategic and commercial reasons come in second and third place.

Among the first 12 respondents, 11 would agree to licence their resources to researchers, while 9 would license to tool developers and 8 to end-users.

### 8.3.10 Market

Several questions relating to the market were asked to respondents to get their point of view on the current and future market of NIMM LRs.

It seems that 40% of them spend a fair amount of money for various data acquisition (between 6,000 to 300,000 EURO - only 1 respondent answered 0) and production (between 6,000 and 1,500,000 EURO). The expected budget within 3-5 years time seem to show an increasing need for NIMM LRs (between 5 and 10 times more than the present budget).

Though only 6 respondents answered the question about the actors who are leading the development in this market sector, 4 out of 6 agree that there are several leading actors (instead of key actors).

Regarding the other questions concerning the market (*Which is your estimate of the Multimodal LR market today?, How do you see the future in your area?*), the number of answers is too low to extract significant figures.

## 8.4 Conclusion

The survey confirmed the point that there is a significant interest in NIMM data resources, although these resources seem to be currently under-estimated in terms of budget or funding. It also highlights the idea that within 2-5 years time, these resources would increase in terms of availability and common usage. We regret that the output of the survey with regard to the specific open questions on future needs and market seems too low to draw reasonable figures. A second survey oriented on these issues is needed to improve the results.

The main results are listed below:

- the type of resources acquired or offered the most are audio resources, then come video resources and static image resources.
- the main 6 applications of current interest are: Information retrieval, Speech recognition, Speech verification, Face recognition, Speech/lips correlation, Voice control.
- the application area mentioned the most by the respondents is Research, before Information Systems (e.g. banking, tourism, telecommunication), Web Applications, Education/Training, and Entertainment.
- the main languages of interest are: English UK and US(14), French (10), German (7), European and Latin American Spanish (7) and Italian (5).
- the majority of the producers of NIMM LRs follow internal specifications at 60%.

- most of the respondents carry out validation internally but only 24% of them produce validation reports.
- 40% of the respondents spend a fair amount of money for various data acquisitions (between 6,000 to 300,000 EURO) and production (between 6,000 and 1,500,000 EURO). The expected budget within 3-5 years for NIMM LRs seem to be between 5 and 10 times higher than the present budget.

The survey work is a means to develop an improved strategy with regard to the planning of future market and distribution policies, together with set-up of promotion activities. Therefore, the results will be used as a basis for the preparation of guidelines for the creation of NIMM data resources within the ISLE project, which will be disseminated to the interested community. These guidelines will in particular take into consideration the validation issues which do not seem to be widely addressed at present, although they could be a value-added support to the extension of the NIMM data resources market. The guidelines will also encourage the development of standards in this market area, which in itself tends to reinforce the activity and its market share.

# Acknowledgements

We gratefully acknowledge the support of the ISLE project by the European Commission's HLT Programme. We would also like to thank data resource creators for their willingness to make descriptions of their data resource publicly available in this report as well as for the time they have given in communicating with the European ISLE NIMM team.

# 9 Appendix 1. Questionnaires collected at Dagstuhl November 2001

These questionnaires were informally collected during the Dagstuhl Seminar on “Coordination and Fusion in Multimodal Interaction": http://www.dfki.de/~wahlster/Dagstuhl_Multi_Modality/

The questionnaires were manually collected by the Working Group #2 on Data Collection and Multimodal Annotation Tools (including Lisa Harper, Michael Kipp, Emiel Krahmer, Jean-Claude Martin, Dagmar Schmauks...) on 1 Nov 2001. The handwritten questionnaires were typed by Jean-Claude Martin.

For more information and/or clarification, individual researchers should be contacted directly (see contact information in each questionnaire).

More information can also be found on http://www.limsi.fr/Individu/martin/questionnairesDagstuhl/

## 9.1 Utrecht University - Denk project

### 9.1.1 Your name, e-mail

Robbert-Jan Beun

e-mail: rj@cs.uu.nl

url: http://www.cs.uu.nl/staff/rj.html

### 9.1.2 Name of institute or organization

Utrecht University

Dept. of Computer Science

Padualaan 14

Postbus 80.089

NL-3508 TB Utrecht

NL

fax: +31-30-251-3791

### 9.1.3 Names of projects

Denk

### 9.1.4 How many corpora have you assembled?

5 or 6 (2 multimodal)

### 9.1.5 How many subjects were captured per corpus (average number)?

Speech describing picture monologues : around 5

Speech and multimodal : around 40

### 9.1.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single
(X) dyads
( ) triads
( ) groups larger than 3
( ) any Wizard-of-Oz data?

### 9.1.7 What are the profiles of the subjects? (age groups, profession, gender,...)

Very different in age and profession

### 9.1.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*
*computer: screen, keyboard, touchpad, dataglove ...*
human : arms, hand
Shared workspace of lego blocks

### 9.1.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*
*computer: speech, embodied agent, graphics ...*
human speech

### 9.1.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*
AVI

### 9.1.11 What coding scheme(s) do you use?

own

### 9.1.12 What coding tool(s) do you use primarily?

manuscripts

### 9.1.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*
building blocks

### 9.1.14 Which spoken/sign languages occur in your corpora?

Dutch

### 9.1.15 What are the costs for using your corpora?

(X) free for research use
( ) available for prices like _____ (example)
( ) not available
( ) not yet available, but maybe in the future

### 9.1.16 What purposes have your corpora been used for?

(X) exploratory studies
(X) evaluating hypotheses
(X) computer system evaluation => Denk system : use of expressions, focus space, indirect commands
( ) training data (machine learning, classifiers, recognition)
( ) collecting a body of representative specimen for an inventory/lexicon

### 9.1.17 Are you planning to collect corpora in the future?

no

### 9.1.18 Are you looking for available corpora? If so, what are your needs?

Not yet

# 9.2 LORIA

## 9.2.1 Your name, e-mail

Noelle Carbonell
e-mail: Noelle.Carbonell@loria.fr

## 9.2.2 Name of institute or organization

LORIA
B 218
BP 239
F-54506 Vandoeuvre-lès-Nancy
F
phone: +33 3 83 59 20 32
fax: +33-3-83 41 30 79

## 9.2.3 Names of projects

IS4 All thematic Network

## 9.2.4 How many corpora have you assembled?

5

## 9.2.5 How many subjects were captured per corpus (average number)?

6-8

## 9.2.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single
( ) dyads
( ) triads
( ) groups larger than 3
( ) any Wizard-of-Oz data?

## 9.2.7 What are the profiles of the subjects? (age groups, profession, gender,...)

General public age : >20 <40 professional activity

### 9.2.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*

*computer: screen, keyboard, touchpad, dataglove ...*

human : hand

computer : screen

### 9.2.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*

*computer: speech, embodied agent, graphics ...*

human : speech, gesture

software successive states

### 9.2.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*

primary data on tapes (video, DAT)

no files

30 hours per experiment

attempt at quicktime failed but it was in 1998 (could be possible now)

### 9.2.11 What coding scheme(s) do you use?

Written transcriptions of subjects' and system multimodal utterances

Home-made coding

### 9.2.12 What coding tool(s) do you use primarily?

Some ad hoc tools (unix perl shell)

### 9.2.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*

Graphical design application

Process control

Information centre (flights)

### 9.2.14 Which spoken/sign languages occur in your corpora?

French only

### 9.2.15 What are the costs for using your corpora?

(X) free for research use

( ) available for prices like _____ (example)

( ) not available
( ) not yet available, but maybe in the future

### 9.2.16 What purposes have your corpora been used for?

(X) exploratory studies
(X) evaluating hypotheses
( ) computer system evaluation
( ) training data (machine learning, classifiers, recognition)
( ) collecting a body of representative specimen for an inventory/lexicon

### 9.2.17 Are you planning to collect corpora in the future?

yes

### 9.2.18 Are you looking for available corpora? If so, what are your needs?

# 9.3 MIT Media Lab

## 9.3.1 Your name, e-mail

Justine Cassel
e-mail: justine@media.mit.edu
url: http://justine.www.media.mit.edu/people/justine/

## 9.3.2 Name of institute or organization

MIT
Media Lab.
E15-315
20 Ames Street
MA 02139 Cambridge
USA
fax: +1-617-253-4899

## 9.3.3 Names of projects

## 9.3.4 How many corpora have you assembled?

>10

## 9.3.5 How many subjects were captured per corpus (average number)?

25

## 9.3.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single
(X) dyads
(X) triads
( ) groups larger than 3
(X) any Wizard-of-Oz data?

### 9.3.7 What are the profiles of the subjects? (age groups, profession, gender,...)

Male & female

4-11 years old (> 25)

18-22 years old

### 9.3.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*

*computer: screen, keyboard, touchpad, dataglove ...*

human: face, arms, hand, body.

computer: screen

Toy castle

### 9.3.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*

*computer: speech, embodied agent, graphics ...*

human: speech, hand/arm gesture, posture, facial expression.

computer: speech, embodied agent

### 9.3.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*

QuickTime, MPEG

### 9.3.11 What coding scheme(s) do you use?

McNeill

custom

Schiffrin

### 9.3.12 What coding tool(s) do you use primarily?

MacShapa

### 9.3.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*

Storytelling

Learning

### 9.3.14 Which spoken/sign languages occur in your corpora?

English

### 9.3.15 What are the costs for using your corpora?

(X) free for research use

( ) available for prices like _____ (example)

( ) not available

( ) not yet available, but maybe in the future

### 9.3.16 What purposes have your corpora been used for?

(X) exploratory studies

(X) evaluating hypotheses

(X) computer system evaluation

( ) training data (machine learning, classifiers, recognition)

( ) collecting a body of representative specimen for an inventory/lexicon

### 9.3.17 Are you planning to collect corpora in the future?

yes

### 9.3.18 Are you looking for available corpora? If so, what are your needs?

no

# 9.4 Universität Bielefeld - Situated Verbmobil Artificial Communicators

## 9.4.1 Your name, e-mail

Gernot Fink
e-mail: gernot@techfak.uni-bielefeld.de
url: http://www.techfak.uni-bielefeld.de/~gernot/

## 9.4.2 Name of institute or organization

Universität Bielefeld
Technische Fakultät
Angewandte Informatik
Universitätsstr. 25
10 01 31
D-33501 Bielefeld
D
phone: +49-521-106-2931
fax: +49-521-106-2992

## 9.4.3 Names of projects

SFB 360 “Situated Verbmobil Artificial Communicators”

## 9.4.4 How many corpora have you assembled?

2 speech only car
1 WOZ actions annotated
video and audio 20 dialogues Banfix
video of handwriting captured by tablet

## 9.4.5 How many subjects were captured per corpus (average number)?

2 speech only car => 22 subjects
1 WOZ actions annotated => 50
video and audio 20 dialogues Banfix => 20-40
video of handwriting captured by tablet => 5-10

## 9.4.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single
(X) dyads
( ) triads
( ) groups larger than 3
(X) any Wizard-of-Oz data?

### 9.4.7 What are the profiles of the subjects? (age groups, profession, gender,...)

University environments

### 9.4.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*
*computer: screen, keyboard, touchpad, dataglove ...*
human : upper part of body
workspace

hands, upper part of body
tablet

### 9.4.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*
*computer: speech, embodied agent, graphics ...*
human : speech, overlap, spontaneous speech effects, dialogues acts, gesture (Banfix)

### 9.4.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*
Analog

### 9.4.11 What coding scheme(s) do you use?

Custom mark-up
Orthographic for speech
prosody

### 9.4.12 What coding tool(s) do you use primarily?

Prosodic annotation tool based on Xwaves

### 9.4.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*
Construction, auto-non vital car functions

### 9.4.14 Which spoken/sign languages occur in your corpora?

German

### 9.4.15 What are the costs for using your corpora?

(X) free for research use => ask for Baufix (SFB) Saarbrucken-laughter in speech
( ) available for prices like _____ (example)
(X) not available => Car speech
( ) not yet available, but maybe in the future

### 9.4.16 What purposes have your corpora been used for?

(X) exploratory studies => Baufix
(X) evaluating hypotheses => Baufix
(X) computer system evaluation => Speech data
(X) training data (machine learning, classifiers, recognition) => handwriting
( ) collecting a body of representative specimen for an inventory/lexicon

### 9.4.17 Are you planning to collect corpora in the future?

Not currently

### 9.4.18 Are you looking for available corpora? If so, what are your needs?

Handwriting

# 9.5 MITRE Corporation - Multimodal referent resolution in map-based interaction

## 9.5.1 Your name, e-mail

Lisa Harper
e-mail: lisah@mitre.org

## 9.5.2 Name of institute or organization

The MITRE Corporation
AI Dept. / W15B
1820 Dolley Madison Bldv.
VA 22102-3481 McLean
USA
phone: +1-703-883-5241
fax: +1-703-883-1379

## 9.5.3 Names of projects

Multimodal referent resolution in map-based interaction

## 9.5.4 How many corpora have you assembled?

1

## 9.5.5 How many subjects were captured per corpus (average number)?

2

## 9.5.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

( ) single
(X) dyads
( ) triads
( ) groups larger than 3
( ) any Wizard-of-Oz data?

### 9.5.7 What are the profiles of the subjects? (age groups, profession, gender,...)

½ non specialists, ½ military experience

### 9.5.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*
*computer: screen, keyboard, touchpad, dataglove ...*
computer : large electronic whiteboard
human : body, arms, hands, no face

### 9.5.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*
*computer: speech, embodied agent, graphics ...*

sketch, hand/arm, head, speech (transcription)

### 9.5.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*
Digital quicktime sorenson3

### 9.5.11 What coding scheme(s) do you use?

Not yet decided

### 9.5.12 What coding tool(s) do you use primarily?

Anvil

### 9.5.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*
map task

### 9.5.14 Which spoken/sign languages occur in your corpora?

English

### 9.5.15 What are the costs for using your corpora?

(X) free for research use => Not yet finished
( ) available for prices like _____ (example)
( ) not available

( ) not yet available, but maybe in the future

### 9.5.16 What purposes have your corpora been used for?

(X) exploratory studies
( ) evaluating hypotheses
( ) computer system evaluation
( ) training data (machine learning, classifiers, recognition)
( ) collecting a body of representative specimen for an inventory/lexicon

### 9.5.17 Are you planning to collect corpora in the future?

yes

### 9.5.18 Are you looking for available corpora? If so, what are your needs?

Speech, hand gesture, gaze in collaborative task-based interaction

## 9.6 AIST Tokyo - JEITA Multimodal corpora

### 9.6.1 Your name, e-mail

Koiti Hasida
e-mail: hasida@aist.go.jp
url: http://i-content.org/hasida/

### 9.6.2 Name of institute or organization

AIST Tokyo Waterfront
Cyber Assist Research Center
2-41-6 Aomi
Koto-Ku
135-0064 Tokyo
fax: +81-3-5530-2061

### 9.6.3 Names of projects

JEITA Multimodal corpora

### 9.6.4 How many corpora have you assembled?

2

### 9.6.5 How many subjects were captured per corpus (average number)?

many

### 9.6.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

( ) single
(X) dyads
( ) triads
(X) groups larger than 3
( ) any Wizard-of-Oz data?

### 9.6.7 What are the profiles of the subjects? (age groups, profession, gender,...)

20-30 years

students
male & female

### 9.6.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*
*computer: screen, keyboard, touchpad, dataglove ...*
human : whole body
no computer

### 9.6.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*
*computer: speech, embodied agent, graphics ...*
human : speech

### 9.6.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*
MPEG 1 & 2

### 9.6.11 What coding scheme(s) do you use?

GDA

### 9.6.12 What coding tool(s) do you use primarily?

GDA Tagging Editor

### 9.6.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*
tourism, face-guessing task

### 9.6.14 Which spoken/sign languages occur in your corpora?

Japanese

### 9.6.15 What are the costs for using your corpora?

( ) free for research use
(X) available for prices like __$40-50_____ (example)
( ) not available
(X) not yet available, but maybe in the future

### 9.6.16 What purposes have your corpora been used for?

(X) exploratory studies
( ) evaluating hypotheses
( ) computer system evaluation
( ) training data (machine learning, classifiers, recognition)
( ) collecting a body of representative specimen for an inventory/lexicon

### 9.6.17 Are you planning to collect corpora in the future?

yes

### 9.6.18 Are you looking for available corpora? If so, what are your needs?

End-user presentation

# 9.7 Microsoft Research

## 9.7.1 Your name, e-mail

Derek Jacoby
e-mail: derekja@microsoft.com

## 9.7.2 Name of institute or organization

Microsoft Research
Program Manager Speech Technology Group
One Microsoft Way
WA 98052-6399 Redmond
USA

## 9.7.3 Names of projects

## 9.7.4 How many corpora have you assembled?

1 corpus over 3 iterations

## 9.7.5 How many subjects were captured per corpus (average number)?

200  30 min per corpus

## 9.7.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single
( ) dyads
( ) triads
( ) groups larger than 3
(X) any Wizard-of-Oz data?

## 9.7.7 What are the profiles of the subjects? (age groups, profession, gender,...)

5 ages groups
profession not controlled (education info)
balanced Male / Female

### 9.7.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*

*computer: screen, keyboard, touchpad, dataglove ...*

Human face and hand

Computer touchpad

### 9.7.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*

*computer: speech, embodied agent, graphics ...*

Human speech, gesture (selected field)

### 9.7.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*

AVI & MPEG

### 9.7.11 What coding scheme(s) do you use?

Home developed

Application specific (speech + field)

Coding scheme for speech

### 9.7.12 What coding tool(s) do you use primarily?

Home developed

### 9.7.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*

Email + calendar

### 9.7.14 Which spoken/sign languages occur in your corpora?

English

### 9.7.15 What are the costs for using your corpora?

( ) free for research use

( ) available for prices like _____ (example)

(X) not available

( ) not yet available, but maybe in the future

### 9.7.16 What purposes have your corpora been used for?

( ) exploratory studies
( ) evaluating hypotheses
(X) computer system evaluation
(X) training data (machine learning, classifiers, recognition)
( ) collecting a body of representative specimen for an inventory/lexicon

### 9.7.17 Are you planning to collect corpora in the future?

Yes

### 9.7.18 Are you looking for available corpora? If so, what are your needs?

Not multimodal yet

## 9.8 University of Art and Design Media Lab

### 9.8.1 Your name, e-mail

Kristiina Jokinen
e-mail: kristiina.jokinen@uiah.fi
url: http://www.uiah.fi/ www.fng.fi/hugo.htm

### 9.8.2 Name of institute or organization

University of Art and Design
Media Lab.
Hämeentie 135C
FIN-00560 Helsinki

### 9.8.3 Names of projects

Interact
Natural Interaction in spoken interfaces

### 9.8.4 How many corpora have you assembled?

1

### 9.8.5 How many subjects were captured per corpus (average number)?

100 (one user / one dialogue)

### 9.8.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

( ) single
(X) dyads
( ) triads
( ) groups larger than 3
( ) any Wizard-of-Oz data?

### 9.8.7 What are the profiles of the subjects? (age groups, profession, gender,...)

Professional service agent (female) & information seekers (male/female) age 20-70

### 9.8.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*

*computer: screen, keyboard, touchpad, dataglove ...*

### 9.8.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*

*computer: speech, embodied agent, graphics ...*

Human speech.

### 9.8.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*

### 9.8.11 What coding scheme(s) do you use?

Home grown for dialogue info (dialogue acts, topic, focus)

### 9.8.12 What coding tool(s) do you use primarily?

Home grown

### 9.8.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*

information seeking

### 9.8.14 Which spoken/sign languages occur in your corpora?

Finnish

### 9.8.15 What are the costs for using your corpora?

( ) free for research use

( ) available for prices like _____ (example)

( ) not available

(X) not yet available, but maybe in the future

### 9.8.16 What purposes have your corpora been used for?

(X) exploratory studies

( ) evaluating hypotheses

( ) computer system evaluation

(X) training data (machine learning, classifiers, recognition)

(X) collecting a body of representative specimen for an inventory/lexicon

### 9.8.17 Are you planning to collect corpora in the future?

Yes, for evaluation of the system.

### 9.8.18 Are you looking for available corpora? If so, what are your needs?

For our museum project, multimodal narration would be useful with sign language as one mode.

# 9.9 Linköping University - Swedish Dialogue System

## 9.9.1 Your name, e-mail

Arne Jönsson
e-mail: arnjo@ida.liu.se
url: http://www.ida.liu.se/~arnjo/

## 9.9.2 Name of institute or organization

Linköping University
Computer and Information Science
SE-58183 Linköping
fax: +46 13 14 22 31

## 9.9.3 Names of projects

SDS (Swedish Dialogue System)

## 9.9.4 How many corpora have you assembled?

1 (collected by Göteborg University)

## 9.9.5 How many subjects were captured per corpus (average number)?

Around 30 I think

## 9.9.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

( ) single
(X) dyads
( ) triads
( ) groups larger than 3
( ) any Wizard-of-Oz data?

## 9.9.7 What are the profiles of the subjects? (age groups, profession, gender,...)

Professional travel agent salesmen
Normal customers of various ages

### 9.9.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*

*computer: screen, keyboard, touchpad, dataglove ...*

Human Body

### 9.9.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*

*computer: speech, embodied agent, graphics ...*

Speech & gesture

### 9.9.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*

### 9.9.11 What coding scheme(s) do you use?

Different

### 9.9.12 What coding tool(s) do you use primarily?

Tractor and other tools developed in the project

### 9.9.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*

tourism, travel

### 9.9.14 Which spoken/sign languages occur in your corpora?

Swedish

### 9.9.15 What are the costs for using your corpora?

( ) free for research use

( ) available for prices like _____ (example)

( ) not available

(X) not yet available, but maybe in the future

### 9.9.16 What purposes have your corpora been used for?

(X) exploratory studies

( ) evaluating hypotheses

(X) computer system evaluation + DEVELOPPEMENT

( ) training data (machine learning, classifiers, recognition)
( ) collecting a body of representative specimen for an inventory/lexicon

### 9.9.17 Are you planning to collect corpora in the future?

Yes

### 9.9.18 Are you looking for available corpora? If so, what are your needs?

-

# 9.10 DFKI

## 9.10.1 Your name, e-mail

Michael Kipp
e-mail: kipp@dfki.de
url: http://www.dfki.de/~kipp/

## 9.10.2 Name of institute or organization

Deutsches Forschungszentrum für Künstliche Intelligenz
Stuhlsatzenhausweg 3
D-66123 Saarbrücken

## 9.10.3 Names of projects

## 9.10.4 How many corpora have you assembled?

1

## 9.10.5 How many subjects were captured per corpus (average number)?

2

## 9.10.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

( ) single
(X) dyads
( ) triads
( ) groups larger than 3
( ) any Wizard-of-Oz data?

## 9.10.7 What are the profiles of the subjects? (age groups, profession, gender,...)

~ 50 years, journalists, males

### 9.10.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*

*computer: screen, keyboard, touchpad, dataglove ...*

human: upper body, arms, face, often legs

### 9.10.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*

*computer: speech, embodied agent, graphics ...*

human: speech, hand/arm gesture, posture

### 9.10.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*

AVI

### 9.10.11 What coding scheme(s) do you use?

own

### 9.10.12 What coding tool(s) do you use primarily?

Anvil

### 9.10.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*

Sales presentation generation

### 9.10.14 Which spoken/sign languages occur in your corpora?

German

### 9.10.15 What are the costs for using your corpora?

( ) free for research use

( ) available for prices like _____ (example)

(X) not available

( ) not yet available, but maybe in the future

### 9.10.16 What purposes have your corpora been used for?

(X) exploratory studies

( ) evaluating hypotheses

( ) computer system evaluation

(X) training data (machine learning, classifiers, recognition)
(X) collecting a body of representative specimen for an inventory/lexicon

### 9.10.17 Are you planning to collect corpora in the future?

no

### 9.10.18 Are you looking for available corpora? If so, what are your needs?

no

# 9.11 Tilburg University

## 9.11.1 Your name, e-mail

Emiel J. Krahmer
e-mail: e.j.krahmer@kub.nl
url: http://fdlwww.kub.nl/~krahmer

## 9.11.2 Name of institute or organization

Tilburg University
Computational Linguistics
P.O. Box 90153
NL-5000 LE Tilburg
NL
fax: +31-13-466-3110

## 9.11.3 How many corpora have you assembled?

Only unimodal

## 9.11.4 Are you planning to collect corpora in the future?

Possibly

## 9.11.5 Are you looking for available corpora? If so, what are your needs?

Yes.
Facial expression, gestures + speech

# 9.12 University of Edinburgh HCRC

## 9.12.1 Your name, e-mail

John Lee
e-mail: J.Lee@ed.ac.uk
url: http://www.hcrc.ed.ac.uk/~john/

## 9.12.2 Name of institute or organization

University of Edinburgh
Human Communication Research Center (HCRC)
2 Buccleuch Place
EH8 9LW Edinburgh
GB
phone: +44 131 650 4420
fax: +44 131 650 4587

## 9.12.3 Names of projects

HCRC MapTAsk etc. design dialogues

## 9.12.4 How many corpora have you assembled?

Personally one

## 9.12.5 How many subjects were captured per corpus (average number)?

2 (in design dialogues)

## 9.12.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

( ) single
(X) dyads
( ) triads
( ) groups larger than 3
( ) any Wizard-of-Oz data?

### 9.12.7 What are the profiles of the subjects? (age groups, profession, gender,...)

Undergraduate students, typically.

Design dialogues : architect professors.

### 9.12.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*

*computer: screen, keyboard, touchpad, dataglove ...*

Human face and upper body.

### 9.12.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*

*computer: speech, embodied agent, graphics ...*

Human speech ; in some cases gaze.

### 9.12.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*

Digital Video (I think)

### 9.12.11 What coding scheme(s) do you use?

Various, XML based.

### 9.12.12 What coding tool(s) do you use primarily?

Home-built ; also e.g. MATE tools

### 9.12.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*

MapTask ; architectural design

### 9.12.14 Which spoken/sign languages occur in your corpora?

English

### 9.12.15 What are the costs for using your corpora?

( ) free for research use => MapTask via LDC

( ) available for prices like _____ (example)

(X) not available => DESIGN DIALOGUES

( ) not yet available, but maybe in the future

### 9.12.16 What purposes have your corpora been used for?

(X) exploratory studies
(X) evaluating hypotheses
( ) computer system evaluation
(X) training data (machine learning, classifiers, recognition)
( ) collecting a body of representative specimen for an inventory/lexicon

### 9.12.17 Are you planning to collect corpora in the future?

Yes

### 9.12.18 Are you looking for available corpora? If so, what are your needs?

Any corpora of graphics-based communication would be of interest

# 9.13 European Media Laboratory GmbH

## 9.13.1 Your name, e-mail

Rainer Malaka
e-mail: Rainer.Malaka@eml.villa-bosch.de
url: http://www.eml.org/english/staff/homes/malaka.html

## 9.13.2 Name of institute or organization

European Media Laboratory GmbH
Villa Bosch
Schloß-Wolfsbrunnenweg 33
D-69118 Heidelberg
fax: +49-6221 - 298

## 9.13.3 Names of projects

SmartKom, Embassi, Deep map

## 9.13.4 How many corpora have you assembled?

3 (for each prog.)

## 9.13.5 How many subjects were captured per corpus (average number)?

+/- 50

## 9.13.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single
( ) dyads
( ) triads
( ) groups larger than 3
( ) any Wizard-of-Oz data?

## 9.13.7 What are the profiles of the subjects? (age groups, profession, gender,...)

+/- Representative normal Germans + same US

### 9.13.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*

*computer: screen, keyboard, touchpad, dataglove ...*

Human Body

Embassi : electronic equipment

### 9.13.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*

*computer: speech, embodied agent, graphics ...*

Human gesture, facial expression

Computer speech

### 9.13.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*

### 9.13.11 What coding scheme(s) do you use?

Embassi : MATE+

Others: SmartKom

### 9.13.12 What coding tool(s) do you use primarily?

Own tool

### 9.13.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*

Tourism, Navigation, Map Task, Consumer electronics

### 9.13.14 Which spoken/sign languages occur in your corpora?

German / English

### 9.13.15 What are the costs for using your corpora?

(X) free for research use => AFTER NEGOTIATION

( ) available for prices like _____ (example)

( ) not available

( ) not yet available, but maybe in the future

### 9.13.16 What purposes have your corpora been used for?

(X) exploratory studies => WOZ

( ) evaluating hypotheses
( ) computer system evaluation
(X) training data (machine learning, classifiers, recognition) => FOR MM ANAPHORA RESOLUTION
( ) collecting a body of representative specimen for an inventory/lexicon

### 9.13.17 Are you planning to collect corpora in the future?

Yes

### 9.13.18 Are you looking for available corpora? If so, what are your needs?

Yes.
Deictic gesture + language.

# 9.14 SRI / LIMSI / LINC

## 9.14.1 Your name, e-mail

Jean-Claude Martin
e-mail: martin@limsi.fr
url: http://www.limsi.fr/Individu/martin/

## 9.14.2 Name of institute or organization

LIMSI
CNRS
B.P. 133
F-91403 Orsay
F
phone: +33-6 84 21 62 05
fax: +33-1 69 85 80 88

## 9.14.3 Names of projects

Invited researcher at SRI in 1997 / 1998 on a multimodal map WOZ (finished)
with L. Julia, A. Cheyer, J. Hobbs, A. Kehler, J. Bear
on the STIMULATE NSF funded project
Several other projects starting at Limsi in 2002 (Interactive TV, Multimodal dialogue with avatars) and at LINC-IUT de Montreuil (FAXCOM, MICAME)

## 9.14.4 How many corpora have you assembled?

1 at SRI

## 9.14.5 How many subjects were captured per corpus (average number)?

10

## 9.14.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single
( ) dyads
( ) triads
( ) groups larger than 3
(X) any Wizard-of-Oz data?

### 9.14.7 What are the profiles of the subjects? (age groups, profession, gender,...)

SRI employees

### 9.14.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*
*computer: screen, keyboard, touchpad, dataglove ...*
Human face and hand
Computer screen and pen

### 9.14.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*
*computer: speech, embodied agent, graphics ...*
Human speech, pen gesturing, computer graphics

### 9.14.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*
VHS

### 9.14.11 What coding scheme(s) do you use?

Home made

### 9.14.12 What coding tool(s) do you use primarily?

None

### 9.14.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*
tourism

### 9.14.14 Which spoken/sign languages occur in your corpora?

English

### 9.14.15 What are the costs for using your corpora?

( ) free for research use
( ) available for prices like _____ (example)
(X) not available
( ) not yet available, but maybe in the future

### 9.14.16 What purposes have your corpora been used for?

(X) exploratory studies

( ) evaluating hypotheses

( ) computer system evaluation

( ) training data (machine learning, classifiers, recognition)

( ) collecting a body of representative specimen for an inventory/lexicon

### 9.14.17 Are you planning to collect corpora in the future?

Yes

### 9.14.18 Are you looking for available corpora? If so, what are your needs?

Yes

Human-human multimodal communication with references to objects

# 9.15 MITRE

## 9.15.1 Your name, e-mail

Mark Maybury
e-mail: maybury@mitre.org
url: http://www.mitre.org/resources/centers/it/maybury/mark.html

## 9.15.2 Name of institute or organization

The MITRE Corporation
Information Technologies Division
3k-205
202 Burlington Road
MA 01730-1420 Bedford
USA
phone: +1-781-271-7230
fax: +1-781-271-2780

## 9.15.3 Are you planning to collect corpora in the future?

No

## 9.15.4 Are you looking for available corpora? If so, what are your needs?

Yes. Broadcast data

## 9.15.5 Are you interested to subscribe to our mailing list for future activities on multimodal data collection (workshops, questionnaire results etc.)?

Yes: greiff@mitre.org sboyilin@mitre.org

# 9.16 University of Ulster

## 9.16.1 Your name, e-mail

Paul Mc Kevitt
e-mail: p.mckevitt@ulst.ac.uk
url: http://www.infm.ulst.ac.uk/~paul/

## 9.16.2 Name of institute or organization

University of Ulster
Magee College - Faculty of Informatics
School of Computing and Intelligent Systems
BT48 7JL Derry/Londonderry
GB
phone: +44-28-7137-5433
fax: +44-28-7137-5470

## 9.16.3 Are you planning to collect corpora in the future?

No

## 9.16.4 Are you looking for available corpora? If so, what are your needs?

Yes.
Multimodal annotated: vision (deictic gestures + facial) + language

# 9.17 Universität Erlangen-Nürnberg

## 9.17.1 Your name, e-mail

Elmar Nöth
e-mail: noeth@informatik.uni-erlangen.de
url: http://www5.informatik.uni-erlangen.de/HTML/German/Persons/MA/noe

## 9.17.2 Name of institute or organization

Universität Erlangen-Nürnberg
Informatik 5
Martensstr. 3
D-91058 Erlangen
fax: +49 9131 303811

## 9.17.3 Names of projects

Smartkom (in progress)

## 9.17.4 How many corpora have you assembled?

1

## 9.17.5 How many subjects were captured per corpus (average number)?

<10

## 9.17.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single
( ) dyads
( ) triads
( ) groups larger than 3
( ) any Wizard-of-Oz data?

## 9.17.7 What are the profiles of the subjects? (age groups, profession, gender,...)

Adults, students + scientists

### 9.17.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*

*computer: screen, keyboard, touchpad, dataglove ...*

human face + speech

### 9.17.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*

*computer: speech, embodied agent, graphics ...*

emotions, f.e.

### 9.17.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*

QT

### 9.17.11 What coding scheme(s) do you use?

Partitur

### 9.17.12 What coding tool(s) do you use primarily?

CMU Video

### 9.17.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*

info kiosk

### 9.17.14 Which spoken/sign languages occur in your corpora?

German

### 9.17.15 What are the costs for using your corpora?

( ) free for research use

( ) available for prices like _____ (example)

( ) not available

(X) not yet available, but maybe in the future => request

### 9.17.16 What purposes have your corpora been used for?

(X) exploratory studies

( ) evaluating hypotheses

( ) computer system evaluation

(X) training data (machine learning, classifiers, recognition)
( ) collecting a body of representative specimen for an inventory/lexicon

### 9.17.17 Are you planning to collect corpora in the future?

yes

### 9.17.18 Are you looking for available corpora? If so, what are your needs?

yes

# 9.18 Oregon Graduate Institute

## 9.18.1 Your name, e-mail

Sharon Oviatt
e-mail: oviatt@cse.ogi.edu
url: http://www.cse.ogi.edu/CHCC/Personnel/oviatt.html

## 9.18.2 Name of institute or organization

Oregon Graduate Institute
Center for Human-Computer Communication
20000 N.W. Walker Rd.
OR 97006 Beaverton
USA
fax: +1-503-748-1548

## 9.18.3 Names of projects

NSF mobile interface design
DARPA command post of the future
ONR wearables project + VR

## 9.18.4 How many corpora have you assembled?

~12 corpora, 25-50 hours

## 9.18.5 How many subjects were captured per corpus (average number)?

25-50 subjects (each 1 hour)

## 9.18.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single
(X) dyads
( ) triads
(X) groups larger than 3
(X) any Wizard-of-Oz data?

### 9.18.7 What are the profiles of the subjects? (age groups, profession, gender,...)

Gender balanced

3 children age 7-10, rest adults

broad spectrum (white collar)

military (male)

### 9.18.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*

*computer: screen, keyboard, touchpad, dataglove ...*

child : face + upper body

adult : split screen (face + upper body) hands

### 9.18.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*

*computer: speech, embodied agent, graphics ...*

speech, manual gestures, acoustics -> dialogue level

comp: pen

cross-modal and within-modal synchronicity

### 9.18.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*

VHS

### 9.18.11 What coding scheme(s) do you use?

Study specific

### 9.18.12 What coding tool(s) do you use primarily?

Study specific, formerly by hand

### 9.18.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*

map task, education characters, games

### 9.18.14 Which spoken/sign languages occur in your corpora?

English, Japanese (translator)

### 9.18.15 What are the costs for using your corpora?

( ) free for research use
( ) available for prices like _____ (example)
( ) not available
(X) not yet available, but maybe in the future => 2-3 years

### 9.18.16 What purposes have your corpora been used for?

(X) exploratory studies
(X) evaluating hypotheses
(X) computer system evaluation
(X) training data (machine learning, classifiers, recognition)
( ) collecting a body of representative specimen for an inventory/lexicon

### 9.18.17 Are you planning to collect corpora in the future?

yes

### 9.18.18 Are you looking for available corpora? If so, what are your needs?

Theoretically yes
Own library

# 9.19 Università di Roma La Sapienza

## 9.19.1 Your name, e-mail

Catherine Pelachaud
e-mail: cath@dis.uniroma1.it
url: http://www.dis.uniroma1.it/~pelachau

## 9.19.2 Name of institute or organization

Università di Roma "La Sapienza"
Dipartimento di Informatica e Sistemistica
Via buonaroti, 12
I-00185 Roma
phone: +39-06-48 29 92 13
fax: +39-06-48 29 92 18

## 9.19.3 Names of projects

ISLE (EU)
MAGICSTER (EU)
COMMEDIA (CNR Italy)

## 9.19.4 How many corpora have you assembled?

1 by CNR Padova
many videos from TV shows
U of Rome Tre

## 9.19.5 How many subjects were captured per corpus (average number)?

4-5 speakers CNR Padova

## 9.19.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single => CNR Padova : single
(X) dyads => U of Rome Tre : singles + dyads
( ) triads
( ) groups larger than 3
( ) any Wizard-of-Oz data?

### 9.19.7 What are the profiles of the subjects? (age groups, profession, gender,...)

Large variety

### 9.19.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*

*computer: screen, keyboard, touchpad, dataglove ...*

human : face CNR

human : face, arms, body, gaze, U of Rome Tre

### 9.19.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*

*computer: speech, embodied agent, graphics ...*

U of Rome Tre : hand / arm gesture, facial expression, gaze…

CNR : lip shape parameters + audio (VCV)

### 9.19.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*

AVI + Elite parameters for lip movements

### 9.19.11 What coding scheme(s) do you use?

U of Rome Tre : Isabella Poggi's coding scheme

CNR : phonological parameters (lip width & height…)

### 9.19.12 What coding tool(s) do you use primarily?

U of Rome : hand coding (now with anvil)

CNR : Elite elaboration

### 9.19.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*

lip shape analysis + lip shape generation for talking heads

embodied agent : gesture + facial expression generation

### 9.19.14 Which spoken/sign languages occur in your corpora?

Italian

### 9.19.15 What are the costs for using your corpora?

( ) free for research use
( ) available for prices like _____ (example)
( ) not available
( ) not yet available, but maybe in the future
=> Do not know

### 9.19.16 What purposes have your corpora been used for?

(X) exploratory studies
( ) evaluating hypotheses
( ) computer system evaluation
( ) training data (machine learning, classifiers, recognition)
( ) collecting a body of representative specimen for an inventory/lexicon

### 9.19.17 Are you planning to collect corpora in the future?

Yes, to explore the expressiveness of gesture / facial expressions

### 9.19.18 Are you looking for available corpora? If so, what are your needs?

Yes video + audio

# 9.20 IRST

## 9.20.1 Your name, e-mail

Fabio Pianesi
e-mail: pianesi@itc.it
url: http://ecate.itc.it:1024/People/pianesi.html

## 9.20.2 Name of institute or organization

Istituto Per la Ricerca Scientifica e Tecnologica (IRST)
Ist.Trentino Di Cultura
Via Sommarive
I-38100 Povo
Italy
phone: +39 0461 314570
fax: +39 0461 314591

## 9.20.3 Names of projects

"Nespole !"

## 9.20.4 How many corpora have you assembled?

1

## 9.20.5 How many subjects were captured per corpus (average number)?

21

## 9.20.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single
( ) dyads
( ) triads
( ) groups larger than 3
( ) any Wizard-of-Oz data?

### 9.20.7 What are the profiles of the subjects? (age groups, profession, gender,...)

Adult, gender balanced, computer skilled

### 9.20.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*
*computer: screen, keyboard, touchpad, dataglove ...*
Upper body (not for annotation)
Screen, tablet log, pen

### 9.20.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*
*computer: speech, embodied agent, graphics ...*
log : pen gestures + speech
graphics

### 9.20.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*
digital videos (?)

### 9.20.11 What coding scheme(s) do you use?

Gesture scheme ; our dialogue acts + content

### 9.20.12 What coding tool(s) do you use primarily?

?

### 9.20.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*
Tourism

### 9.20.14 Which spoken/sign languages occur in your corpora?

Italian ; English ; German || French (future)

### 9.20.15 What are the costs for using your corpora?

( ) free for research use
( ) available for prices like _____ (example)
( ) not available

(X) not yet available, but maybe in the future => tablet log and video

### 9.20.16 What purposes have your corpora been used for?

(X) exploratory studies => archt + scenario
( ) evaluating hypotheses
(X) computer system evaluation
( ) training data (machine learning, classifiers, recognition)
( ) collecting a body of representative specimen for an inventory/lexicon

### 9.20.17 Are you planning to collect corpora in the future?

Yes ; spoken patient + doctor tele-interaction + tablet
(also Nespole!)

### 9.20.18 Are you looking for available corpora? If so, what are your needs?

Generally multimodal corpus community

# 9.21 DFKI

## 9.21.1 Your name, e-mail

Norbert Reithinger
e-mail: bert@dfki.de
url: http://www.dfki.de/~bert/

## 9.21.2 Name of institute or organization

Deutsches Forschungszentrum für Künstliche Intelligenz
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
phone: +49 681 302-5346
fax: +49 681 302-5020

## 9.21.3 Names of projects

Vermobil, Smartkom

## 9.21.4 How many corpora have you assembled?

None, only annotated or used

## 9.21.5 How many subjects were captured per corpus (average number)?

Vermobil > 100
Smartkom < 100

## 9.21.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single => Smartkom
(X) dyads => Vermobil (human-human)
( ) triads
( ) groups larger than 3
( ) any Wizard-of-Oz data?

## 9.21.7 What are the profiles of the subjects? (age groups, profession, gender,...)

Comparable to the general population

### 9.21.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*

*computer: screen, keyboard, touchpad, dataglove ...*

Smartkom : face, arms, posture of humans, location of gesture of computer

### 9.21.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*

*computer: speech, embodied agent, graphics ...*

human:

smartkom : speech, gesture, facial expression

vermobil : speech

computer : smartkom : speech

### 9.21.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*

Quicktime

### 9.21.11 What coding scheme(s) do you use?

Smartkom, verbmobil

### 9.21.12 What coding tool(s) do you use primarily?

Annotag, Anvil

### 9.21.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*

travel guide, EPG, cinema, home theatre

### 9.21.14 Which spoken/sign languages occur in your corpora?

Smartkom: German

Verbmobil: German, English, Japanese

### 9.21.15 What are the costs for using your corpora?

( ) free for research use

(X) available for prices like ____please ask BAS_____ (example)

( ) not available

( ) not yet available, but maybe in the future

### 9.21.16 What purposes have your corpora been used for?

(X) exploratory studies
(X) evaluating hypotheses
( ) computer system evaluation
(X) training data (machine learning, classifiers, recognition)
( ) collecting a body of representative specimen for an inventory/lexicon

### 9.21.17 Are you planning to collect corpora in the future?

Not in my group, as we are mostly annotators and consumers

### 9.21.18 Are you looking for available corpora? If so, what are your needs?

Yes. We need corpora with annotated gestures and postures.

# 9.22 DFKI

## 9.22.1 Your name, e-mail

Thomas Rist
e-mail: rist@dfki.de
url: http://www.dfki.de/~rist

## 9.22.2 Name of institute or organization

Deutsches Forschungszentrum für Künstliche Intelligenz
Geb. 43, Raum 0.19
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
phone: +49-681-302-5266
fax: +49-681-302-5020

## 9.22.3 Names of projects

Mlounge, Magicster, HMJ-RM (driver scenario)

## 9.22.4 How many corpora have you assembled?

Our group does not assemble corpora but our project partners

## 9.22.5 How many subjects were captured per corpus (average number)?

Teleconferencing scenario : 2-3 people
Groupmeeting scenario: 10 people

## 9.22.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single
( ) dyads
( ) triads
(X) groups larger than 3
( ) any Wizard-of-Oz data?

### 9.22.7 What are the profiles of the subjects? (age groups, profession, gender,...)

Adults, mixed gender, age 25-50 ?, professionals

### 9.22.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*

*computer: screen, keyboard, touchpad, dataglove ...*

room setting

### 9.22.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*

*computer: speech, embodied agent, graphics ...*

### 9.22.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*

? video tapes, 1 wav audio data at speech over IP audio conference (NISLab, Denmark)

### 9.22.11 What coding scheme(s) do you use?

### 9.22.12 What coding tool(s) do you use primarily?

### 9.22.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*

group work, teleconferencing, meeting

### 9.22.14 Which spoken/sign languages occur in your corpora?

### 9.22.15 What are the costs for using your corpora?

( ) free for research use

( ) available for prices like _____ (example)

( ) not available

( ) not yet available, but maybe in the future

=> need to check with partners

### 9.22.16 What purposes have your corpora been used for?

(X) exploratory studies => to get information for generation of behaviours for animated characters

( ) evaluating hypotheses

( ) computer system evaluation

( ) training data (machine learning, classifiers, recognition)

( ) collecting a body of representative specimen for an inventory/lexicon

### 9.22.17 Are you planning to collect corpora in the future?

Not myself

### 9.22.18 Are you looking for available corpora? If so, what are your needs?

Negotiation dialogues, multiparty, faces, full embodiment

# 9.23 LORIA

## 9.23.1 Your name, e-mail

Laurent Romary
e-mail: Laurent.Romary@loria.fr
url: http://www.loria.fr/~romary/

## 9.23.2 Name of institute or organization

LORIA - INRIA
Equipe Langue et Dialogue
BP 239
F-54506 Vandoeuvre-lès-Nancy

## 9.23.3 Names of projects

MIAMM, Langue et Dialogue

## 9.23.4 How many corpora have you assembled?

4

## 9.23.5 How many subjects were captured per corpus (average number)?

5-10

## 9.23.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single
(X) dyads
( ) triads
( ) groups larger than 3
(X) any Wizard-of-Oz data?

## 9.23.7 What are the profiles of the subjects? (age groups, profession, gender,...)

Mainly students (first years)

### 9.23.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*

*computer: screen, keyboard, touchpad, dataglove ...*

human face

computer touch pad

### 9.23.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*

*computer: speech, embodied agent, graphics ...*

human speech, hand/arm gesture

computer speech, graphics

### 9.23.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*

### 9.23.11 What coding scheme(s) do you use?

TEI, Mate reference annotation module, …

### 9.23.12 What coding tool(s) do you use primarily?

emacs

### 9.23.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*

instructional dialogues

### 9.23.14 Which spoken/sign languages occur in your corpora?

Spoken French

### 9.23.15 What are the costs for using your corpora?

(X) free for research use

( ) available for prices like _____ (example)

( ) not available

( ) not yet available, but maybe in the future

### 9.23.16 What purposes have your corpora been used for?

(X) exploratory studies

( ) evaluating hypotheses

( ) computer system evaluation
( ) training data (machine learning, classifiers, recognition)
(X) collecting a body of representative specimen for an inventory/lexicon

### 9.23.17 Are you planning to collect corpora in the future?

yes

### 9.23.18 Are you looking for available corpora? If so, what are your needs?

Transcribed MM dialogue with designation gestures

# 9.24 TU Berlin

## 9.24.1 Your name, e-mail

Dagmar Schmauks
e-mail: dagmar.schmauks@tu-berlin.de
url: http://www.tu-berlin.de/~afs/person/schmaukd.htm

## 9.24.2 Name of institute or organization

TU Berlin
Arbeitsstelle für Semiotik, Sek. TEL 16-1
Ernst-Reuter-Platz 7
D-10587 Berlin
phone: +49-30-314-79 440
fax: +49-30-314-27 638

## 9.24.3 Names of projects

Berlin dictionary of everyday gestures

## 9.24.4 How many corpora have you assembled?

1

## 9.24.5 How many subjects were captured per corpus (average number)?

2 subjects x 150 gestures

## 9.24.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single
( ) dyads
( ) triads
( ) groups larger than 3
( ) any Wizard-of-Oz data?

## 9.24.7 What are the profiles of the subjects? (age groups, profession, gender,...)

Young/old, male/female

### 9.24.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*

*computer: screen, keyboard, touchpad, dataglove ...*

human: face, arms, hand, body...

### 9.24.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*

*computer: speech, embodied agent, graphics ...*

human: hand/arm gesture

### 9.24.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*

VHS video

### 9.24.11 What coding scheme(s) do you use?

HamNoSys, natural language descriptions

### 9.24.12 What coding tool(s) do you use primarily?

### 9.24.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*

everyday gestures (emblems)

### 9.24.14 Which spoken/sign languages occur in your corpora?

NL

Everyday gestures

### 9.24.15 What are the costs for using your corpora?

( ) free for research use

( ) available for prices like _____ (example)

( ) not available

(X) not yet available, but maybe in the future

### 9.24.16 What purposes have your corpora been used for?

( ) exploratory studies

( ) evaluating hypotheses

( ) computer system evaluation

( ) training data (machine learning, classifiers, recognition)

(X) collecting a body of representative specimen for an inventory/lexicon

### 9.24.17 Are you planning to collect corpora in the future?

Yes : emblems in other languages

### 9.24.18 Are you looking for available corpora? If so, what are your needs?

### 9.24.19 Are you interested to subscribe to our mailing list for future activities on multimodal data collection (workshops, questionnaire results etc.)?

Yes (Massimo Serenari serenari@cs.tu-berlin.de)

# 9.25 Lotus

## 9.25.1 Your name, e-mail

Candy Sidner
e-mail: sidner@merl.com
url: http://www.merl.com/people/sidner/

## 9.25.2 Name of institute or organization

Previous affiliation where data was collected: Lotus Development Corp

## 9.25.3 Names of projects

Email collaborative agent

## 9.25.4 How many corpora have you assembled?

1

## 9.25.5 How many subjects were captured per corpus (average number)?

20

## 9.25.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

( ) single
(X) dyads
( ) triads
( ) groups larger than 3
( ) any Wizard-of-Oz data? => done as WOZ

## 9.25.7 What are the profiles of the subjects? (age groups, profession, gender,...)

Adults, Male & Female, Professionals

## 9.25.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*
*computer: screen, keyboard, touchpad, dataglove ...*

Video of user (keyboard, screen, hands), speech of user + wizard

### 9.25.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*
*computer: speech, embodied agent, graphics ...*
Speech annotated
Speech transcribed to text

### 9.25.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*
None

### 9.25.11 What coding scheme(s) do you use?

None

### 9.25.12 What coding tool(s) do you use primarily?

None

### 9.25.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*
User using own email

### 9.25.14 Which spoken/sign languages occur in your corpora?

English

### 9.25.15 What are the costs for using your corpora?

( ) free for research use
( ) available for prices like _____ (example)
(X) not available
( ) not yet available, but maybe in the future

### 9.25.16 What purposes have your corpora been used for?

( ) exploratory studies
( ) evaluating hypotheses
( ) computer system evaluation
(X) training data (machine learning, classifiers, recognition) => for speech recognition
( ) collecting a body of representative specimen for an inventory/lexicon
=> collecting data on task operation for building tasks models

### 9.25.17 Are you planning to collect corpora in the future?

### 9.25.18 Are you looking for available corpora? If so, what are your needs?

# 9.26 Mitsubishi Electric Research Laboratories

## 9.26.1 Your name, e-mail

Candy Sidner
e-mail: sidner@merl.com
url: http://www.merl.com/people/sidner/

## 9.26.2 Name of institute or organization

Mitsubishi Electric Research Laboratories
201 Broadway
MA 02139 Cambridge
USA

## 9.26.3 Names of projects

Collagen / Entertainment system

## 9.26.4 How many corpora have you assembled?

Corpus collection now underway

## 9.26.5 How many subjects were captured per corpus (average number)?

So far 5 users ; planned 40

## 9.26.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single
( ) dyads
( ) triads
( ) groups larger than 3
( ) any Wizard-of-Oz data?

## 9.26.7 What are the profiles of the subjects? (age groups, profession, gender,...)

Male & Female
English native speakers and a few non native speakers
Largely aged 30-50, some 20-30

Mostly computer knowledgeable users (researchers, people who use computer as tool).

### 9.26.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*
*computer: screen, keyboard, touchpad, dataglove ...*
none

### 9.26.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*
*computer: speech, embodied agent, graphics ...*
human: speech as said ; speech as recognition system hears
computer : speech output ; summary of gui actions

### 9.26.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*
Not yet chosen

### 9.26.11 What coding scheme(s) do you use?

None yet

### 9.26.12 What coding tool(s) do you use primarily?

None yet

### 9.26.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*
Entertainment system for TV viewing

### 9.26.14 Which spoken/sign languages occur in your corpora?

English

### 9.26.15 What are the costs for using your corpora?

( ) free for research use
( ) available for prices like _____ (example)
(X) not available
(X) not yet available, but maybe in the future

### 9.26.16 What purposes have your corpora been used for?

( ) exploratory studies
( ) evaluating hypotheses
( ) computer system evaluation
( ) training data (machine learning, classifiers, recognition)
( ) collecting a body of representative specimen for an inventory/lexicon

=> will be used to evaluate one aspect of NL design

### 9.26.17 Are you planning to collect corpora in the future?

Yes

### 9.26.18 Are you looking for available corpora? If so, what are your needs?

Yes, any 2 persons dialogues

# 9.27 DFKI

## 9.27.1 Your name, e-mail

Wolfgang Wahlster
e-mail: wahlster@dfki.de
url: http://www.dfki.de/~wahlster/

## 9.27.2 Name of institute or organization

Deutsches Forschungszentrum für Künstliche Intelligenz
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
phone: +49-681-302-5252
fax: +49-681-302-5341

## 9.27.3 Names of projects

Embassy (Sony, Grundig)
READY, REAL, Smartkom

## 9.27.4 How many corpora have you assembled?

3

## 9.27.5 How many subjects were captured per corpus (average number)?

READY: 60
REAL : 60
Smartkom : > 100

## 9.27.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single
( ) dyads
( ) triads
( ) groups larger than 3
(X) any Wizard-of-Oz data?

### 9.27.7 What are the profiles of the subjects? (age groups, profession, gender,...)

### 9.27.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*

*computer: screen, keyboard, touchpad, dataglove ...*

human: face, arms, hand, body...

Smartkom : all human + computer screen

### 9.27.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*

*computer: speech, embodied agent, graphics ...*

READY : human speech and gesture

Smartkom : human speech, gesture and facial expression

Smartkom : computer : speech and screen

### 9.27.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*

Quicktime

### 9.27.11 What coding scheme(s) do you use?

Smartkom scheme (BAS, BITS)

### 9.27.12 What coding tool(s) do you use primarily?

Anvil, BAS Partitur

### 9.27.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*

READY, REAL : Navigation in airport

READY: car breakdown

### 9.27.14 Which spoken/sign languages occur in your corpora?

## 9.27.15 What are the costs for using your corpora?

(X) free for research use => READY REAL SMARTKOM : 1 year delay after exploitation by consortium via ELRA/LDC

( ) available for prices like _____ (example)

( ) not available

( ) not yet available, but maybe in the future

## 9.27.16 What purposes have your corpora been used for?

(X) exploratory studies

(X) evaluating hypotheses

(X) computer system evaluation

(X) training data (machine learning, classifiers, recognition) => BAYESIAN NET TRAINING IN READY

( ) collecting a body of representative specimen for an inventory/lexicon

## 9.27.17 Are you planning to collect corpora in the future?

Yes in PEACH (joint project IRST-DFKI)

## 9.27.18 Are you looking for available corpora? If so, what are your needs?

Facial expression corpora combined with speech

# 9.28 IBM T.J. Watson Research Center

## 9.28.1 Your name, e-mail

Michelle Zhou
e-mail: mzhou@us.ibm.com
url: http://www.cs.columbia.edu/~zhou/

## 9.28.2 Name of institute or organization

IBM T.J. Watson Research Center
Intelligent Multimedia Interaction
30 Sawmill River Road
NY 10532 Hawthorne
USA

## 9.28.3 Names of projects

Responsive Information Architect (RIA)

## 9.28.4 How many corpora have you assembled?

3 corpora
#1: speech corpora (for speech generation)
#2: graphics + visual corpora (for graphics generation)
#3: multimedia corpora (for multimedia generation)

## 9.28.5 How many subjects were captured per corpus (average number)?

#1: speech : 60+ patterns
#2: graphics : 100 pictures : around 4000 objects
#3: multimedia : 80 slides

## 9.28.6 How many subjects were recorded at the same time (visible in the same frame)?

*(check all applicable options)*

(X) single FOR SPEECH
(X) dyads FOR SPEECH
( ) triads
( ) groups larger than 3
( ) any Wizard-of-Oz data?

=> our data (esp. multimedia data) is collected for the main purpose of generation so the captured scene is quite abstract

### 9.28.7 What are the profiles of the subjects? (age groups, profession, gender,...)

### 9.28.8 Which human body parts and/or computer media hardware are visible?

*human: face, arms, hand, body...*

*computer: screen, keyboard, touchpad, dataglove ...*

Computer screen

### 9.28.9 Which modalities are annotated?

*human: speech, hand/arm gesture, posture, facial expression ...*

*computer: speech, embodied agent, graphics ...*

Computer speech, text, graphics, images

### 9.28.10 What file formats do you use for primary data (video/sound)?

*(AVI, QuickTime, MPEG, Digital Video...)*

MPEG, JPEG, WAV

### 9.28.11 What coding scheme(s) do you use?

Own scheme

### 9.28.12 What coding tool(s) do you use primarily?

XML Spy

### 9.28.13 What is the application area?

*(tourism, navigation, museum, arts, map task ...)*

tourism, navigation

### 9.28.14 Which spoken/sign languages occur in your corpora?

English

### 9.28.15 What are the costs for using your corpora?

( ) free for research use

( ) available for prices like _____ (example)

( ) not available

(X) not yet available, but maybe in the future

### 9.28.16 What purposes have your corpora been used for?

(X) exploratory studies

( ) evaluating hypotheses

( ) computer system evaluation

(X) training data (machine learning, classifiers, recognition)

(X) collecting a body of representative specimen for an inventory/lexicon

### 9.28.17 Are you planning to collect corpora in the future?

Yes

### 9.28.18 Are you looking for available corpora? If so, what are your needs?

Yes

Need user multimodal input data (WOZ or from video) that could possibly capture the user interaction patterns so we can use it to learn these patterns for interpretation & generation purposes.

# 10 Appendix 2. Questionnaire

The letter and questionnaire used for the survey are presented below:

Dear [Contact Name],

The European Language Resources Association (ELRA) is in the process of surveying user needs and market prospective with regard to Multimodal Language Resources.

We would appreciate if you could take about 5 minutes to complete the questionnaire given below.

The information you provide will be part of a market needs study within the ISLE (International Standards for Language Engineering) project from the IST Programme of the European Commission.

All specific answers will remain strictly confidential. Only general statistics from this study will be distributed to survey participants and reproduced in reports on the topic of user needs.

Please return your completed questionnaire to mapelli@elda.fr

Valerie Mapelli
Project Leader - Technical & Commercial Services

Your name:
Your organisation:
E-mail:

Please place an X in the boxes below that are relevant to your organization.

1. MULTIMODAL LANGUAGE RESOURCES (LR) MARKET

Are you a user and/or a provider of Multimodal Language Resources (LR)?
[ ] User
[ ] Provider

Does your organization use LR:
[ ] that are produced internally?
[ ] that are produced by specific contracted vendors?

2. DATA

The resources you
[ ] acquire
[ ] offer
comprise:
[ ] Audio
[ ] Image
[ ] Video
Other (please specify) _____

3. RESOURCES WELL-SUITED FOR:

AUTHENTICATION
[ ] Face verification
[ ] Speech verification
[ ] User authentication
Other (please specify) _____

RECOGNITION
[ ] Face recognition
[ ] Speech recognition
[ ] Person recognition
[ ] Expression recognition
Other (please specify) _____

ANALYSIS
[ ] Body movements tracking (lips, hands, head, arms, legs, etc.)
[ ] Speech/lips correlation
Other (please specify) _____

SYNTHESIS
[ ] Talking heads
[ ] Avatars
[ ] Humanoid agents
[ ] Multimedia development
Other (please specify) _____

CONTROL
[ ] Voice control
[ ] Speech assisted video
Other (please specify) _____

OTHER
[ ] Information retrieval
Other (please specify) _____

4. APPLICATION AREAS

[ ] Education/Training
[ ] Research
[ ] Entertainment
[ ] Information Systems (e.g. banking, tourism, telecommunication)
[ ] Web Applications
Other (please specify) _____

5. LANGUAGES

[ ] Language independent
[ ] Language dependent

Which language(s) do you use/need LR for (please specify): _____

6. FOR PROVIDERS ONLY

Do you produce:
[ ] Language Resources
[ ] Prototypes
[ ] Tools
[ ] Generic Systems
[ ] Customized Applications
(Please add a description as attached file, brochure, etc.)

VALIDATION OF LR
When producing Multimodal LR, do you follow specific guidelines?
[ ] Internal specifications
[ ] External specifications - Please name them and give references:
[ ] None

Do you follow specific standards?
[ ] Yes
[ ] No
If Yes, please name them and give references:

Do you validate your LR (i.e. do you check deviation of the produced LR from initial specifications)?
[ ] Validation carried out internally
[ ] Validation carried out by independent/external organization/expert
[ ] No

If yes, does this lead to the production of validation reports?
[ ] Yes
[ ] No

7. EVALUATION FREQUENCY OF LR NEEDS

How often do you re-evaluate your LR needs and seek available databases?

[ ] Monthly
[ ] Once per quarter
[ ] Once per semester
[ ] Once per year
[ ] Once every 1-2 years
[ ] Never
[ ] Not applicable

8. DISTRIBUTION

Would you be willing to make your resources available to others according to a negotiated standardised distribution agreement?
[ ] Yes
[ ] No

If no, what are the reasons for not distributing your resources?
[ ] Technical
[ ] Commercial (pricing policy)
[ ] Legal (Copyright, Industrial/intellectual property rights)
[ ] Strategic
Other (please specify) _____

If yes, whom would you be ready to license your resources to?
[ ] End-users
[ ] Tool developers
[ ] Researchers
Other (please specify) _____

9. MARKET

How much do you spend for data acquisiton? _____ EURO/year
How much do you spend for data production? _____ EURO/year
How big is your expected purchasing budget for Multimodal LR?:
- now: _____
- in 3-5 years time: _____
Which is your estimate of the Multimodal LR market today?
Are there key-actors which are leading the development in your market sector, or are there several leading actors?
How do you see the future in your area?:
- size:
- new actors:
- new product development:
- investments required:
Other comments:

10. CONCLUDING QUESTIONS

Would you be willing to participate in future ELRA surveys?
[ ] Yes
[ ] No

Would you recommend this questionnaire to be mailed to someone from another organisation?
If yes, complete the following information:
Contact name:
Contact organisation:
Contact e-mail:

Please indicate here the amount of time it took you to complete this questionnaire (approximately _____ minutes)

We hope that ELRA can improve its services for you in the future through your participation in this survey. Thank you for taking the time to respond.

---Please return this questionnaire to mapelli@elda.fr ---

# 11 Appendix 3. Statistics

The statistics obtained from the Multimodal Language Resources Market Survey are given below. The number of responses per question are indicated in square brackets [ ], followed by the corresponding percentage with regard to the total number of replies.

**Number of respondents = 25**

**1. MULTIMODAL LANGUAGE RESOURCES (LR) MARKET**

*Are you a user and/or a provider of Multimodal Language Resources (LR)?*

| | # answers | % |
|---|---|---|
| - User | [17] | 68% |
| - Provider | [17] | 68% |

*Does your organization use LR:*

| | # answers | % |
|---|---|---|
| - that are produced internally? | [19] | 76% |
| - that are produced by specific contracted vendors? | [14] | 56% |

**2. DATA**

*The resources you*

| | # answers | % |
|---|---|---|
| - acquire | [15] | 60% |
| - offer | [9] | 36% |

*comprise:*

| | # answers | % |
|---|---|---|
| - Audio | [21] | 84% |
| - Image | [7] | 28% |
| - Video | [13] | 52% |
| Other (please specify) _____ | [8] | 32% |

Answer 1. Annotations (speech, gesture coding,…)

Answer 2. We are acquiring annotated audio for research and are currently collecting video ourselves

Answer 3. annotations of transcriptions of the speech part

Answer 4. Interface Surfaces

Answer 5. infrared video of gestures, coordinates of pointing gestures on the workspace

Answer 6. magnetometer and 3-d optical

Answer 7. written language

Answer 8. Text (annotated at various levels)

**3. RESOURCES WELL-SUITED FOR:**

*AUTHENTICATION*

| | # answers | % |
|---|---|---|
| - Face verification | [6] | 24% |
| - Speech verification | [8] | 32% |
| - User authentication | [5] | 20% |
| Other (please specify) _____ | [2] | 8% |

Answer 1. finger print and signature

Answer 2. Biometric Authentication (speech, signature)

*RECOGNITION*

| | # answers | % |
|---|---|---|
| - Face recognition | [7] | 28% |
| - Speech recognition | [14] | 56% |
| - Person recognition | [3] | 12% |
| - Expression recognition | [3] | 12% |
| Other (please specify) _____ | [4] | 16% |

Answer 1. mimic

Answer 2. music and other sounds

Answer 3. gesture recognition

Answer 4. gestures on a touch screen

*ANALYSIS*

| | # answers | % |
|---|---|---|
| - Body movements tracking (lips, hands, head, arms, legs, etc.) | [6] | 24% |
| - Speech/lips correlation | [7] | 28% |
| Other (please specify) _____ | [3] | 12% |

Answer 1. Cooperation between gesture and speech

Answer 2. acoustic, video, 3-d optical, midsagital magnetometry

Answer 3. written language analysis

*SYNTHESIS*

| | # answers | % |
|---|---|---|
| - Talking heads | [5] | 20% |
| - Avatars | [2] | 8% |
| - Humanoid agents | [5] | 20% |
| - Multimedia development | [6] | 24% |
| Other (please specify) _____ | [1] | 4% |

Answer 1. text generation

*CONTROL*

| | # answers | % |
|---|---|---|
| - Voice control | [7] | 28% |
| - Speech assisted video | [1] | 4% |
| Other (please specify) _____ | | |

*OTHER*

| | # answers | % |
|---|---|---|
| - Information retrieval | [14] | 56% |
| Other (please specify) _____ | [5] | 20% |

Answer 1. multimodal command languages (speech + gestures)

Answer 2. research into cross-modality issues, for the above applications

Answer 3. multimodal dialogue (speech + gesture)

Answer 4. Linguistic research

Answer 5. Information extraction, Text Summarisation

**4. APPLICATION AREAS**

| | # answers | % |
|---|---|---|
| - Education/Training | [9] | 36% |
| - Research | [21] | 84% |
| - Entertainment | [6] | 24% |
| - Information Systems (e.g. banking, tourism, telecommunication) | [14] | 56% |

| | | |
|---|---|---|
| - Web Applications | [10] | 40% |
| Other (please specify) _____ | [3] | 12% |

Answer 1. security

Answer 2. control of consumer devices

Answer 3. Media Archiving for Content Provider

**5. LANGUAGES**

| | # answers | % |
|---|---|---|
| - Language independent | [13] | 52% |
| - Language dependent | [18] | 72% |

*Which language(s) do you use/need LR for (please specify): _____*

| | # answers | % |
|---|---|---|
| French | [10] | 40% |
| English | [9] | 36% |
| German | [7] | 28% |
| Italian | [5] | 20% |
| Spanish | [4] | 16% |
| English US | [3] | 12% |
| English UK | [2] | 8% |
| Japanese | [2] | 8% |
| Arabic | [1] | 4% |
| Chinese | [1] | 4% |
| French Quebecois | [1] | 4% |
| Greek | [1] | 4% |
| Korean | [1] | 4% |
| Mandarin | [1] | 4% |
| Netherlands | [1] | 4% |
| Polish | [1] | 4% |
| Portuguese | [1] | 4% |
| Russian | [1] | 4% |
| Spanish Castilian | [1] | 4% |
| Spanish Catalan | [1] | 4% |
| Spanish Colombian | [1] | 4% |
| Spanish Latino | [1] | 4% |
| Swedish | [1] | 4% |

**6. FOR USERS ONLY**

*Which resources or applications does your organisation believe to need within 2-5 years (you may use the items listed in section 3)?*

| | # answers | % |
|---|---|---|
| Number of answers | [4] | 16% |

Answer 1. 16kHz (desk mic), 1000+ speaker corpora, Read news + spelling + digit (or equivalent) for above listed languages.

Answer 2. video with transcribed speech identification of other sounds, identification of objects and persons in images (with the position) coding of incrusted texts

Answer 3.

1. Spontaneous speech in many languages
2. Emotional or affected speech (maybe with additional physiological tracks
3. Multimodal recordings
4. Foreign accent
5. Natural conversation (between people, e.g. logs from call centres, but also multimodal)
6. Multi-speaker for Auditory Scene Analysis

7. Noisy speech/reverberated speech

8. Pronunciation Lexica

Answer 4. corpus of gesture and speech, face/gaze/hand gesture/speech speech/emotion

**7. FOR PROVIDERS ONLY**

*Do you produce:*

| | # answers | % |
|---|---|---|
| - Language Resources | [12] | 48% |
| - Prototypes | [6] | 24% |
| - Tools | [8] | 32% |
| - Generic Systems | [3] | 12% |
| - Customized Applications | [6] | 24% |
| (Please add a description as attached file, brochure, etc.) | [4] | 16% |

Answer 1. [Research databases]

Answer 2. see www.mpi.nl, www.mpi.nl/ISLE, www.mpi.nl/DOBES

Answer 3. for more information, you can consult: http://clic.fil.ub.es

Answer 4. www.limsi.fr

*VALIDATION OF LR*

*When producing Multimodal LR, do you follow specific guidelines?*

| | # answers | % |
|---|---|---|
| - Internal specifications | [15] | 60% |
| - External specifications | [5] | 20% |
| - Please name them and give references: | [4] | 16% |

Answer 1. ISLE guidelines

Answer 2. Eagles

Answer 3. Speecon, Orientel

Answer 4. SPEECON

| | # answers | % |
|---|---|---|
| - None | [1] | 4% |

*Do you follow specific standards?*

| | # answers | % |
|---|---|---|
| - Yes | [11] | 44% |
| - No | [4] | 16% |
| If Yes, please name them and give references: | [10] | 40% |

Answer 1. BAS, Eagles

Answer 2. TEI

Answer 3. Speecon, Orientel

Answer 4. XML/Unicode

Answer 5. see recent ISLE overview

Answer 6. EAGLES

Answer 7. MPEG-7

Answer 8. ESPS

Answer 9. SPEECON

Answer 10. XML

*Do you validate your LR (i.e. do you check deviation of the produced LR from initial specifications)?*

| | # answers | % |
|---|---|---|
| - Validation carried out internally | [14] | 56% |
| - Validation carried out by independent/external organization/expert | [1] | 4% |

| | | |
|---|---|---|
| - No | [3] | 12% |

*If yes, does this lead to the production of validation reports?*

| | # answers | % |
|---|---|---|
| - Yes | [6] | 24% |
| - No | [10] | 40% |

**8. EV ALUATION FREQUENCY OF LR NEEDS**

*How often do you re-evaluate your LR needs and seek available databases?*

| | # answers | % |
|---|---|---|
| - Monthly | [2] | 8% |
| - Once per quarter | [5] | 20% |
| - Once per semester | [3] | 12% |
| - Once per year | [4] | 16% |
| - Once every 1-2 years | [1] | 4% |
| - Never | - | - |
| - Not applicable | [5] | 20% |

**9. DISTRIBUTION**

*Would you be willing to make your resources available to others according to a negotiated standardised distribution agreement?*

| | # answers | % |
|---|---|---|
| - Yes | [12] | 48% |
| - No | [8] | 32% |

*If no, what are the reasons for not distributing your resources?*

| | # answers | % |
|---|---|---|
| - Technical | - | - |
| - Commercial (pricing policy) | [2] | 8% |
| - Legal (Copyright, Industrial/intellectual property rights) | [8] | 32% |
| - Strategic | [5] | 20% |
| Other (please specify) _____ | [2] | 8% |

Answer 1. privacy

Answer 2. some are already freely available on our web site for non commercial use

*If yes, whom would you be ready to license your resources to?*

| | # answers | % |
|---|---|---|
| - End-users | [8] | 32% |
| - Tool developers | [9] | 36% |
| - Researchers | [11] | 44% |
| Other (please specify) _____ | - | - |

**10. MARKET**

*How much do you spend for data acquisition? (in EURO/year)*

| | # answers | % |
|---|---|---|
| Number of answers | *[10]* | *40%* |

Answer 1. 40000

Answer 2. $50000

Answer 3. 20000

Answer 4. 0

Answer 5. 285910

Answer 6. 3000

Answer 7. 10000

Answer 8. 50000
Answer 9. 300000
Answer 10. 6000

*How much do you spend for data production? (in EURO/year)*

| | # answers | % |
|---|---|---|
| Number of answers | *[10]* | *40%* |

Answer 1. 20000
Answer 2. $75000
Answer 3. 72000
Answer 4. 15000
Answer 5. 1429551
Answer 6. 50000
Answer 7. 150000
Answer 8. 100000
Answer 9. huge amount
Answer 10. 6000

*How big is your expected purchasing budget for Multimodal LR?:*

| | # answers | % |
|---|---|---|
| - now: | [7] | 28% |

Answer 1. 4000
Answer 2. 1715461
Answer 3. 2000
Answer 4. 20000
Answer 5. 50000
Answer 6. 20000
Answer 7. small

| | # answers | % |
|---|---|---|
| - in 3-5 years time: | [5] | 20% |

Answer 1. 40000
Answer 2. 1715461-3500000
Answer 3. 100000
Answer 4. 50000
Answer 5. 100000
Answer 6. hope for increase

*Which is your estimate of the Multimodal LR market today?*

| | # answers | % |
|---|---|---|
| Number of answers | *[1]* | *4%* |

Answer 1: not studied yet

*Are there key-actors which are leading the development in your market sector, or are there several leading actors?*

| | # answers | % |
|---|---|---|
| Number of answers | *[6]* | *24%* |

Answer 1. not studied yet
Answer 2. several leading players
Answer 3. key actors
Answer 4. I think there are several leading actors
Answer 5. several
Answer 6. several

*How do you see the future in your area?:*

| | # answers | % |
|---|---|---|
| - size: | [3] | 12% |

Answer 1. increasing

Answer 2. unclear

Answer 3. big

| | # answers | % |
|---|---|---|
| - new actors: | [2] | 8% |

Answer 1. unclear

Answer 2. publishers will be key actors

| | # answers | % |
|---|---|---|
| - new product development: | [4] | 16% |

Answer 1. mainly

Answer 2. increasing

Answer 3. probable

Answer 4. will depend

| | # answers | % |
|---|---|---|
| - investments required: | [3] | 12% |

Answ er 1. very big

Answer 2. will be driven by new product developments (expected answer unclear)

Answer 3. small if content is already digital

| | # answers | % |
|---|---|---|
| Other comments: | [-] | 0% |

**11. CONCLUDING QUESTIONS**

*Would you be willing to participate in future ELRA surveys?*

| | # answers | % |
|---|---|---|
| - Yes | [18] | 72% |
| - No | [1] | 4% |

*Would you recommend this questionnaire to be mailed to someone from another organisation?*

| | # answers | % |
|---|---|---|
| If yes, complete the following information: | [2] | 8% |

Contact name:

Contact organisation:

Contact e-mail:

*Please indicate here the amount of time it took you to complete this questionnaire (approximately _____ minutes)*

| | # answers | % |
|---|---|---|
| 5 minutes | [3] | 18,75% |
| 7 minutes | [1] | 6,25% |
| 10 minutes | [3] | 18,75% |
| 15 minutes | [6] | 37,5% |
| 10-15 minutes | [1] | 6,25% |
| 20 minutes | [1] | 6,25% |
| 60 minutes | [1] | 6,25% |
| Total of answers | [16] | 64% of 25 respondents |