The Workshop Programme

Multimodal Corpora: Models Of Human Behaviour For The Specification And Evaluation Of Multimodal Input And Output Interfaces Tuesday 25th May 2004

http://lubitsch.lili.uni-bielefeld.de/MMCORPORA/

Centro Cultural de Belem, LISBON, Portugal, 25th may 2004
In Association with the 4th INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION LREC2004
http://www.lrec-conf.org/lrec2004/index.php
Main conference 26-27-28 May 2004

09:00 Welcome

Recommendations for Multimodal Annotation Tools and Schemes

09:15 Yang Shi, Travis Rose and Francis Quek
A System for Situated Temporal Analysis of Multimodal Communication

09:45 Laila Dybkjær, Niels Ole Bernsen
Recommendations for Natural Interactivity and Multimodal Annotation Schemes

Multimodal Systems Design and Evaluation

10:15 Knut Kvale, Jan Eikeset Knudsen, and John Rugelbak
A Multimodal Corpus Collection System for Mobile Applications

10:45 Louis Vuurpijl, Louis ten Bosch, Stéphane Rossignol, Andre Neumann, Norbert Pfleger, Ralf Engel
Evaluation of multimodal dialog systems

11:15 – 11:30 : Coffee break

Multimodal Systems Design and Evaluation

11:30 Niels Ole Bernsen
Measuring Relative Target User Group Success in Spoken Conversation for Edutainment Applications

12:00 Johanna Höysniemi, Perttu Hämäläinen
Describing children's intuitive movements in a perceptive adventure game

12:30 – 13:15 : Discussion

13:15 – 14:30 : Lunch

Invited Talk

14:30 Annelies Braffort
Corpora for Sign Language Studies

Coding Schemes and Multimodal Communication

15:15 Loredana Cerrato
A coding scheme for the annotation of feedback phenomena in conversational speech

15:45 – 16:15 : Coffee break

16:15 Emanuela Magno Caldognetto, Isabella Poggi, Piero Cosi, Federica Cavicchio, G. Merola
Multimodal Score: an ANVIL Based Annotation Scheme for Multimodal Audio-Video Analysis

16:45 Jonas Beskow, Loredana Cerrato, Björn Granström, David House, Magnus Nordstrand, Gunilla Svanfeldt
The Swedish PF-Star Multimodal Corpora

17:15 Jan Peter de Ruiter
On the primacy of language in multimodal communication

17:45 – 18:30 : Panel discussion

18:30 : End of workshop

Workshop Organisers

Jean-Claude MARTIN, LIMSI-CNRS
Elisabeth Den OS, MPI
Peter KÜHNLEIN, Univ. Bielefeld
Lou BOVES, Univ. Of Nijmegen
Patrizia PAGGIO, CST
Roberta CATIZONE, Univ. Sheffield

Workshop Programme Committee

Elisabeth Ahlsén Univ. Göteborg
Jens Allwood Univ. Göteborg
Elisabeth André Univ. Augsburg
Niels Ole Bernsen NIS Lab
Lou Boves Univ. Nijmegen
Stéphanie Buisine LIMSI-CNRS
Roberta Catizone Univ. Sheffield
Loredana Cerrato TMH-CTT
Piero Cosi ISTC-SPFD CNR
Jan-Peter de Ruiter MPI
Els den Os MPI
Laila Dybkjær NIS Lab
David Horowitz Vox Generation
Bart Jongejan CST
Alfred Kranstedt SFB 360
Steven Krauwer Univ. Utrecht
Peter Kühnlein SFB 360
Knut Kvale Telenor R&D
Myriam Lamolle LINC-IUT
Joseph Mariani LIMSI-CNRS
Jean-Claude Martin LIMSI-CNRS
Jan-Torsten Milde FH Aalen
Sharon Oviatt CHCC
Patrizia Paggio CST
Catherine Pelachaud Univ. Paris
Janienke Sturm Univ. Nijmegen

Table of Contents

Author Index

Introduction

Multimodal Corpora: Models Of Human Behaviour For The Specification And Evaluation Of Multimodal Input And Output Interfaces

The primary purpose of this one day workshop is to share information and engage in the collective planning for the future creation of usable multidisciplinary multimodal resources. It will focus on the following issues regarding multimodal corpora: how researchers build models of human behaviour out of the annotations of video corpora, how they use such knowledge for the specification of multimodal input (e.g. merging users' gestures and speech) and output (e.g. specification of believable and emotional behaviour in Embodied Conversational Agents) in human computer interfaces, and finally how they evaluate multimodal systems (e.g. full system evaluation and glass box evaluation of individual system components).

The topics which the workshop aimed to address are:

- Models of human multimodal behaviour in various disciplines
- Integrating different sources of knowledge (literature in socio-linguistics, corpora annotation)
- Specifications of coding schemes for annotation of multimodal video corpora
- Parallel multimodal corpora for different languages
- Methods, tools, and best practice procedures for the acquisition, creation, management, access, distribution, and use of multimedia and multimodal corpora
- Methods for the extraction and acquisition of knowledge (e.g. lexical information, modality modelling) from multimedia and multimodal corpora
- Ontological aspects of the creation and use of multimodal corpora
- Machine learning for and from multimedia (i.e., text, audio, video), multimodal (visual, auditory, tactile), and multicodal (language, graphics, gesture) communication
- Exploitation of multimodal corpora in different types of applications (information extraction, information retrieval, meeting transcription, multisensorial interfaces, translation, summarisation, www services, etc.)
- Multimedia and multimodal metadata descriptions of corpora
- Applications enabled by multimedia and multimodal corpora
- Benchmarking of systems and products; use of multimodal corpora for the evaluation of real systems
- Processing and evaluation of mixed spoken, typed, and cursive (e.g., pen) language processing
- Automated multimodal fusion and/or generation (e.g., coordinated speech, gaze, gesture, facial expressions)
- Techniques for combining objective and subjective evaluations, and for making evaluations cost-effective, predictive and fast

Multimodal resources feature the recording and annotation of several communication modalities such as speech, hand gesture, facial expression, body posture, graphics. Several researchers have been developing such multimodal resources for several years, often with a focus on a limited set of modalities or on a given application domain.

A number of projects, initiatives and organisations have addressed or will address multimodal resources with a federative approach:

- 2005 : Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research, 30 Aug. - 2 Sep. 2005 Wageningen, The Netherlands http://www.noldus.com/events/
- 2004 : A tutorial "Your Next Usability Lab:Tools for Data Collection and Analysis" will be given by dr. Lucas Noldus and Tobias Heffelaar during the CHI 2004 conference http://www.chi2004.org/program/prog_tutorials.html#t17
- 2004 : At the Dagstuhl seminar on Evaluating Embodied Conversational Agents 15-19 March, 2004, a Working group on ECA and Human-Human interaction http://homepages.cwi.nl/~zsofi/eeca/WG1.ppt
- 2004 : The European 6th Framework program (FP6), includes multilingual and multisensorial communication as one of the major R&D issue, and the evaluation of technologies appears as a specific item in the Integrated Project instrument presentation http://www.cordis.lu/ist/so/interfaces/home.html. Recently started NoE and IP :
    - Humaine NoE http://emotion-research.net/ WP5 on databases
    - AMI project http://www.amiproject.org/
    - Similar NoE http://www.similar.cc/
    - PASCAL Pattern Analysis statistical modeling and computational learning http://www.cs.rhul.ac.uk/colt/pascalprop.doc
- 2002 : At LREC2002, a workshop had addressed the issue of "Multimodal Resources and Multimodal Systems Evaluation" http://www.limsi.fr/Individu/martin/wslrec2002/MMWorkshopReport.doc
- 2002 : Measuring Behavior 2002 Special Interest Group « Tools and techniques for the study of multimodal communication: speech, gesture and facial expression » http://www.noldus.com/events/mb2002/program/sig_4.html
- 2001 : A Working Group at the Dagstuhl Seminar on Multimodality recorded, in November 2001, 28 questionnaires from researchers on multimodality, from which 21 have been announcing their attention to record other multimodal corpora in the future. (http://www.dfki.de/~wahlster/Dagstuhl_Multi_Modality/)
- 2000 : At LREC2000, a 1st workshop had addressed the issue of multimodal corpora, focussing on meta-descriptions and large corpora http://www.mpi.nl/world/ISLE/events/LREC%202000/LREC2000.htm
- 2000 : NIMM was a work group on Natural Interaction and MultiModality which ran under the IST-ISLE project (http://isle.nis.sdu.dk/). In 2001, NIMM compiled a survey of existing multimodal resources (more than 60 corpora are described in the survey), coding schemes and annotation tools. The ISLE project was developed both in Europe and in the USA (http://www.ldc.upenn.edu/sb/isle.html)
- Surveys about multimodal annotation coding schemes and tools :
    - EcorporaA (European Language Resources Association) launched in November 2001 a survey about multimodal corpora including marketing aspects (http://www.icp.inpg.fr/EcorporaA/).
    - COCOSDA http://www.slt.atr.co.jp/cocosda/beijing/multi-modal.files/frame.htm
    - LDC, MITRE
- Institutes and consortium
    - National Institute of Standards and Technology, http://www.nist.gov
    - Linguistic Data Consortium, http://www.ldc.upenn.edu
    - International Committee for the Co-ordination and Standardisation of Speech Databases and Assesment Techniques, http://www2.slt.atr.co.jp/cocosda
    - European Language Resources Association, http://www.elda.fr. The European Language Resources Distribution Agency, http://www.elda.fr, is ELRA's operational body.
    - EcorporaA (European Language Resources Association) launched in November 2001 a survey about multimodal corpora including marketing aspects (http://www.icp.inpg.fr/EcorporaA/).
- Projects
    - Natural Interactivity Tools Engineering, http://nite.nis.sdu.dk
    - Architecture and Tools for Linguistic Analysis Systems, http://www.nist.gov/speech/atlas
    - Intera project
    - NIST Automatic Meeting Transcription Project (http://www.nist.gov/speech/test_beds/mr_proj): "The National Institute of Standards and Technology (NIST) held an all-day workshop entitled "Automatic Meeting Transcription Data Collection and Annotation" on 2 November 2001. "The workshop addressed issues in data collection and annotation approaches, data sharing, common annotation standards and tools, and distribution of corpora. ... To collect data representative of what might be expected in a functional meeting room of the future, [NIST has] created a media- and sensor-enriched conference room containing a variety of cameras and microphones."
    - ATLAS (http://www.nist.gov/speech/atlas): Also at NIST, "ATLAS (Architecture and Tools for Linguistic Analysis Systems) is a recent initiative involving NIST, LDC and MITRE. ATLAS addresses an array of applications needs spanning corpus construction, evaluation infrastructure, and multimodal visualisation."

- TALKBANK (http://www.talkbank.org): TALKBANK is funded by the National Science Foundation (NSF). Its goal "is to foster fundamental research in the study of human and animal communication. TalkBank will provide standards and tools for creating, searching, and publishing primary materials via networked computers." One of the six sub-groups is concerned with communication by gesture and sign.

Yet, existing annotation of multimodal corpora until now have been done mostly on an individual basis, each researcher or team focusing on its own needs and knowledge about modality specific coding schemes or application examples. Thus, there is a lack of real common knowledge and understanding of how to proceed from annotations to usable models of human multimodal behaviour and how to use such knowledge for the design and evaluation of multimodal input and embodied conversational agent interfaces.

Furthermore, the evaluation of multimodal interaction poses different (and very complex) problems than the evaluation of monomodal speech interfaces or WYSIWYG direct interaction interfaces. There are a number of recently finished and ongoing projects in the field of multimodal interaction in which attempts have been made to evaluate the quality of the interfaces in all meanings that can be attached to the term 'quality'. There is a widely felt need in the field for exchanging information on multimodal interaction evaluation with researchers in other projects. One of the major outcomes of this workshop should be better understanding of the extent to which evaluation procedures developed in one project generalise to other, somewhat related projects.

Out of 15 submitted papers, 10 papers were accepted for long presentation. They enable the workshop to cover several dimensions of multimodal corpora:

- *Multimodal phenomena* : verbal and gestural feedback, visual correlates of emotional speech, facial animation, human movement notation
- *Multimodal corpora collection and analysis* : guidelines, annotation schemes
- *Multimodal system design and evaluation* : wizard of oz prototyping, animated agent systems and multimodal spoken dialogue systems, evaluation metrics
- *Application areas* : Edutainment systems (computer games, children), multi-participant meetings

We expect the output of the workshop to be the following: a deeper knowledge of the potential of models of human multimodal behaviour for the specification and evaluation of multimodal input and output interfaces, better understanding of challenging issues in the usability of multimodal corpora, and the fostering of a multidisciplinary community of multimodal researchers and multimodal interface developers.

Jean-Claude MARTIN, LIMSI-CNRS, France, martin@limsi.fr
Elisabeth Den OS, MPI, Netherlands, Els.denOs@mpi.nl
Peter KÜHNLEIN, Univ. Bielefeld, Germany, p@uni-bielefeld.de
Lou BOVES, Univ. Of Nijmegen, Netherlands, L.Boves@let.kun.nl
Patrizia PAGGIO, CST, Denmark, patrizia@cst.dk
Roberta CATIZONE, Sheffield, United Kingdom, roberta@dcs.shef.ac.uk

A System for Situated Temporal Analysis of Multimodal Communication

Yang Shi, Travis Rose and Francis Quek

Vision Interface & Sys. Lab. (VISLab) CSE Dept., Wright State U., Dayton, OH
quek@cs.wright.edu

Abstract

We present our multimedia *Visualization for Situated Temporal Analysis* (VisSTA) system that facilitates analysis of multi-modal human communication incorporating video, audio, speech transcriptions, and coded multimodal (e.g. gesture and gaze) data. VisSTA is based on the *Multiple Linked Representation* strategy and keeps the user temporally situated by ensuring tight linkage among all representational components. The system features multiple representations, which include a hierarchical video-shot organization, a variety of animated graphs, animated multi-tier text transcripts, and an avatar representation. VisSTA is a multi-video system permitting simultaneous playing of multiple synchronized video streams that are time-locked to other data components. An integrated observation database system is included in VisSTA for storing the results of data analysis.

Introduction

We present a system for coding, annotating and visualizing multimodal language data that comprises multi-stream video, audio, time-continuous signal data, and symbolic time-occupying entities. The *Visualization for Situated Temporal Analysis* (VisSTA) system is designed through the course of a series of collaborative projects involving psycholinguists, computer vision and speech processing researchers and human-computer interaction researchers. Each of these researchers contributed requirements and insights to the design of this evolving system. As a result, this system is a novel data visualization and video annotation tool that has been designed to support analysis of multimodal communication. Our intended users range from psycholinguists to anyone with a need for general purpose video annotation and analysis.

We shall discuss the overarching design, architecture, and interface components of VisSTA in turn. We will also show an example of VisSTA's use as a coding tool. VisSTA is X-windows-based and currently runs on Silicon Graphics O2 workstations. A new version of VisSTA is being developed for Mac OS X.

Design

VisSTA is an integration framework that combines a variety of representational components to manipulate specific data types, and ties all the components together using a common control bus.

An interactive system may be viewed as a conduit of communication between the human and the machine (Mayhew 1992). Modern psychology and linguistics theories of discourse stress the importance of maintaining a state of 'situatedness' for communication to be successful (Gordon, Grosz et al. 1993; Brennan 1995; Delin 1995). For multimodal communication, temporal synchrony and relationships are critical. Hence, VisSTA employs temporal cohesion as the key dimension of situatedness. All representational components are linked by time synchrony.

VisSTA is an example of *Multiple Linked Representations* of dynamic components (Kozma, Russel et al. 1996; Quek, Bryll et al. 2002) in which each representation reinforces the situatedness condition. Furthermore, each representation in the system is active, thereby enabling *multiple-point access* to the underlying data. Within the multiple linked representations scheme, all components are able to function as both controllers and displays. For example, the user can mark a run of text in the speech transcription representation and have all other components play the segment. As the video is played, the transcription window also highlights each word as it comes into temporal focus. A component called the *VCR-style Control Panel* is the visual representation of the central control component.

Figure 1 shows the VisSTA interface with all the major components open. The essence of our *Multiple Linked Representation* strategy is that all the representational components are synchronized with the *current time focus*.

Interface Design & Implementation

The essential feature of VisSTA is that all components constantly animate to keep a common current time focus throughout the system. We discuss how the relevant data and current time focus are represented for each component. VisSTA provides a unified environment for coding and displaying a range of synchronized datatypes

Hierarchical Shot-Keyframe Representation

The *Hierarchical Shot-Keyframe Representation* and *Shot-Keyframe Hierarchy Editor* permit the interactive segmentation of video into a hierarchy of shots. A shot s is defined simply as a consecutive series of video frames f_B ...

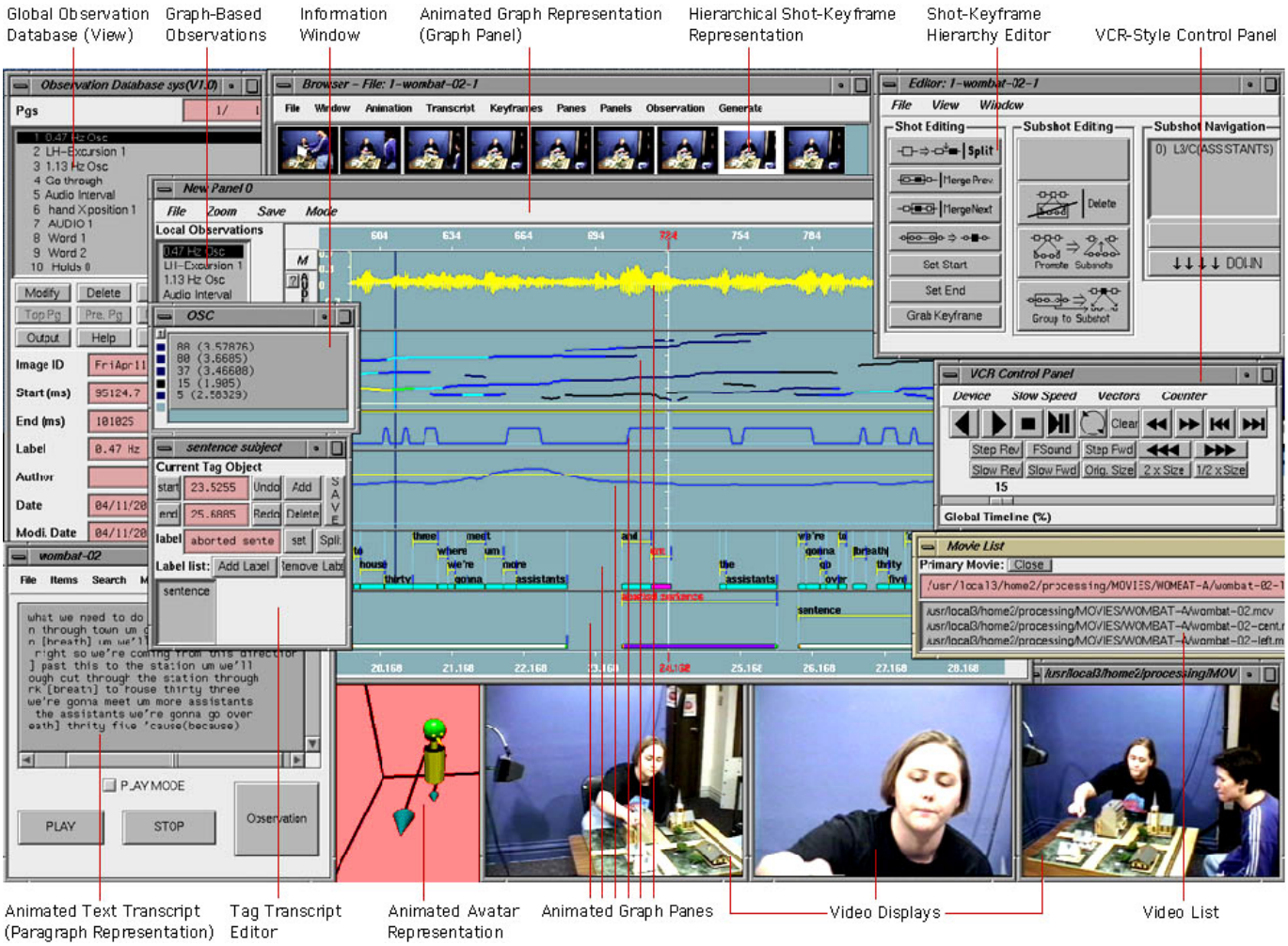


Figure 1: VisSTA with All Components

f_E. Although the frame is the smallest physical unit of video, the shot is the smallest possible semantic video unit (Taskiran, Bouman et al. 2000). Each shot may be given a short label and a longer textual description, and is represented visually by a keyframe. Shots may be subdivided hierarchically. At any time, the Hierarchical Shot-Keyframe Representation displays the keyframes of a consecutive set of shots at some level of the hierarchy. This hierarchical representation is useful for dividing video into larger discourse segments for analysis and for representing strict discourse hierarchies.

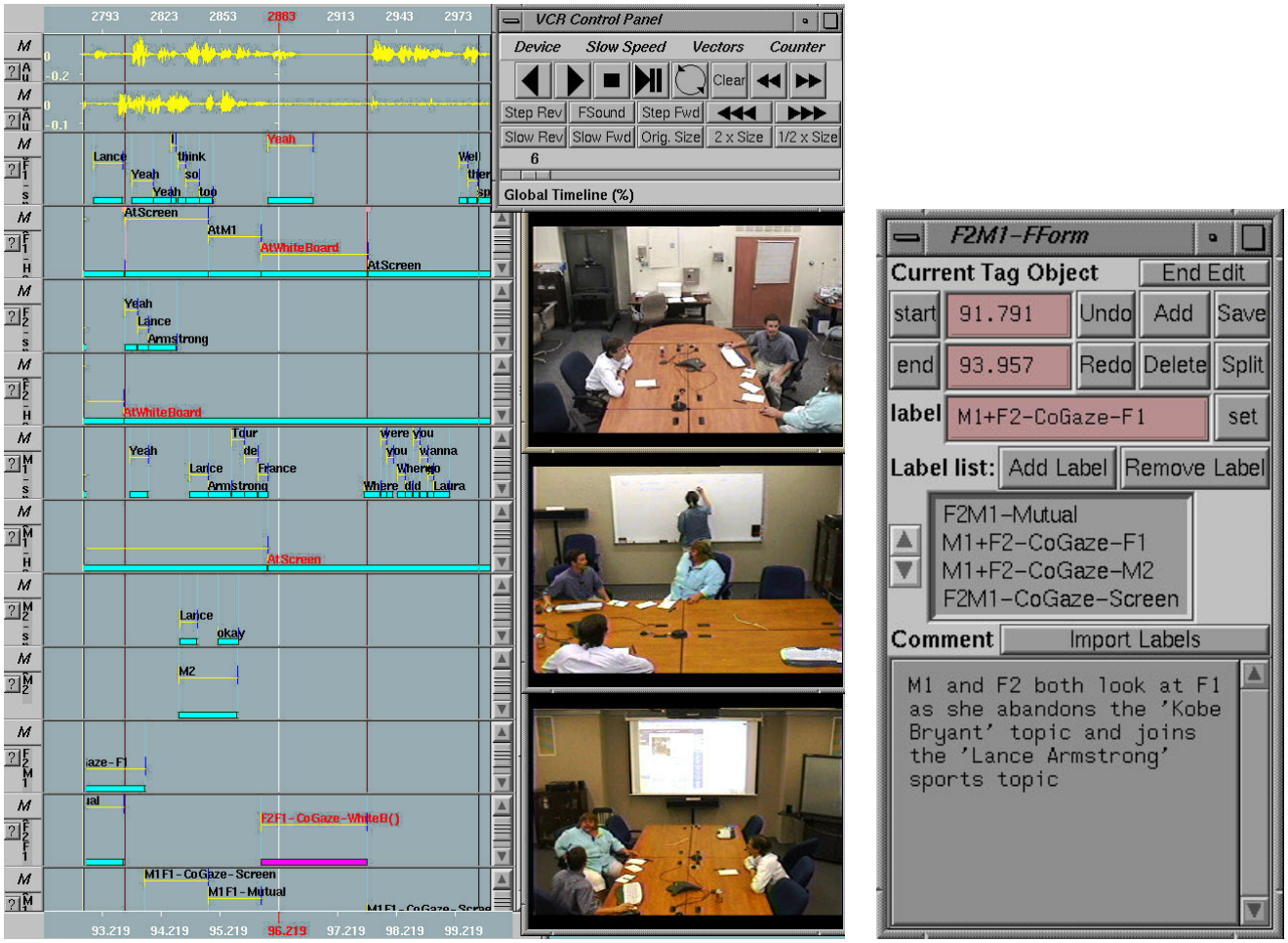
Animated Text Transcript Window

The *Animated Text Transcript* window, shown in Figure 1, displays speech transcripts in paragraph format. As the video plays the appropriate word is highlighted and the window scrolls to keep temporally situated in accordance with the *Multiple Linked Representation* strategy.

Animated Graph Representation Panel

The Animated Graph Representation Panel, shown in Figure 1, is the key component of VisSTA for multimodal analysis. The panel is a container that can hold a set of Animated Graph Panes, which are used to plot various types of data. Any data plot can be displayed in a pane and the panes can be freely arranged in the panel. The panes are vertically aligned to reflect their temporal relationship (i.e. they are all time-aligned). The current time focus is represented as a single vertical line drawn through the center of all the panes. Each pane allows random access to the data via the user clicking any point on a data plot.

Figure 1 shows a panel containing four different panes. Panes can be used both for visualization of graphical data and for annotation of video. This approach allows us to freely mix graphical plots and annotations inline and in time-aligned fashion. The topmost pane displays an audio waveform. The second pane displays a special multiple trace data format that shows a frequency-time plot of gestural oscillation. The third and fourth panes are continuous graph plots, which for example can show right and left hand positions. The last two panes display annotation data. Each panel uses a simulated notebook to manage its own local observation database. An observation can be set graphically in any pane by clicking and dragging to define its time-occupancy. Panels, and the panes within them, can be thought of as music score representations.



(a) (b)

Figure 2. The VisSTA Coding Interface: (a) Multi-view meeting videos and panel for annotation; (b) Editor for time-occupying entities

Global Observation Window

The *Global Observation Window* displays a list of global observations. The Global Observation Database contains observations imported from various local observations made in panels. The Global Observation Database can be searched through a query window.

Sample Analysis

We show an example of using VisSTA to do annotation of a real meeting. This annotation was intended as a preliminary analysis of multimodal communication. VisSTA can be used for coding and analysis of the multimodal communication and interaction in multi-participant meetings.

Our data comes from a multi-camera pilot microcorpus collected at the National Institutes for Standards and Technology (NIST) meeting room facility on July 27, 2003. Four individuals (two male, two female) were tasked with choosing the top five news stories for the week of July 22, 2003.

Figure 2(a) shows VisSTA displaying the meeting videos from multiple time-synchronized viewpoints, and the graphical representation of events as time-occupying entities. The start time, end time, label and comment of specific events can be edited via the interface shown in Figure 2(b). Figure 2(a) shows our coding interface with multiple tiers in a music score representation as follows: two channels of audio, participant's speech and head orientation, as well as gaze patterning between subjects in a pairwise fashion (e.g. co-gaze at each other, sharing the same gaze-attractor, etc.). From this example it can be seen that VisSTA provides a powerful interface for multimodal analysis.

Advantages

Besides VisSTA, other tools designed for video visualization and annotation include ANVIL (Kipp 2003), SignStream (Neidle 2002), MacSHAPA (Sanderson, Scott et al. 1994), MultiTool (Nivre, Allwood et al. 1998),

SyncWriter (Hanke and Prillwitz 1994), CLAN (CHILDES 2003), and MediaTagger (Wittenberg 2000). VisSTA is designed to be a flexible and general-purpose application that integrates annotation and data visualization. VisSTA supports a variety of annotation methods including observations, text transcripts and video shot hierarchies. VisSTA can handle arbitrary time-series data (e.g. hand position, head position and orientation), audio, voice F_0 etc., and all the plots can be freely arranged.

Conclusion

We have presented a multimedia database system for analysis of natural multi-modal language. VisSTA is a principled interface based on the *Multiple Linked Representation Model*. The model helps the user to analyze and visualize multi-modal data in multiple representations in a time-situated fashion. The system maintains temporal situatedness by keeping all components synchronized with the current time focus. While VisSTA is a much larger system comprising more components for visualizing a variety of data types, we focused on the music score representation as the key component for interactive multimodal coding.

Acknowledgements

This research has been supported by the U.S. National Science Foundation STIMULATE program, Grant No. IRI-9618887: *Gesture, Speech, and Gaze in Discourse Segmentation*; the NSF KDI program, Grant No. BCS-9980054: *Cross-Modal Analysis of Signal Sense: Multimedia Corpora and Tools for Gesture, Speech, and Gaze Research*; and the NSF ITR program, Grant No. ITR-0219875: *Beyond the Talking Head and Animated Icon: Behaviorally Situated Avatars for Tutoring*; and the Advanced Research and Development Activity ARDA VACEII grant 665661: *From Video to Information: Cross-Model Analysis of Planning Meetings*. We thank John Garofolo at NIST for providing access to the meeting room data. We also thank Susan Duncan and David McNeill for their continuing collaboration and their comments and insights into the NIST video coding.

References

Brennan, S. E. (1995). "Centering attention in discourse." *Language & Cog. Processes* 10(2): 137-167.

CHILDES (2003). *Using CLAN*, CHILDES Project, Carnegie Mellon U. http://childes.psy.cmu.edu/clan/.

Delin, J. (1995). "Presupposition & shared knowledge in it – clefts." *Language & Cog. Processes* 10(2): 97-120.

Gordon, P. C., B. J. Grosz, et al. (1993). "Pronouns, names, and the centering of attention in discourse." *Cog. Science* 17(3): 311-347.

Hanke, T. and S. Prillwitz (1994). "SyncWRITER: Integrating Video into the Transcription & Analysis of Sign Language". *Proc 4th European Congress on Sign Language Research*, Munich, Germany.

Kipp, M. (2003). *Anvil: Annotation of Video and Spoken Language.*

Kozma, R. B., J. Russel, et al. (1996). "The Use of Multiple, Linked Representations to Facilitate Science Understanding". *Int'l Perspectives on the Design of Tech-Supported Learning Environments*. S. Vosniadou, et.al (Eds). Mahwah, New Jersey.

Mayhew, D. (1992). **Principles and Guidelines in Software User Interface Design.** Prentice-Hall Inc.

Neidle, C. (2002). "SignStream™: A Database Tool for Research on Visual-Gestural Language." *Sign Language and Linguistics* 4(1/2): 203-214.

Nivre, J., J. Allwood, et al. (1998). "Towards Multimodal Spoken Language Corpora: TransTool & SyncTool." *Proc. Wksp on Partially Automated Techniques for Transcribing Naturally Occurring Speech at COLING-ACL '98*, Montreal, Canada.

Quek, F., R. Bryll, et al. (2002). "A multimedia database system for temporally situated perceptual psycholinguistic analysis." *Multimedia Tools & Apps.* 18(2): 91-113.

Sanderson, P. M., J. J. P. Scott, et al. (1994). "MacSHAPA and the enterprise of Exploratory Sequential Data Analysis (ESDA)." International Journal of Human-Computer Studies 41: 633-668.

Taskiran, C. M., C. A. Bouman, et al. (2000). "The ViBE Video Database System: an Update & Further Studies". *SPIE/IS&T Conf. Storage & Retrieval for Media Databases*, San Jose, CA.

Wittenberg, P. (2000). *MediaTagger*, Max Planck Inst. for Psycholinguistics. http://www.mpi.nl/world/tg/CAVA/mt/MTandDB.html.

Recommendations for Natural Interactivity and Multimodal Annotation Schemes

Laila Dybkjær and Niels Ole Bernsen
NISLab, University of Southern Denmark
Campusvej 55, 5230 Odense M
laila@nis.sdu.dk nob@nis.sdu.dk

Abstract

Standards and guidelines for creating natural interactivity and multimodal (NIMM) annotation schemes are becoming vital factors in ensuring usability and re-usability of annotation schemes as well as of the tools supporting the use of annotation schemes. This paper presents and discusses recommendations for the creation, documentation, representation, evaluation, selection, and adaptation of NIMM annotation schemes.

Keywords

Guidelines, annotation schemes, natural interactivity, multimodality.

1. Introduction

The field of natural interactivity and multimodal (NIMM) annotation covers spoken interaction, gaze, facial expression, gesture, body posture, use of referenced objects and artefacts during communication, interpersonal (physical) distance, etc., and combinations of any of these. Annotation (or coding) schemes in the NIMM area have so far been fairly anarchistic with little standardisation except for sub-areas, such as speech transcription and facial expression. However, standards and guidelines for creating NIMM annotation schemes more generally are becoming vital factors in ensuring usability and re-usability not only of the annotation schemes themselves but also of the tools which support the use of the annotation schemes.

This paper presents and discusses recommendations for the development and evaluation of NIMM annotation schemes in terms of five points addressed in the following five sections: how to create NIMM coding schemes; how to document NIMM coding schemes; how to represent NIMM coding schemes and annotations in a computer-readable format; how to evaluate NIMM coding schemes; and how to locate, select, and adapt an appropriate existing coding scheme. The proposed recommendations are heavily based on work done in the ISLE (International Standards for Language Engineering) NIMM Working Group [7], cf. [5].

2. Coding Scheme Creation

A coding scheme is designed to enable corpus tagging of instances of a particular class of phenomena expressed in one or several modalities. Coding scheme creation involves, at least, conceptual/theoretical work, tag set creation, and coding scheme testing and evaluation. Coding scheme creation often serves a particular initial purpose but this does not exclude that, once created, the coding scheme could benefit other coders and many different coding purposes.

The following rules of thumb address conceptual/theoretical work and tag set creation. Testing and evaluation is discussed in Section 5. The coding scheme creator should at least consider the following points:

- What is/are the coding purpose(s), what will the annotations be used for, etc.
- Which modality/modalities should be marked up;
- Which phenomena are of interest.
- Is the identified class of phenomena sufficient for the purpose(s) for which it is intended.
- Is the class of phenomena kept as general as allowed by the coding purpose(s).
- Often but not always, the class of phenomena to be coded is based on a theory which claims closure for the class, such as, for instance, that the class of phenomena includes all possible, different human facial expressions. This theory needs testing and validation.
- Sometimes the coding scheme is merely intended to capture a subset of some larger class of phenomena for some purpose, such as when speech transcribers use a subset of a larger set of transcription tags. In such cases, there should be clear rules for how to add new phenomena to the coding scheme, should that be needed later, so that these will be consistent with the already existing ones.
- Each phenomenon must be clearly exemplified and described, so that both the coding scheme creator and others are always able to decide, given a certain token in a corpus, whether or not that token is an instance of that phenomenon. This point is crucial to inter-coder agreement on how to apply the coding scheme to a given corpus, cf. Section 5. Lack of clarity and coverage in the description of phenomena translates

into reduced inter-coder agreement, reduced consistency of codings, and quickly into a coding scheme which is too unreliable for practical use.

- Each phenomenon must be assigned a syntactic tag whose presence in the corpus, or whose reference to a particular token in the corpus, indicates its presence.
- The tag set representing the relevant class of phenomena should preferably be defined using some kind of standard format for coding tool use, e.g. XML. The tag set to be interpreted by machine does not need to have the same format as the tag set used by the human coder, one-to-one correspondence is sufficient (se also Section 4).
- The tag set should be extensible following well-defined rules.

The guidelines above are closely connected with coding scheme documentation and coding scheme formats, as discussed in Sections 3 and 4.

3. Coding Scheme Documentation

Experience shows that many coding schemes are poorly documented, which makes their retrieval and re-use very difficult. There is not yet any standards as regards which kind of documentation (meta-data) to include with a coding scheme. The MATE [8] and NITE [9] projects have proposed the concept of a coding module which extends the notion of a coding scheme with documentation that should be sufficient for colleagues to understand and use the coding scheme, cf. [3, 4]. At the same time, this documentation is structured in a way which makes it easy to search through if available on the web. The contents of a coding module is listed below:

- Name of coding module
  (E.g. my_gestures.)
- Author(s) of coding module
  (E.g. Tom Jones.)
- Version
  (E.g. v1.2.)
- Notes
  (References to literature, validation information, comments, etc.)
- Purpose of the coding module
  (Description of the purpose for which the coding module was first created.)
- Coding level(s) covered by the coding module
  (E.g. dialogue acts, hand gesture, nose wrinkles, …)
- Description of data source type(s) required for use of the coding module
  (E.g., an orthographic transcription may be a pre-condition for applying a particular coding scheme.)
- Explanation of references to other coding modules
  (If the coding module assumes that there are references to other levels of markup then these references should be explained.)
- Coding procedure
  (Description of how the coding module should be applied to a corpus in order to produce a reliable coding. The coding procedure is important to ensure the reliability of the coding and thus to its quality. The coding procedure should include, cf. [4]:
  - Description of the coders: their number, roles and required training.
  - The steps to be followed in the coding.
  - Intermediate results, such as temporary coding files.
  - Quality measures (the non-satisfaction of which may require re-coding).
- Coding example showing the coding scheme markup in use
  (This could be a snippet from an annotated file or a constructed example. The purpose is to give users of the coding module an idea of what the markup looks like when applied.)
- Clear description of each phenomenon, example(s) of each phenomenon
  (The descriptions provided are essential to a clear and sufficient explanation of how each concept-tag pair should be applied during markup. Any uncertainty left by the descriptions and examples provided will translate into unreliable coding, inter-coder disagreement, etc.)
- A markup declaration, possibly hierarchically ordered, of the tags for the (individually named) phenomena which can be marked up using the coding module
  (The tag set declaration can be presented in several different ways, e.g., as a DTD, cf. Section 4.)

It takes time to create and document good coding schemes but we believe it is worth the effort. Don't expect that anyone will be able to reliably use a "coding scheme" which only consists of, e.g., a tag set and a sparse description. You may have been able to use it yourself at creation time having it all in your head, but if you want to return to it just a few month later it will not be that easy even for you.

4. Coding Scheme Representation

This section addresses which formats to use for coding scheme representation. We need to distinguish between computer-readable formats and human-readable formats.

As for computer-readable formats, there is a strong trend today towards using XML. Coding scheme definitions are very often provided via an XML DTD (Document Type Definition) or via XML Schemas. We recommend to follow this de facto standard since XML, DTDs and Schemas are machine-readable, extensible, and widespread. Also for annotated data, XML is widely used. This means that using XML for this purpose as well will facilitate the exchange of annotated corpora and the use of tools based on XML corpus representation. It should be noted, however, that

XML is only syntax. Translation of tags into the set used by a specific tool may be needed in order to use that tool. Usually this is still much less work compared to translating a home-grown language into XML, e.g. the same parser tools can be used. For more information on XML, see, e.g., [10].
Whereas XML DTDs and Schemas are excellent for computers, they are less easy to read and write for humans. If tool support is available when one makes a markup declaration, it may be possible to use a format which is more friendly and easy for humans without special programming skills. Behind the user interface, the tool may then, e.g., convert the markup declaration into an XML DTD. To the user, however, the markup declaration may just be in terms of, e.g., well-defined form-filling. The special XML tags are then added by the tool behind the scene.
We recommend the development of tools which facilitate easy indication of markup declarations and support the use of an underlying standard representation format.

5. Coding Scheme Evaluation

Coding scheme evaluation follows coding scheme creation and documentation. The purpose of evaluation is to test the quality of the coding scheme and the results produced by using the coding scheme as intended. Precise and informative evaluation results provide very useful information to those looking for an existing coding scheme to use, cf. Section 6.
The coding scheme should be applied according to the prescriptions in the coding procedure, cf. Section 3. Thus, e.g., the annotators must have the background and expertise recommended and the number of annotators prescribed must be used to ensure the quality of the coding.
The ease-of-use and reliability of the coding scheme may be measured by:

- asking coders their opinion (interview, questionnaire);
- checking if different coders use tags consistently;
- measuring the time taken to code;
- measuring the quality of the annotations, cf. below.

Similarly, the ease-of-use of coding tools may be evaluated by asking coders their opinion and by measuring the time it takes them to code. Measuring the quality of codings is also relevant for tools evaluation if markup is done semi-automatically or automatically.
Coding scheme quality is a research area of its own. A coding scheme may be evaluated by:

- comparing different corpus samples coded by means of the scheme to assess *coverage;*
- comparing the results produced by different coders to assess *inter-coder reliability;*
- comparing the results produced by the same coder on the same corpus sample at different times, for instance with a one-week delay, to assess *consistency*.

Coding scheme quality may be evaluated:

- qualitatively through discussion of the choices made by coders when they differ;
- quantitatively through scoring measures.

A frequently used method to compare the results produced by different coders (inter-coder agreement) is called *kappa*:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

P(A) is the proportion of times that the coders agree and *P(E)* is the proportion of times that they are expected to agree by chance. A problem with this method is that there is no sound interpretation of which kappa values are good enough. Moreover, kappa presupposes independent events which is far from always the case in NIMM contexts, see also [2].
Two other measures, precision and recall, may be used if there is an 'authoritative source' to which the codings may be compared. *Precision* expresses the proportion of the occurrences found that have been correctly coded:

$$\text{precision} = \frac{\text{found - incorrect}}{\text{found}}$$

Found represents everything that was marked by the coder. *Incorrect* represents the incorrect markups made by the coder, as determined by the authority.
Recall expresses the proportion of occurrences that have been found:

$$\text{recall} = \frac{\text{all - missing}}{\text{all}}$$

All represents all occurrences present in the corpus, as determined by the authority, and *missing* represents those occurrences that were not identified by the coder [1].
We recommend that:

- any evaluation made of a coding scheme is referenced from the documentation of the coding scheme, so that it is easy to find;
- evaluation methods used and the evaluation process are clearly described;
- evaluation results are clearly documented.

6. Coding Scheme Selection And Adaptation

We have discussed recommendations related to the creation, documentation, and evaluation of coding schemes. However, it is of course much easier if there is already a well-documented and evaluated coding scheme available somewhere which fits the needs one may have. It is better still if this coding scheme comes with tools support.
No matter if one is going to create a coding scheme or select an already existing scheme, one should consider the issues listed in Section 2. Moreover, one should know who will be doing the coding, i.e. which level of expertise is available for this task.

When this is done, we recommend to look for an existing coding scheme which satisfies the identified constraints before a possible decision is made to create one's own coding scheme. Locating existing coding schemes is not necessarily easy to do for the moment since there are many sources which one may consult, including, e.g., survey reports, proceedings of conferences such as LREC, the ELRA/ELDA website [6], and free-style web search.
The checking of which coding schemes exist and what they are meant to be used for could be greatly facilitated if coding schemes are:

- well-documented, following the recommendations in Section 3;
- available on the web in the form of collections maintained at a small number of sites.

Documentation following the recommendations above (Section 3) would also greatly facilitate comparison of different coding schemes.
If one or several coding schemes are found which could be candidates for selection, we recommend to consider at least the following criteria before selection is made, and to weight the criteria according to their importance in the specific case:

- Coding scheme documentation.
- Coding scheme evaluation.
- Coding scheme extensibility, if applicable (Section 2).
- Coding scheme adaptability.

By *extensibility* we mean that new tags and their conceptual descriptions can easily be added. Extensibility becomes easier if the coding scheme includes a description of how this should be done. *Adaptation* of a coding scheme may include coding scheme extension but may also include other forms of changes to the original scheme, such as partial replacement of the tag set, a different coding procedure, or other/more coding files referenced. Whether adaptation - which is typically a larger operation than extending a scheme –is the right choice, depends at least on:

- how many changes are needed to make the coding scheme fit one's purpose;
- how easy it will be to make the adaptation; and
- what will be gained from making the adaptation compared to creating a new coding scheme.

Ease of adaptation depends on the coding scheme itself as well as on the available documentation.
The gain by making adaptation may range from not having to create an entirely new coding scheme and not having to do the coding scheme documentation from scratch, to getting access to tools support which may greatly facilitate the annotation and analysis process. If the gain is small, it may, in fact, pay off to create a new coding scheme instead, one which completely fits one's purposes. Available tools support, on the other hand, is a great advantage and may make adaptation the optimal choice.

7. Conclusion

(De facto) coding scheme standards mainly exist for speech and text annotation, especially in the area of transcription, and for media production-related issues. For other NIMM sub-areas, no real standards seem yet to exist. The standards which do exist have typically been brought forward by projects or international groups of people with a shared interest in some area, and sufficient need and momentum to get the consensus-building process started. Most existing standards are accompanied by supporting software, which makes them even more attractive to use since their use is facilitated by the software.
The recommendations for NIMM annotation scheme development and evaluation presented in this paper are based on best practice studies made in the European NIMM Working Group in the ISLE project. We hope that they may serve as a basis for further work in the NIMM annotation area, eventually leading to standardisation.

Acknowledgements

Much of the work reported above was carried out in the ISLE project. We gratefully acknowledge the support by the European Commission's HLT Programme. We would also like to thank Malene Knudsen, Joachim Llisterri, Maria Machuca, Jean-Claude Martin, Catherine Pelachaud, Montse Riera, and Peter Wittenburg for their contributions to ISLE report D9.2.

References

1. Bernsen, N.O., Dybkjær, H. and Dybkjær, L. *Designing Interactive Speech Systems. From First Ideas to User Testing.* London, Springer Verlag 1998.
2. Dybkjær, H. and Dybkjær, L. Measuring Transaction Success in Spoken Dialogue Information Systems. *Proc. of the Nordtalk Symposium on Relations between Utterances,* Copenhagen, 2002, 110-131.
3. Dybkjær, L., Bernsen, N.O., Carletta, J., Evert, S., Kolodnytsky, M. and O'Donnell, T. The NITE Markup Framework. *NITE Report* D2.2, 2002.
4. Dybkjær, L., Bernsen, N.O., Dybkjær, H., McKelvie, D. and Mengel, A. The MATE Markup Framework. *MATE Report* D1.2, 1998.
5. Dybkjær, L., Bernsen, N.O., Knudsen, M.W., Llisterri, J., Machuca, M., Martin, J.-C., Pelachaud, C., Riera, M. and Wittenburg, P. Guidelines for the Creation of NIMM Annotation Schemes. *ISLE Report* D9.2, 2003.
6. ELRA: http://www.elda.fr/
7. ISLE NIMM: isle.nis.sdu.dk
8. MATE: mate.nis.sdu.dk
9. NITE: nite.nis.sdu.dk
10. XML: www.w3.org/XML

A Multimodal Corpus Collection System for Mobile Applications

Knut Kvale, Jan Eikeset Knudsen, and John Rugelbak
Telenor R&D, Snarøyveien 30, N-1331 Fornebu, Norway
knut.kvale@telenor.com je-knuds@online.no john.rugelbak@telenor.com

Abstract

In this paper we describe a flexible and extendable corpus collection system for multimodal applications with composite speech and pen inputs, and composite audio and display outputs.

The corpus collection system can handle several pen clicks on the touch screen during an utterance and it can easily be extended to handle other modalities than speech and pen (e.g. gestures). The advantages of the corpus collection system are demonstrated with a scenario-based user experiment where non-expert users were asked to solve tasks in a tourist guide domain using our multimodal PDA-based application.

Keywords

Multimodal corpus, composite inputs, flexible design.

Introduction

In multimodal human-computer interfaces multiple input and output modalities can be combined in several different ways. This gives the users the opportunity of choosing the most natural interaction method depending on context and task.

Multimodal systems have the different parallel input channels active at the same time. We distinguish between sequential and composite multimodal inputs. In a *sequential* multimodal system only one of the input channels is interpreted at each dialogue stage (e.g. the first input). In a *composite* multimodal system all inputs received from the different input channels within a given time window are interpreted jointly [1].

Composite multimodal interaction is natural between humans, but it is by far one of the most complicated scenarios to implement for human-computer interaction.

For the purpose of investigating multimodal human-computer interaction, a test platform has been developed for speech-centric multimodal interaction with small mobile terminals, offering the possibility of composite pen and speech input and composite audio and display output. In the main parts of this work we cooperated with researchers at France Télécom, Portugal Telecom, Max Planck Institute for Psycholinguistics, and the University of Nijmegen in the EURESCOM-project MUST – "Multimodal and Multilingual Services for Small Mobile Terminals" [2,3,4].

This paper focuses on the multimodal corpus collection system of the test platform. In the following sections we first describe the platform architecture. Then we elaborate on the multimodal corpus collection system. Finally a sample corpus from a user experiment is discussed.

System Overview And Architecture

The multimodal test platform

Our test platform consists of a server and a thin client (i.e. the Mobile Terminal) as shown in figure 1.

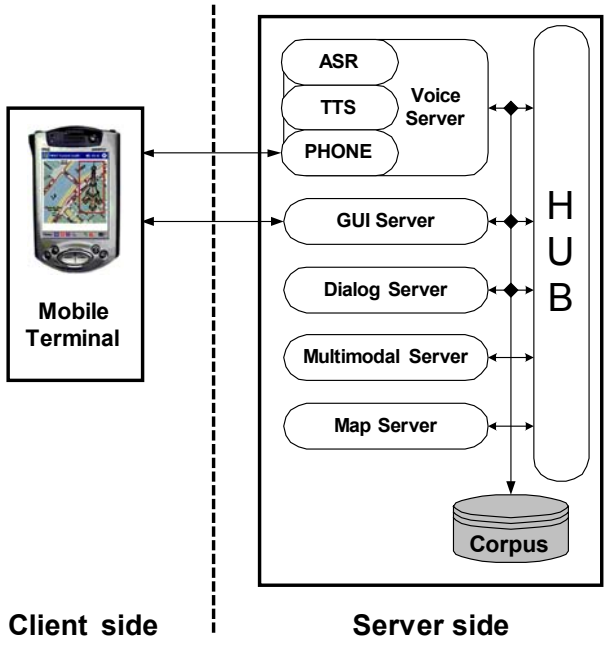


Figure 1: The overall architecture of the test platform.

The Server side comprises five main autonomous modules which inter-communicate via a central facilitator module (HUB). The modules are:

Voice Server – comprises Automatic Speech Recognition (ASR), Text to Speech Synthesis (TTS) and Telephony Server (PHONE) for the speech/audio modalities.

GUI Server – handles the graphical user interface (GUI) signals between the terminal (display) and the server side for the pen/display modalities.

Dialog Server - performs the dialog/context management.

Multimodal Server - performs multimodal integration of the incoming signals (fusion), and distributes the response through the output channels (fission).

Map Server - acts as a proxy interface to the map database

HUB – manages the inter-communication for the modules.

The requests from the user are represented in terms of textual or abstract symbols by the boundary modules, i.e. the Voice- and GUI Server that handle the interaction with the user. The Dialog Server combines and processes the inputs (late fusion), and acts accordingly to fulfil the users request (typically a query to a database). The feedback is sent to the boundary modules via the Multimodal Server in order to be presented in different modalities on the Mobile Terminal (early fission).

The Norwegian version of the multimodal test platform is based on the Telenor R&D voice platform [4]. The Automatic Speech Recognition is based on Philips SpeechPearl® 2000 for Norwegian with a fixed 65 word open grammar covering 10 concepts. For Norwegian Text-to-Speech Synthesis we use Telenor R&D's Talsmann®.
The Client side is implemented on a PDA with audio and touch screen. For the experiments reported here we applied a Compaq iPAQ Pocket PC running Microsoft CE 3.0/2002. The PDA is communicating with the Server side via WLAN in order to obtain mobility for the terminal. More technical details of the multimodal platform are provided in [2,3,4,5,7].

The applications

We have implemented two different map applications: "Tourist guide to Paris" [2,3,4,6], and "Bus travel information for Oslo" [8]. These map-based applications require use of both pen and speech actions to accomplish the tasks, but the users are free to interact either *sequentially*, i.e. to tap with the pen first and then talk, or *simultaneously*, defined as a pen action in the time window from e.g. one second before start of speech to one second after end of speech (called composite inputs).
These multimodal map-applications are *fully user driven*. Thus, the system must always be in the ready state of obeying and serving the user, i.e. receiving queries from the user at any time and in any dialog state, and respond accordingly. This complicates the multimodal dialogue control and management.

The user interface

For the "Tourist guide to Paris" application the graphical part of the user interface consists of two different types of maps: An overview map for Paris showing all Points Of Interest (POI), such as the Eiffel Tower, Notre Dame and Hotel de Ville, and detailed maps with the respective POI in the center and optionally with facilities such as restaurants, metro stations or hotels around the POI. Figure 2 shows the PDA screen-layout with the detailed map for the Eiffel Tower.

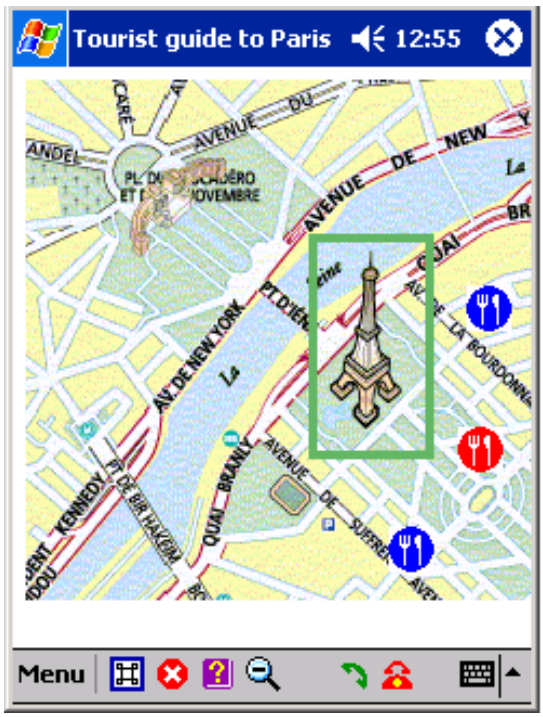


Figure 2: The PDA-screen layout of the "Tourist guide to Paris" showing the detailed map for the Eiffel Tower with nearby restaurants.

The Multimodal Corpus Design

The design of a multimodal corpus, i.e. the content and data structure of the corpus, depends on the application and the aim of the user experiment analysis. Our intention was to analyze and evaluate multimodal man-machine dialogues with small mobile terminals. We were interested in finding to what extend users really combined the different modalities (sequential or composite inputs). To do this we defined metrics as timing, user response time and success rate (time and number of turns to complete a task).

Corpus data and parameters

The main parameters in the multimodal corpus data set are listed in table 1. All the parameters in this table have a timestamp attribute. The time resolution is parameter dependent. For the most time critical parameters such as input voice utterances and pen clicks the resolution is 50–100 ms. This time resolution is needed for evaluating the coordination of the composite speech and pen inputs, and user response times in general.

In the corpus a dialog turn is defined as one user input action and the corresponding system output.

Parameter	Description/Attributes
Header information	Administrative information about the user experiments such as host laboratory, signature and information about the user (e.g. age, gender etc).
Audio input	The audio (speech) input to the system during the whole dialog session is recorded to an audio file.
Audio output	The audio output to the user during the whole dialog session is recorded to an audio file.
Input speech	The input speech utterances that are forwarded to the ASR engine are also recorded to audio files.
ASR symbols	The recognized textual or abstract symbols from the ASR engine. Information about the grammar. Technical information about the ASR engine.
Text prompts	The text that is synthesized and played. Technical information about the TTS engine.
Audio prompts	The pre-recorded audio files played to the user. Type of audio such as voice, music and sound effects.
Input pen	Data field(s) associated with the input pen clicks from the terminal, such as screen coordinates and name of the clickable object (i.e. icon).
Output display	The XML/HTML files representing the GUI display. Graphical type (text, forms, icons, images etc).
Dialog state	The current dialog state.

Table 1: The multimodal corpus parameters with attributes.

Directory and file structure

The directory structure of the multimodal corpus is shown in figure 3. Only the Dialog-, GUI- and Voice Server modules store data to the corpus, and the respective corpus files are stored to each module's corpus directories. A sub-directory is created for each dialog session, and the name of the sub-directory is the timestamp at the beginning of the session (i.e. the session ID), and the format is: YYYY-MM-DD_hh_mm_ss.

The Dialogue Server creates a *Main Corpus File* (*main_corpus_file.xml*) for each dialog session. This file contains information about all parameters listed in table 1. The format of the Main Corpus File is XML, and a Document Type Definition (DTD) validates the format. XML eases the process of retrieving, inspecting and processing the data in the corpus. Below is a sample portion of the Main Corpus File:

```
<Turn number="1">
 <UserInput dialogstate="HOME">
  <Pen>
    <Hotspot type="POI" category="church" name="Notre Dame"/>
    <Timestamp>2002_06_18_13_42_17__892</Timestamp>
  </Pen>
 </UserInput>
<SystemOutput dialogstate="POI">
 <Graphical>
  <XMLFilename>
    ./GuiServer/Corpus/2002_06_18_13_42_04/gui_display_2.xml
  </XMLFilename>
  <HTMLFilename>
    ./GuiServer/Corpus/2002_06_18_13_42_04/gui_display_2.html
  </HTMLFilename>
  <Timestamp>2002_06_18_13_42_18__489</Timestamp>
 </Graphical>
</SystemOutput>
</Turn>
```

In this case the user taps on a POI (here: "Notre Dame") on the overview map, and traverses to the corresponding detailed map represented by the content of the files *gui_display_2.xml* and *gui_display_2.html.*

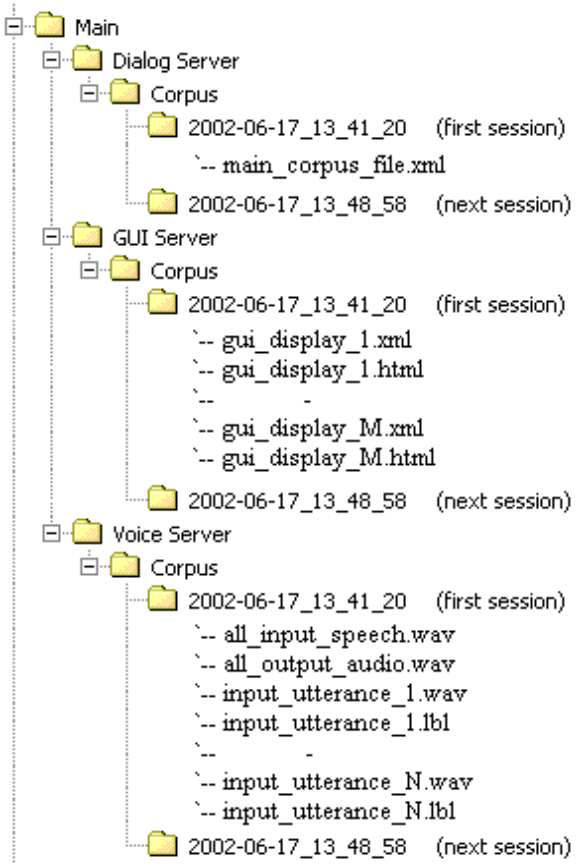


Figure 3: Directory and file structure of the multimodal corpus.

An XML- and an HTML-body define the graphics displayed at the Client side. The GUI Server stores these bodies in the corpus for each displayed image to an XML file (*gui_display_*.xml*) and an HTML file (*gui_display_*.html*) respectively.

The Voice Server records all input and output speech to audio files (**.wav*) in Microsoft WAV-format (A-law, 8 kHz, mono). The Voice Server creates a *Label File* (*input_utterance_*.lbl*) for each input speech utterance. The corresponding recognized symbols from the ASR engine (i.e. words and concepts with confidence scores and timestamps) are stored in the Label File. The format is XML and complies with the DTD for the Main Corpus File.

A Sample Corpus From A User Experiment

Our test platform has been applied in a scenario-based user experiment where non-expert users were asked to solve different tasks in a tourist guide domain [6]. Since the test subjects were unfamiliar with using multimodal inputs, we first had to explain the functionality. The main aim of the experiment was to

investigate whether users' interaction style (sequential versus composite pen and speech input) depended on the *format* of the introduction to the system. We also studied *learning effects*, i.e. whether the users' interaction style changed over time, and *timing issues* such as whether tapping tended to be near the start of utterances, near the end of utterances, or near deictic words.
In this section we briefly discuss the Norwegian part of the corpus with respect to the flexibility of the collection system and the possibility of reusing the corpus for further research on multimodal interaction.

The corpus

The 21 test users were divided into three groups that got the same introduction to the system. Parts of the introduction were presented to the groups on different formats (one text version and two different videos). Each subject was presented to 3 scenarios. All scenarios had exact the same structure, and the users had to solve 6 tasks during each scenario. To complete all tasks both pen and speech inputs were required, but the users were free to choose either sequential or composite pen and speech input at each step in the dialogue. The corpus for this experiment consists of 507 pen taps and 758 speech utterances.

Using the corpus for analysis

Based on the corpus parameters and attributes (e.g. timestamps) listed in table 1 we may calculate different metrics, such as the number of dialog turns for solving a task or to complete a scenario, utterance length, and overall task completion time. The corpus can be used to investigate the multimodal interaction patterns in different contexts and tasks, e.g. how users apply pen inputs nearby spoken deictic words.

Conclusions And Further Work

We have described a flexible multimodal corpus collection system, and shown how it can be used for studying multimodal interaction. The corpus contains timestamps for all system outputs and several input events. New hypotheses can be tested on the corpus by defining new thresholds and metrics.
The flexibility of our corpus system gives several benefits:

- The platform can easily be adapted to new applications and has been extended to allow two taps within one utterance, e.g. "when does the next bus go from here <tap 1> to here <tap 2>. The corpus collection system handles this too [8].
- The corpus collection system can easily be extended to handle other modalities.
- The corpus collection system is well designed for annotation.
- The corpus collection system is well designed for the reconstruction of the dialog session e.g. by means of an XML processor and a media player.

For future work we plan to develop an analysis tool that comprise an annotation- and a reconstruction-module. This analysis tool may ease the investigation of multimodal interaction patterns in different contexts and tasks.

Acknowledgments

We would like to thank our colleagues in the MUST project and in the Speech Technology Group at Telenor R&D for valuable and fruitful discussions and cooperation.
This work has been financed by Telenor R&D, EURESCOM and the BRAGE-project of the research program "Knowledge development for Norwegian language technology" (KUNSTI) of the Norwegian Research Council.

References

1. World Wide Web Consortium (W3C) Multimodal Interaction Requirements. Available at http://www.w3.org/TR/mmi-reqs/
2. Almeida L, et al., "User friendly multimodal services,
   - A MUST for UMTS". In: Proc. EURESCOM summit 2002, Heidelberg, Germany, Oct 2002.
3. Almeida, L. et al. "Implementing and Evaluating a Multimodal Tourist Guide", Proc. International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems, pp.1-7, Copenhagen, Denmark, 2002.
4. Almeida, L., et.al.: "The MUST guide to Paris - Implementation and expert evaluation of a multimodal tourist guide to Paris", Proc. ISCA Tutorial and Research Workshop (ITRW) on Multi-Modal Dialogue in Mobile Environments, (IDS 2002), pp. 49-51, Kloster Irsee, Germany, 2002.
5. Knudsen, J.E., Johansen, F.T. and Rugelbak, J., "Tabulib 1.4 Reference Manual", Telenor R&D N 36/2000, 2000.
6. Kvale, K., Rugelbak, J., Amdal, I.: "How do non-expert users exploit simultaneous inputs in multimodal interaction?", In: Proc. International Symposium on Human Factors in Telecommunication, pp. 169-176, Berlin, 1.-4. December 2003.
7. Kvale, K., Warakagoda, N.D. and Knudsen, J.E., "Speech centric multimodal interfaces for mobile communication systems", in Telektronikk, 2.2003, pp. 104-117.
8. Lium, A.S., "A speech-centric Multimodal Application for Bus Traffic based on a Handheld, Mobile terminal", Master thesis, the Norwegian University of Science and Technology, spring 2003 (in Norwegian).

Evaluation of multimodal dialog systems

Louis Vuurpijl[1], Louis ten Bosch[1], Stéphane Rossignol[1], Andre Neumann[1], Norbert Pfleger[2], Ralf Engel[2]

[1]NICI, The Netherlands {vuurpijl, rossignol}@nici.kun.nl, l.tenbosch@let.kun.nl
[2]DFKI, Germany {pfleger,rengel}@dfki.de

Abstract

This paper presents the results of an elaborate study on pen and speech-based multimodal interaction systems. The performance of the "COMIC" system is assessed through human factors analyses and evaluation of the acquired multimodal data. The latter requires tools that are able to monitor user input, system feedback, and performance of the multimodal system components. Such tools can bridge the gap between observational data and the complex process of the design and evaluation of multimodal systems. The evaluation tool presented here is validated in a human factors study on the usability of COMIC for design applications and can be used for semi-automatic transcription of multimodal data.

Keywords

Multimodal system design and evaluation, multimodal corpora, human factors.

Introduction

Experience has shown that the design and evaluation of multimodal interactive systems poses a complex, multi-disciplinary problem [1,2]. In large projects such as SmartKom [3] or COMIC [4], it requires a collaboration between researchers from psychology and cognitive science, up to computer science and artificial intelligence. On the one hand, the study of human subjects interacting with the system yields tons of data that can now be explored by means of "traditional" annotation and transcription tools. On the other hand, these data reveal no details about the performance of individual or mutually communicating system components on the basis of particularities in the multimodal inputs. One could state that the main problem is caused by the gap between annotating data acquired through human factors studies and using these data in the process of system design and evaluation. This paper reports on our findings in this matter in the context of the design and evaluation of the COMIC multimodal system for bathroom design.

In bathroom design, (non-expert) customers have to provide the salesperson with shape, dimensions and additional features of a bathroom. Recordings of dialogs between salespersons and customers have shown that these dialogs are inherently multimodal. In the IST project COMIC (www.hcrc.ed.ac.uk/comic/), we are developing a system that supports non-expert users with specifying the bathroom of their desire, in a way that approximates natural human-human interaction and dialog. To build such a system, and to be able to advance our understanding of the issues involved in interaction with such a system, we need to explore how people enter data about a bathroom with pen and speech as input channels [6]. In this paper we report on an experiment that was aimed at investigating the performance of individual components of the COMIC system. To that end we performed a usability study in which naive subjects interacted with the system, and in doing so, generated a large amount of data that can be used to measure the performance of the individual system components.

Previous research (e.g. [4, 6]) has shown that it is very difficult to make sense of the data recorded in multimodal interaction systems. Even if, as is the case in the present experiment, the interaction strategy is designed to constrain the user actions, multimodal interaction appears to offer many alternative ways to approach the goal. This large degree of freedom is especially important in the analysis of interactions with naive subjects, who lack the telepathic knowledge of the system's expectations that the system designers do have, and that helps tremendously in finding the most efficient interaction strategy and to avoid situations in which the system may not be robust. In addition, objective data (the input and output of the individual modules in a system, including time stamps attached to actions of the system and the user) form a kind of cascade. In order to analyze the performance of individual modules, for each module its complete set of input and output messages must be considered. For speech and pen input this involves manual annotation of the physical input signals. Speech input must be transcribed verbatim, as well as in the form of the concept values expressed by the words. For pen input {x,y,z} coordinate streams must be annotated with the semantic labels that are relevant in the specific application. To assess the performance of modules that have no direct relations with physical input or output, such as FUSION, which receives symbolic input of the speech and pen input processors and passes symbolic data to the dialog action manager (DAM), {input,output} pairs must also be annotated for correctness (or type of error). In the past, the development of multimodal systems has been hindered by the absence of suitable tools for annotating and analyzing interaction data. A tool for the analysis of these interaction data would greatly facilitate the evaluation of the entire dialogue system. It is the aim of this paper to present the tool that we developed to support experiments with the COMIC system.

The COMIC System

The eventual COMIC system will comprise decoders for speech (ASR) and pen input (PII), a FUSION module that merges pen and speech input, a dialog and action manager (DAM), a Fission module that decides what information must be rendered in the form of speech, text or graphics, and output modules that generate the actual output, including an avatar with an advanced facial expression generator. The provisional system used in the experiment described in this paper had full-fledged input and fusion modules, a rudimentary DAM and simple, fixed procedures for output generation and rendering. The user interacts with the system via a head mounted close-talk microphone and a Wacom Cintiq 15X LCD tablet that acts as a paper-and-pen metaphor. COMIC employs the MULTI-PLATFORM communication architecture (MP), which is developed by DFKI, one of the partners in COMIC [5]. All data communicated between modules are encoded in XML and logged. These data provide a means for system debugging and tuning and typically are not considered when transcribing video, audio, or pen data. In the remainder of this paper, we describe our approach for combining both types of data: observational recordings and system loggings. A specification of typical multimodal system loggings and the evaluation tool "μeval" are discussed. Subsequently, we present the results from the human factors experiments that were obtained by using the new tool. We will show that μeval provides a means for semi-automatic annotation of the acquired observational data, while providing statistics on the system performance based on annotated system logs.

General structure of multimodal system loggings

Most communication platforms like Galaxy, the Open Agent Architecture and MP provide means to log system messages. Given the multi-modular nature of multimodal systems, and because modules are typically developed by different persons, system logs can end up in a mess of messages that are only interpretable by the producer. Logs of inter-module messages nowadays are mostly encoded in XML. Messages are structured in a header, containing the source of the message, a message identifier, and timing information. The latter is extremely important and time should be synchronized over all modules. The contents of the body of a message is defined by the developers of the module that writes the message and must be parsed by all modules that read it. Loggings can become extremely large, making it very difficult to investigate failures in the communication protocols by hand. Today, no tools exist that support module developers who use MP as the integration platform in the process of debugging the distributed system messages.

The tool we developed contains knowledge about the message content and is able to parse messages produced by all current COMIC modules. It is designed such that it can monitor any message log that contains:

```
header: <timestamp> <id> <source>
body: any xml-encoded string sequence
```

For example, if a user interacting with the system would draw a wall and speak out its length, the following message sequence would be recorded:

```
<msg>t0 id0 pen-tablet
  some-sequence-of-coordinates</msg>
<msg>t1 id1 microphone
  some-audio-input</msg>
<msg>t2 id2 PII
  some-wall-encoding</msg>
<msg>t3 id3 ASR
  some-lattice-containing-length</msg>
<msg>t4 id4 FUSION
  some-wall-with-length-encoding</msg>
<msg>t5 id5 DAM
  some-rendering-and-next-state</msg>
```

In this example, it is assumed that all input data are communicated, including audio signals. In most cases however, audio and video signals do not pass through communication channels in order to reduce bandwidth. This is also the case in COMIC, where the ASR system is directly coupled to a microphone and stores audio fragments on disk. Pen coordinates are communicated and are thus contained in the multimodal system logs.

Fast semi-automated annotation of MM interaction

When annotating multimodal interaction dialogs, the annotation process in general takes at least as long as the interaction itself. By using μeval, this process can be sped up considerably, while recording performance statistics for the individual modules. The tool considers header information present in the system logs, and sorts messages by their source and timestamp. So, messages from all PII, ASR, and other sources can easily be identified and categorized. For each message, messages from other sources that temporally correspond to it, can be detected. User input can be monitored by depicting pen input coordinates and playing audio inputs stored on disk. The latter is possible when ASR messages are marked up with the filename of the corresponding audio fragment. Now, during the processing of the recorded loggings by μeval, for each sequence of messages, the user input is rendered and the corresponding output of each module is presented in a manner that is easily readable and interpretable for a human evaluator.. The evaluator of the interaction turns can judge each output in terms of categories, such as 'ok', 'false', 'rejected by the module', 'rejected by the user', as 'noise', or as 'out-of-grammar' or 'ignore'. All correct interpretations labeled 'ok' can directly be used as the label of the unknown user input, and require no further involvement of the evaluator. All other classes of input can be stored for later processing or can be transcribed manually. We have used μeval effectively for evaluating data while human factor experiments were ongoing. It appeared that the evaluation of each experiment took about 15 minutes, whereas the original interaction took on the average 60 minutes. The next sections describe the experiments and the results obtained through μeval.

Dialog design and turn taking

Since no comprehensive taxonomy of possible speech and pen repertoires in the bathroom domain are available, it was decided to design a fully system-driven dialog. A system-driven design narrows down the set of expected user dialog acts and avoids large numbers of out-of-domain or out-of-dialog speech and pen gestures. To that end, a synchronous turn-taking protocol was developed, in which (i) the system prompts the user for information (using canned speech); (ii) the user is allowed a certain time window to enter the requested information; (iii) the input decoders process the entered information, (iv-a) the interpreted information is *beautified* or (iv-b) rejected in case the decoders cannot recognize the input. Beautification, i.e. rendering sketches in the form of straight lines and fixed patterns, or rendering measures in ascii text, is the major way the system uses to show its interpretation of the user input.

If the input can be interpreted, beautification is followed immediately by the prompt for the next information item. If the input cannot be interpreted, a more elaborate prompt is played for the previous information element.

After any system prompt, two situations can occur. If the user is satisfied with the recognition result, he can reply to the next prompt, thereby implicitly confirming the interpretation. Alternatively, the user can explicitly reject this system interpretation, either by pen or speech. One compound turn in the dialog starts with an audio prompt generated by the system, followed by a reply or reject from the user, and terminated by the interpretation (and beautification) of the system. Theoretically, all confirmed system interpretations can be used as transcription of the input [2], but in actual practice subjects accept wrong recognition results when repeated attempts to correct errors are not successful.

Experimental design

The experiment consists of a free and a system-driven phase. In the free phase, subjects are requested to draw three bathrooms from memory, e.g., their parents', their own, and from a friend. No automatic recognition is involved. This condition serves two aims. First, natural, unconstrained, dialog acts provide essential material to further develop the various modules in the COMIC system. Second, the subjects get acquainted with the task: drawing on a tablet while using speech.

Next, they have to copy the same data into a computer system, using the tablet to sketch and write, and using speech to support their graphical input. Now, the computer does try to recognize all input gestures and utterances, using a system driven interaction strategy. Subjects are first instructed (by instructions on paper and by a video) about the automatic system. After entering the data for the three bathrooms, subjects are requested to fill in a questionnaire. In total, 28 native speaking German subjects participated with varying computer experience.

Data collection and labeling using μeval

All logged data have been processed using our evaluation tool. For each system prompt, the expected class of user response is known (i.e. wall, window, door, or some measure). For each individual module, a label was assigned by the human evaluator to indicate the correctness of the module output ('ok','false','noise', 'oog'='out of grammar'). Rejects or confirmations by the user or by the system were also labeled accordingly.

All data that were interpreted by a decoder and were labeled as "ok" by the evaluator can be considered as a candidate for automatic transcription. Depending on the recognition performance of the decoding systems, this can speed up the transcription process considerably, as both segmentation and labeling are performed automatically.

Cases where the system is unable to handle the input correctly are of special interest for improvements. Also data that are rejected by the recognizer, e.g., because the user draws an unknown shape, or in cases where the user employs out-of-context speech, are interesting. For speech, these data are used to refine the language model and to tune acoustic garbage models. For pen input, these cases form examples that require new pattern recognition algorithms. Evaluators from different labs (DFKI, NICI) have used μeval for labeling and debugging purposes. It has proven to speed up both processes considerably.

Evaluation of multimodal input

The results presented here are based on the information generated through μeval. Using the information available in the header of logged messages, the difference between two subsequent semantic expectations (broadcast by the DAM) is defined by the total turn time. Average turn time was computed for 4 input concepts and for each of the three entered bathrooms (n=28). For each concept, the average time per turn (tt), the time for recording pen inputs (tp) and speech inputs (ts) is given below. No significant decrease in turn time was observed, which indicates that subjects quickly understood the task and that the instructions they received are sufficient.

	Bathroom1			Bathroom2			Bathroom3		
	tt	tp	ts	tt	tp	ts	tt	tp	ts
wall	11.4	4.1	2.9	11.0	3.6	3.0	10.5	3.6	2.8
door	13.3	3.4	3.0	11.9	3.5	3.0	12.0	3.6	3.0
window	11.8	3.7	3.3	11.4	3.9	3.0	10.6	3.5	2.9
size	12.2	3.9	3.4	11.8	4.1	3.2	11.8	3.9	3.2
all	12.3	3.6	3.1	11.8	3.6	3.1	11.5	3.5	3.3

When considering recognition results per input category, the tables depicted below indicate whether users improve their pen and speech input over time. Since the semantic interpretation of ASR output depends on the entire recognized sentence, string error rates rather than word errors rates are reported ("zwei meter zehn" recognized as "zwei meter achtzehn" counts as one error). For sizes interpreted by PII, also string error rates (e.g., "7.13 m" incorrectly recognized as "7.18 m") are reported.

Bathroom I	PII				ASR			
	n	ok	fa	r	n	ok	fa	r
WALL	119	117	0	2	41	19	3	19
DOOR	68	47	7	14	45	16	11	18
WINDOW	34	28	0	6	40	22	7	11
SIZE	190	123	61	6	201	66	81	54

Bathroom II	PII				ASR			
	n	ok	fa	r	n	ok	fa	r
WALL	117	114	0	3	37	25	5	7
DOOR	50	40	0	10	33	18	6	9
WINDOW	39	34	0	5	34	22	2	10
SIZE	216	139	72	5	219	84	105	30

Bathroom III	PII				ASR			
	n	ok	fa	r	n	ok	fa	r
WALL	116	116	0	0	52	30	5	17
DOOR	61	46	4	11	28	18	8	2
WINDOW	39	34	0	5	28	20	2	6
SIZE	198	149	48	1	235	89	109	37

Each row (four numbers) corresponds to respectively the total number of inputs (n), the number of correctly recognized input fragments (ok), the number of errors (fa) and the remaining (r) classes of input (rejects, noise, oog). Recognition performance for pen input interpretation is quite well in case of the recognition of drawings. The few errors represent rather complex drawings that PII was not designed for. For sizes, it is noticeable that the performance of ASR increases in the second trial but decreases for the third bathroom. (Main factors constraining the performance of the ASR are the one-line use of the ASR, the quality of the automatic end-of-speech detection, and the used language model). Also note that there is a correspondence between the number of errors and the total number of turns. For each recognition result that is rejected by the user, the system re-phrases the question and another turn is recorded, hence the different number of inputs (n) in the tables.

Monitoring user replies after errors

Subjects showed a variety of attitudes after an incorrect system interpretation in the speech modality. In the beginning of a test, most subjects are inclined to just repeat the utterance or repeat it slower. Rephrasing is not often used. Over sessions, the tendency to switch to the pen modality after an ASR error increases. Using the annotated system logs, such user behavior related to system responses can be monitored efficiently as below:

```
msgid expectation PII ASR FUS DAM USR

00834 WALL_LENGTH  -   f   o  drei R
00835 WALL_LENGTH  -   f   o  zwei R
00836 WALL_LENGTH  -   f   o  zwei R
00837 WALL_LENGTH  o   -   o  3 m  F
```

In this example, the user said “Drei meter” and rejected the output of ASR three times. FUSION made no errors in passing on the interpreted inputs and only after the third try, the user switched to the pen modality, which was judged as “ok” by the evaluator, corresponding to the confirmation “F” (fixed) by the user.

Discussion and conclusions

This paper discusses the possibility of combining the tasks of data transcription and system evaluation in one process. The approach presented here was used in a real human factors evaluation of the multimodal interaction system COMIC. Significant amounts of multimodal interaction data have been processed using the newly developed tool μeval. Although the tool can use many improvements, it has been validated and used effectively for system evaluation and debugging purposes. All module developers involved in input decoding (PII, ASR and FUSION) were able to browse and debug their loggings in a much more efficient way.

To our knowledge, the approach of transcribing multimodal data while annotating the corresponding session logs, has not been reported before in the literature. This approach opens up possibilities for fast transcription of observational data.

We have demonstrated that μeval is a flexible tool for evaluating dialogue turns in a complex human-system interaction, based on observational data and system log files. Although μeval is developed within the particular context of the COMIC bathroom design application and thereby implicitly makes use of the structure of the dialogue, it is basically a general-purpose tool that enables the evaluator to flexibly annotate {input, output} pairs of dialogue turns coded in XML-coded messages.

Acknowledgements

This work is sponsored by the COMIC IST-2001-32311 project.

References

1. Oviatt, S.L., Cohen, P. R., Wu, L., et al, Designing the U-I for Multimodal Speech and Pen-based Gesture Applications: State-of-the-Art Systems and Future Research Directions in *HCI in the new millennium*, pp 419-456, 2000
2. Potamianos, A., Kuo, H., Pargellis, A., et al, Design principles and tools for multimodal dialog systems, in *Proc. ESCA Workshop, IDS-99*, pp 22-24, 1999
3. Wahlster, W., Reithinger, N., Blocher, A. Smartkom: Multimodal Communication with a life-like Character, *Eurospeech*, Aalborg, Denmark, 2001
4. den Os, E.A and Boves, L. Towards Ambient Intelligence: Multimodal computers that understand our intentions*, Proc. eChallenges*, Bologna, 22 – 24, 2003.
5. Herzog, G., H. Kirchmann, Poller, P. et al. MULTIPLATFORM Testbed: An Integration Platform for Multimodal Dialog Systems, *Proc. HLT-NAACL'03,* Edmonton, Canada, 2003.
6. Rossignol, S., ten Bosch, L., Vuurpijl, L., et al, Human-Factors issues in multi-modal interaction in complex design tasks. HCI International, Greece, pp 79-80, June 2003.

Measuring Relative Target User Group Success in Spoken Conversation for Edutainment

Niels Ole Bernsen
Natural Interactive Systems Lab
Campusvej 55, DK 5230 Odense M Denmark
+45 6550 3544, nob@nis.sdu.dk

Abstract

The paper presents corpus data obtained from a relatively large field test of a Wizard of Oz (WoZ)-simulated specification of a multimodal domain-oriented spoken conversation system for edutainment. As the system design targets 10-18 years old users, a metrics is proposed for measuring the extent to which the simulated system specifically manages to appeal to its target user group. The metrics are applied to the WoZ corpus data, focusing on how to handle the observed differences between native and non-native English speaking users. This leads to a derived metrics which seems useful for system development progress evaluation.

Keywords

Evaluation metrics for edutainment systems, animated agent systems evaluation, multimodal spoken dialogue systems.

Introduction

This paper presents corpus-based results on the extent to which we have reached our target user group in a system aimed to have edutaining conversation with primarily 10-18 years old users. The system enables spoken English domain oriented conversation between users and life-like embodied fairytale author Hans Christian Andersen (HCA) and is being developed in the EU NICE project on Natural Interactive Communication for Edutainment [2]. Based on the design specification of the first system prototype, a Wizard of Oz (WoZ) simulation was carried out in the summer of 2003 at the HCA Museum in his native city, Odense, Denmark. During 10 days, approx. 500 conversations were recorded yielding 30 hours of spoken conversation data. This data has been transcribed and transcription coded. Each conversation has been evaluated with respect to the English language proficiency of the user. Topic-tagging of the corpus is in progress in order to identify all conversational topics addressed in the corpus.

By contrast with task-oriented spoken dialogue systems, whether unimodal (speech-only) or multimodal, domain-oriented systems do not help the user accomplish any particular task(s). Rather, the user can talk to the system spontaneously about anything, in any order, within the system's knowledge domain(s). Such systems raise novel issues of corpus-based evaluation, in particular, perhaps, if they have entertainment as one of their primary goals. For instance, classical dialogue efficiency metrics are probably irrelevant to their evaluation [1,3]. Rather, issues such as entertainment and edutainment success move to the forefront.

In the NICE HCA system, the target user group is 10-18 years old kids and adolescents. It is important to be able to evaluate the extent to which the target users are actually being entertained by the system, both in absolute terms and relative to non-target system users. This paper addresses the latter, relative, evaluation issue.

In the following, we briefly describe the NICE HCA system specification and the WoZ simulation. We then propose a metrics for measuring target group success in conversational systems for edutainment, discuss how to apply the metrics when the large majority of the users are non-native English speakers, present the results of applying the metrics, and discuss how to use the metrics for progress evaluation during continued system development.

NICE HCA System Specification

The system specification which was WoZ-simulated provides HCA with six domains of conversation: the childhood part of his life, the fairytale part of his work, his personality and visible physical presence in his study, gathering knowledge about the user, and his role as "gatekeeper" for access to the fairytale world in which users can interact with some of his fairytale characters. In addition, HCA has the "meta" domain of handling meta-communication caused by, e.g., user repeat requests or low input confidence scores. The following system aspects were not simulated: (i) the details of the system's error-handling meta-communication had not been specified at the time. In general, realistic system error behaviour as well as user and system error handling behaviour tend to be difficult to simulate using WoZ [1]; (ii) due to limitations of the graphical animation platform at the time, it was not possible to simulate the 2D user gesture input and its processing which form part of the (now implemented and running) first NICE HCA prototype; (iii) for the same reason, the simulation did not include HCA's conversational listening behaviour which, in the first prototype, enables HCA to show real-time attention to the user's spoken and gesture input. Finally, (iv) the details on exactly when HCA would exhibit emotional behaviour had not been designed at the time.

Wizard Of Oz Simulation Details

In technical Wizard of Oz terms, the simulation may be described as a full, field, close-to-complete specification, messy-experiment WoZ. A *full WoZ simulation* is one which does not include any implemented system components. A *field WoZ simulation* is conducted in the field rather than in a controlled laboratory setting. Users simply walk up and use the system with little or no introduction to its purpose or capabilities, and no requirements on the users whatsoever. A *close-to-complete specification WoZ simulation* is based on the system specification rather than on a more or less loosely defined purpose of gathering interesting data for a system which still has to be specified or which is under specification. The non-simulated specification aspects were described above. Finally, a *messy-experiment WoZ simulation* is one in which interaction experimentation is being carried out under less than strict textbook experimental conditions. Thus, in the simulations reported, the wizards were instructed to make, at their discretion, particular kinds of conversational improvisations which went beyond the system specification. These improvisations served as "messy" experiments intended to elicit user behaviours in addition to those which would be elicited by an uncompromising adherence to the system specification. For example, the wizards could talk out-of-specification in order to query the users about technical inventions made after HCA's times.

Figure 1. HCA addressing the user.

In the Museum, HCA was installed on a laptop which was wirelessly connected to the wizard working in the basement. A student had the task to round up kids and adolescents, inviting them to talk to "a nice person". In addition, a small poster in Danish and English invited the 10-18 year olds and other visitors to talk to this person, describing the system as a spoken computer game and informing users that their conversations would be recorded for research purposes. The user just had to don the headset and get started. Two wizards took turns simulating HCA through speech and movement control. Their main support was a hypertext document organised hierarchically by domain and topic, enabling quick navigation to find appropriate output to the user in the discourse context. The wizards were trained in advance, the training being supported by a written Wizard's guide, and instructed to make notes which were discussed in day-to-day briefing sessions.

Figure 2. A wizard in action.

Basic Data

The basic turn-level simulation data are shown in Table 1. Turn numbers measure the total number of turns made by the user and HCA in a conversation. Since they take turns communicating, each of them will produce half of the turns +/- a single turn.

The total of 498 conversations excludes four conversations of <4 turns and two conversations in which the transcribers thoroughly mixed up the users. All other recorded conversations are included in Table 1. The reason for leaving out the <4 turns exchanges is that there is hardly any conversation if what happens is merely a user saying, e.g., "Hello" and HCA responding, e.g., "Hello, welcome to my study". The reason why Table 1 provides information on users' age, gender and nationality, is that HCA has as a priority in conversation to gather this information from the users in order to use it as the conversation proceeds. He will thus try to collect this information either up front or, at least, early on in each conversation. Age information was provided by 91.0% of all users, gender information by 89.2%, and nationality information by 87.1%. The most common reason, by far, for not providing age, gender, and/or nationality information was that the user broke off the conversation before HCA could gather this data. This is evidenced by the facts that the average number of turns for age-unknown users is as low as 13 and the average number of turns of gender-unknown users is similarly low at 14 (Table 1). In a few cases, the wizards forgot to ask for the information. Few users refused to tell HCA their age or gender, and only in a couple of cases is there reason to believe that a user gave deliberately wrong information. An example is Maria on Day 9 who first had a 98-turn conversation as Maria, an 11 years old female from Denmark, and then came back to have a 24-turn conversation as Maria, a 13 years old boy from Denmark wanting to discuss girls with HCA, unfortunately with limited success.

Table 1 shows a rather close gender balance of 210 (47.3%) female users and 234 (52.7%) male users, as well as near-

identical turn averages for female and male users, i.e. 30 and 29, respectively.

Item counted	Totals
No. conversations	**498**
Age <10	**49**
Age 10-18	**240**
Age >18	**164**
Age unknown	**45**
Male	**234**
Female	**210**
Gender unknown	**54**
No. countries	**29**
No. turns all	**13739**
Av. no. turns all	**28**
No. turns <10	**1267**
Av. no. turns <10	**26**
No. turns 10-18	**7563**
Av. no. turns 10-18	**32**
No. turns >18	**4328**
Av. no. turns >18	**26**
No. turns age unknown	**581**
Av. no. turns age unknown	**13**
No turns male	**6689**
No. turns female	**6310**
Av. no. turns male/female	**29/30**
No turns gender unknown	**740**
Av. no. turns gender unknown	**14**

Table 1. Basic simulation data.

To enable analysis of the extent to which the specified first HCA prototype actually does reach its target user audience, Table 1 splits the users into three age groups: the under-10 year olds, the 10-18 year olds, and the over-18 year olds, representing approx. 10.8%, 52.9%, and 36.2 of the users who told HCA their age, respectively. The relatively low proportion of under-10 year olds may be explained by the fact that most under-10 year old users come from nations in which English is not a first language and hence do not speak English well enough (yet) to engage HCA in conversation. The top-five nationalities in per cent of those who told HCA their nationality, are: Denmark (28.3%), The Netherlands (15.2%), Sweden (11.3%), Norway (9.2%), and Germany (6.7%). The first nation having English as first language is the USA in 6th place (5%). In conformance with the explanation above, we find a higher proportion of speakers from countries having English as first language among the under-10 year olds, i.e. 14/40=35.0%, than of speakers from English speaking countries in proportion to all speakers of known nationality, i.e. 54/434=12.4%.

Reaching The Target Users

Let us define a turn-level metrics called *relative target group success* (RTGS) in order to quantify how well the simulated application manages to appeal to its target users as compared with its appeal to other user groups. Since the application is designed for edutainment, we consider length of conversation a component measure of success: the longer a user wants to talk to the system, the more successful is the system in meeting its edutainment objectives. We therefore propose to initially measure target group success as the percentage difference between average turn length for the target group and for each of the non-target user groups, i.e.:

$$\mathrm{RTGS} = \frac{\mathrm{TG\text{-}OG(n)}}{\mathrm{OG(n)}}\,\%$$

where TG is the target group and OG(n) is some non-targetted user group.

Although we will be applying the metrics to target and other *age* groups, the metrics itself is independent of group definition. It may just as well be applied to, e.g., male and female users of an application targetted at female users.

For the three age groups, i.e. the <10, 10-18, and >18 year olds, the average turn number is 26, 32, and 26, respectively (Table 1). Thus, overall, the target user group conversations are, on average, 23.1% longer than the conversations with both non-target user groups. However, before considering this result an authoritative measure of RTGS, we need to consider the following problem.

	NNE <10	**NNE 10-18**	**NE <10**	**NE 10-18**	**NE >18**	**NGE <19**
No. users	26	203	14	23	17	29
No. turns	670	6396	514	878	468	1019
Av. no. turns	**26**	**32**	**37**	**38**	**28**	**35**

Table 2. Speaker origins. NNE is non-native English speakers, NE is native English speakers, NGE is native or good English speakers.

The 226 10-18 years old users with known nationality in the corpus are mostly non-native English speakers. Only 10.2% (=23) are native English speakers by country, i.e. come from countries which have English as a first language (Table 2). The rest, i.e. 89.8%, may be presumed to be in the process of learning English as a second language. These users are likely to be less articulate than native English speakers in conversation with HCA. We hypothesise that they might therefore tend to stop the conversation earlier than they would have done had their English been more fluent. This would make it difficult for them to match the turn average of their native English speaking counterparts of the same age. The hypothesis, thus, is that the simulated application may well have a higher-than-23% RTGS since most target users may have had a somewhat briefer conversation with HCA than they would have had, had their English skills been more mature.

To test the hypothesis, let us first compare the turn averages of the 10-18 years old native and non-native English speakers (-by-nation). Table 2 shows that the native English speaking target users have a considerably higher turn average, i.e. 38, than the non-native English speaking target users whose turn average, i.e. 32, is the same as the one for all 10-18 years old users (Table 1). This effect of mastering the English language is confirmed when we look at the turn averages of the <10 year olds. The native English speaking kids have a turn average of 37 whereas their non-native counterparts are down at 26 turns. To control for the possibility that mastery of English could be the key factor in making users speak longer with HCA, we may compare the turn average for native English speaking target users with that of native English speaking adults. Table 2 shows that the native English speaking adults had 28-turn conversations with HCA on average. This is only two turns, or 7.7%, above the average number of turns for adults in general (Table 1), showing that, although English mastery may have an effect on the length of user-HCA conversations, this effect is far smaller than the effect of belonging to the target user group. As a final test of the hypothesis of the effect of language mastery on RTGS, we may consult the linguistic grading of the English proficiency of all users on a four-point scale from bad through medium to good and native. Table 2 shows the turn average of all <19 years old native and good English speakers from Day 1 through Day 5 in the corpus. The average of 35 would seem to smoothly fit the hypothesis that, the better the English of the target users, the higher their turn average.

In conclusion, whether or not a user is in the target age group, the better the user's English skills, the longer that user is likely to speak with HCA up to 38 turns on average per conversation. Considering native English speakers only, the <10/10-18 RTGS is only 2.7% whereas the 10-18/>18 RTGS is 35.7%. These figures are +/- an estimated factor <0.1 since approx. 10% of all users did not tell HCA their age and/or nationality and since those users had far briefer conversations with him.

The marked RTGS difference just described between, on the one hand, the <10/10-18 years old and, on the other, the 10-18/>18 years old, suggests that the application clearly has stronger appeal to the <19 years old than to adults. This conclusion is supported by another finding, i.e. that the top-ten user-HCA conversations, which have a staggering average of 111 turns, all involve 6-17 years old youngsters.

Conclusion

This paper has proposed a simple turn-level metrics called relative target group success (RTGS) for quantifying how well an edutainment or entertainment application manages to appeal to its target users. The metrics were then applied to a relatively large (13.739 turns) WoZ corpus. It was shown that the RTGS was highly dependent on whether the defined user groups could or could not be assumed to have English as a first language. This led to the conclusion that RTGS must be measured for native speakers.

Assuming significant numbers of native English speakers in future field tests of the system, the RTGS metrics can be used directly for progress evaluation. However, even in the absence of significant numbers of native speakers, we might use the figures reported above heuristically as incremental constants. We have seen (Table 2) that: native English speaking <10 year olds talk 42.3% longer with HCA than their non-native English speaking counterparts; 10-18 years old native speakers talk 18.8% longer with HCA; and native speaking adults talk 7.7% longer with HCA than all adults (Table 1). In the absence of hard data on, e.g., <19 years old native speakers, we might compute the <10/10-18 years old RTGS for the application using non-native data as:

$$\text{RTGS TG:<10} = \frac{(\text{TG}+18.8\%)-(<10+42.3\%)}{<10+42.3\%}\,\%$$

The more future test <10/10-18/>18 non-native English turn average proportions mirror those found in the WoZ corpus, the more reliable this heuristics might be.

We obviously aim to maximise TG/non-TGs RTGSs in future work, especially the TG/adult RTGS. However, we have no idea of what might be a satisfactory RTGS in absolute terms. In fact, this question may be undecidable. A hard question which does require an answer, on the other hand, concerns *absolute* entertainment success evaluation. For instance, does an average of 38 turns (Table 2) demonstrate edutainment success in absolute terms? If not, how high must the average be? We hope that the upcoming controlled target user test with the first HCA prototype will provide part of the answer, among other things because that test will allow us to interview the target users, something which is notoriously difficult to do in field trials such as the one reported above.

When applying the RTGS metrics, care must of course be taken to exclude other possible factors. In the present case, e.g., wizard differences do not seem to influence the results.

Acknowledgments

The NICE HCA work is being supported by the EU Human Language Technologies programme under contract IST-2001-35293. We gratefully acknowledge the support. Thanks are also due to the excellent work of the wizards and to Thomas Hansen who graded the users' English language proficiency.

References

1. Bernsen, N.O., Dybkjær, H. and Dybkjær, L. *Designing Interactive Speech Systems. From First Ideas to User Testing.* London, Springer Verlag, 1998.
2. NICE: http://www.niceproject.com/
3. Walker, M., Kamm, C., and Litman, D. Towards developing general models of usability with PARADISE. *Nat. Lang. Engineering* 6, 3, 2000.

Describing children's intuitive movements in a perceptive adventure game

Johanna Höysniemi
Tampere Unit for Computer-Human Interaction,
Department of Computer Sciences
FIN-33014 University of Tampere, Finland
johanna@cs.uta.fi

Perttu Hämäläinen
Telecommunications software and multimedia laboratory,
Helsinki University of Technology
P.O.Box 5400, FIN-02015 HUT
pjhamala@tml.hut.fi

Abstract

The controls of computer vision based action games need to be intuitive and physically appropriate in order to provide a pleasant gaming experience. The current research, however, does not provide enough data on what movements children find natural in specific game contexts. We have used the Wizard of Oz methodology to gather children's movements and experimented with different ways to study and describe these gestures. Labanotation was used to notate the movements, but as the method turned out to be time-consuming, more easily applicable methods were employed to inform the design of the computer vision algorithms as well as game character animations.

Keywords

Human movement, notation, intuitive movements, wizard of oz prototyping, computer games, children

Introduction

In the recent years physically more activating control mechanisms such as dance mats and video cameras have entered the market [5,13]. Our work [8,14] focuses on the design of perceptive children's action games that are played using body movements and voice, and designed to support children's physical development. The game is both perceptual and multimodal: the user controls an animated 2D character that mimics the user's movements and use of voice. The game works on a Windows PC equipped with a low-cost web camera and a microphone. This study, however, focuses only on the physical modality of the game.

The key requirements for computer vision based game controls are robustness, responsiveness, intuitiveness and physical appropriateness; the last two in particular since they make the learning phase shorter and also facilitate an enjoyable playing experience. However, there is not enough research on what movements children find intuitive in different game contexts. Moreover, the development of computer vision is laborious, and thus the game concepts need to be evaluated with children before putting extensive effort into building functional prototypes. Due to the lack of usable prototyping tools, we have used the Wizard of Oz (WOz) methodology to gather children's movements during simulated game playing sessions as shown in Figure 1. The WOz study was carried out with 34 children of ages 7 to 9 in a local elementary school and several hours of video material were recorded during the children's play activities.

Figure 1. The wizard controls the game prototypes with keyboard and mouse according to the player's actions.

The gathered video data needs to be studied carefully to facilitate the design process later on. Unfortunately, there are no generalized standards for analyzing and annotating human movements in HCI, possibly due to the high variation in gestures used with different systems and input techniques. These discrepancies also make it difficult to use and employ existing coding schemes fully and effectively. The fast-paced and iterative nature of computer game development also sets requirements for selecting an appropriate method and level of detail in describing the gestures. Moreover, the building of extensive human movement databases or the employment of motion tracking tools presented in the literature is time-consuming and labor intensive. As the main challenge of the study is to find the appropriate movements for a large number of players, the categorization and comparison between children's movements is crucial. Even though inexpensive annotation tools such as Anvil [9] do exist, we also needed tools that allow the presenting of each child's movement sequences simultaneously and the grouping and maneuvering of these sequences. Another requirement for the analysis process is to be able to quickly build the video library for computer vision design purposes and to provide descriptions for the game character design. The main challenges in describing the movements and composing the video corpora were the following:

- What characteristics do children's game control gestures have and how do these affect the applied methodology?
- How to describe time-dependent and multidimensional movement data with sufficient accuracy even with "semi low-tech" tools and in reasonable time?

This paper discusses how preliminary video analysis influenced the applied methodology and what experiences we obtained using two different and atypical approaches to describe children's movements.

Human Movement Analysis

The disciplines that study human movement vary from psychology to sport sciences to choreography to human-computer interaction. Human motion analysis and representation has drawn the interest of the HCI community already since the 1970's [2], especially in the fields of computer vision and animation [1,6,12]. The visualization techniques of human movement, for example 3D animation, have made significant progress in the recent years. Nevertheless, there is still a lot to study in how to notate, describe and analyze human movement and how that data can be applied in the design of perceptual user interfaces.

Wizard of Oz approach

The Wizard of Oz method has been widely used to design and collect language corpora in speech-based systems [4]. We also were interested in collecting a corpus, but based on 7 to 9 year old children's body language and their intuitive game controlling gestures. In our WOz setup the wizard controlled action game prototypes with keyboard and mouse according to the player's movements as shown in Figure 1. The game prototypes were swimming, jumping, running and 'scaring the spiders' games. The children were not given any hints on what movements were expected because that could have constrained their physical expressions. The test setup was designed to be as unrestrictive and natural as possible, for instance, no markers were attached to the children's bodies. No measuring rods or tapes were used either since we did not want the children to feel that they would have to "perform", even though this would have facilitated more accurate movement descriptions. Two video cameras were used: camera 1 was positioned diagonally behind the player, and camera 2 was right in front of the player (same location as the web camera in a real playing situation). The footage of camera 2 can be used directly to evaluate the computer vision algorithms.

Preliminary Video Analysis

Human movements can be represented in a digital form in various ways [2,3,7]; videotapes, notation systems and movement databases. The difficulty in analyzing and representing human motion is often caused by the large size of the collected time-based data and a very specialized application area. Due to the nature of our WOz setup, we could only anticipate what kind of data we would obtain, and thus it was difficult to make any pre-test decisions on what tools and methods should be used to describe the data. Therefore, the recorded and digitized video material was first analyzed to find the requirements and means to simplify the description process, and then two different description approaches were tried out in order to find a suitable, yet time-saving way to define the movement data. After preliminary video analysis, we were able to define the characteristics that affect the selection of description methods described in the following:

The control of a fast-paced physically interactive action game usually involves both control movements and rapid transitions from one type of movement to another in order for the game to be challenging and balanced. The physical game "commands" can be divided into three categories by their nature: 1) *continuous* (e.g. swimming, running), 2) *sporadic* (e.g. jumping), and 3) *transitional* (e.g. from running to jumping).

Additionally, the movements appearing in the videos could be divided into two other categories; *obvious and non-obvious movements*, according to how easily most participants adopted the movement style and whether it could have been anticipated beforehand, for example, because of the game character animations. Running is an example of an obvious movement: most of the children ran when the game character was running. However, in the swimming game children adopted various swimming styles, such as dog stroke or crawl that both belong to the "swimming" category but are very different movements from the perspective of computer vision design. The analysis process of obvious and non-obvious movements is different. The main emphasis in the definition of obvious movements is to find accurate descriptions that define, for example, how participants run when they control the game character, and then collect a set of video sequences that illustrate how most of the children performed that type of movement. The analysis of the non-obvious movements is slightly different; the focus is on categorizing the children's movements and finding the underlying patterns in order to define guidelines for further computer vision and animation design. Additionally, the video corpora based on the non-obvious movements are more versatile; even one child can change his or her movement from one style to another during game play.

Another characteristic of movement description is *accuracy/ambiguity*. These qualities are closely related to the specific needs of the computer vision design. The level of accuracy is dependent on input device technology, for example, whether the developed system is based on one or two cameras, and what body parts are influencing the visual input. It is also important to define what parts of the data can be described in a more ambiguous manner in order to save time, for example, in cases where the computer vision design is still open to big changes and alternative options are available. As mentioned by Badler et al. [2], "natural

language descriptions are subject to ambiguity and unavoidable imprecision in specifying positions, dynamics, styles and other aspects of the movement". However, in categorizing the non-obvious movements, natural language can prove to be a valuable tool for preliminary analysis as long as a systematic categorization is maintained.
The videos of the 34 children playing the game contained a substantial amount of data. One means to simplify the description process is *sampling*, i.e. only selected sequences of a child's movements are described. This is especially useful in situations where a child's movement pattern remains similar for longer periods of time. The sampling rate was defined to cover all variations in the children's movements.

Movement Description Methods

After the preliminary data analysis two approaches that allow visual comparisons between children's movements were tried out: 1) using Labanotation to notate the movements, and 2) describing the movements using simplified logging techniques and visualizations.

Labanotation

Labanotation, developed by Rudolf Laban [2,10], is a symbol system for representing movement of the human body in space and time. Similar to music notation, Labanotation uses a staff which consists of columns for indicating the body part that moves, as shown in Figure 2. Time runs vertically from bottom to top, and the duration of a gesture is represented by the length of the symbol. Labanotation can be described as a skeletal model where all descriptions may be formulated in terms of positions of the joints, coupled with an understanding of how these joints are inter-connected [2].

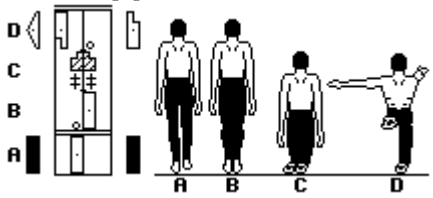


Figure 2. An example of a Labanotation staff. The center column represents the center of weight. The other columns (left and right side of the body) indicate the movement of the body parts that do not carry weight (www.rz.uni-frankfurt.de/~griesbec/LABANE.HTML).

The benefits of Labanotation are evident: it allows visual comparisons between the participants' movements. The symmetry and asymmetry of the movements as well as the rhythm and the length of the motion can easily be seen. However, the notation system could not inform us on how much power, tension or space children use while performing the movements. Laban's effort and shape theory [11] could have provided us with valuable parameters for the qualitative aspects of movements. However, that would have engaged us in an even more time-consuming description process.
The Labanotation staffs were also too detailed compared to the design needs of the system being developed. Labanotation describes movements in three dimensions, but this is often unnecessary for perceptual game user interfaces. We have previously designed interfaces based on two-dimensional motion analysis, using the game context to simplify the technology [8], e.g. in some games the center of mass of detected motion can be interpreted as the position of the user. This is a computationally simple and robust method even with a single web camera. The biggest drawback of the method is not the laborious nature of Labanotation, but to master the notation system, a lot of practice is needed. It is also difficult to use the staffs as a communication tool in the design process, if all team members do not fully understand the notation.

Low-level description techniques

As using Labanotation required such an extensive effort, we decided to experiment with less complicated description and visualization methods. The tools used were image and video editors, spreadsheets, and pen and paper. First, all the events appearing in the videos were listed, and then the appropriate movement sequences were sampled (such as a child's running movements in the beginning, middle and/or end of the game in addition to the places where there were distinctive variations in the running pattern). However, for non-obvious movements one preceding step was required to find all variations of these movement types. For example, in the swimming game shown in Figure 3 all swimming stroke variations were first listed (in all 17 different stroke types were found) before they could be further examined and compared to find patterns and similarities between them. Finally, the swimming types were summarized into four main movement styles; breast stroke, dog stroke, crawl and "mole" stroke. In addition, each non-obvious movement type was analyzed based on its popularity, i.e. whether it was the first movement type a child tried, if it was the main movement type for that child, and how many children actually used the style in question. We acknowledge that swimming styles may be culturally dependent. However, the study helped us find styles that we did not anticipate beforehand due to Finnish swimming education practices.

Figure 3. Three different swimming styles.

The coding schemes were further defined according to the requirements of the computer vision design. It was important to describe the movement on a 2D plane and focus on

the speed and amount of the movement, with particular focus on the movement of the upper body (since children varied their distance from the display and web camera, which hid their legs partially). The computer vision algorithms that, for example, detect a child's running can be based mainly on physical cues appearing above the waist level. Additionally, all anomalous happenings, such as rapidly moving braids of hair and clothing, knees raised high and so forth needed to be reported. The challenges for defining the coding schemes were to define the codes for each different movement and to measure and often approximate the numeric values such as the angles of the joints and the child's vertical movement during the movement cycle. For example, we used the proportions of the head as a measuring unit for comparing vertical movement along the running cycle. Fortunately, precise values are not necessary from the point of view of computer vision design which in the end mainly relies on the video sequences produced during the description process.

The detailed analysis of the movements was mainly based on image sequences of the videos which are typically only 5 to 20 PAL video frames, i.e. 200 to 800 milliseconds in length (such as one running or jump cycle). The sequences are usually easier to analyze by having all the frames visible side by side compared to the frame-by-frame manipulation of video editors or annotation tools. Also, preliminary comparison and grouping of the children's movements was more straightforward when all movement cycles printed and categorized manually. Another benefit of having the movement cycle visible in its entirety was that the frequency of steps, different phases in the movement cycle and their duration were relatively easy to measure. Additionally, the sequences can be used more easily than the Labanotation staffs to inform the game character design later on.

Summary

This paper focused on the preliminary analysis of the video data obtained during simulated physically interactive game play and the experiences of trying out two approaches that allow the visual comparison of children's movements. We admit that the results of our study are context dependent and possibly not applicable in other game contexts. The detailed descriptions of the movements and the experiences obtained when applying the video libraries and movement descriptions in the design of the computer vision algorithms and game character animations will be reported in the future.

Acknowledgments

We would like to thank all the children who participated in our tests. We are deeply grateful to our sponsors for providing us with the financial support needed to carry out the study.

References

1. Aggarwal, J. K, and Cai, Q. Human motion analysis: A review, *Computer Vision and Image Understanding*, Vol. 73, No. 3, March 1999, pp. 428-440.
2. Badler, N.I., and Smoliar, S.W. Digital representations of human movement, *Computing Surveys,* Vol. 11, No. 1, March 1979, pp. 19-38.
3. Ben-Arie, J., Pandit, P., and Rajaram, S. Design of a digital library for human movement, In *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*, Roanoke, Virginia, United States, 2001.
4. Dahlbäck, N., Jönsson, A., and Ahrenberg, L. Wizard of Oz Studies -- Why and How. Proc. of the International Workshop on Intelligent User Interfaces, 1993.
5. D'Hooge, H. Game design principles for the Intel Play Me2Cam* virtual game system, *Intel Technology Journal Q4*, 2001.
6. Gavrila, D.M. The visual analysis of Human Movement: A survey, *Computer Vision and Image Understanding*, Vol. 73, No. 1, January 1999, pp. 82-98.
7. Grünvogel, S., Piesk, J., Schwichtenberg, S., and Büchel, G. AMOBA: A database system for annotating captured human movements. In *Proceedings of Computer Animation 2002*, IEEE Computer Society, Los Alamitos, pp. 98 - 102. Geneva, June 2002.
8. Hämäläinen, P., and Höysniemi, J. A computer vision and hearing based user interface for a computer game for children. In *Proceedings of the 7th ERCIM Workshop "User Interfaces For All"*, 23-25 October 2002, Paris.
9. Kipp, M. "Anvil - A Generic Annotation Tool for Multimodal Dialogue". In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, September 2001, Aalborg, Denmark, 1367-1370,
10. Laban, R. *Laban's Principles of dance and movement notation* (2nd Edition). Macdonald & Evans Ltd., London, 1975.
11. Laban, R., Lawrence, F. C. Effort: Economy in Body Movement. Plays, Inc., Boston, 1974.
12. Neagle, R.J. and Ng, K.C., Machine-representation and Visualisation of a Dance Notation, in *Proceedings of Electronic Imaging and the Visual Arts*. London, UK, July 2003.
13. Sony EyeToy. http://www.eyetoy.com/. January 2004.
14. QuiQui's Giant Bounce. http://www.kukakumma.net

A coding scheme for the annotation of feedback phenomena in conversational speech

Loredana Cerrato

TMH-CTT
Department of Speech Music and Hearing,
KTH Stockholm, Sweden
loce@speech.kth.se

Abstract

A coding scheme specifically developed to label feedback phenomena in conversational speech is presented in this paper. The coding scheme allows the categorization of feedback phenomena according to their typology, direction, and communicative function in the given context.

The results of the reliability tests run to verify the appropriateness of the coding scheme to code feedback phenomena in different languages and across different modalities are also presented.

Keywords

Multimodal annotation, verbal and gestural feedback phenomena, coding scheme.

1. Introduction

One of the most important phenomena in human communication and probably an index of conversation fluency is the production of feedback. Participants in a conversation continuously exchange feedback as a way of providing signals about the success of their interaction. Feedback can therefore be considered as a sort of "running commentary to what the current speaker is saying or doing" as Poyatos states [12, p. 241]. As a consequence the success or failure of a conversation relies much on feedback, for this reason the study of feedback phenomena is achieving popularity in the field of human-machine interfaces and speech technology development. This is due to the fact that researchers, being aware of the important role that feedback phenomena play in communicative exchanges, try to integrate some of them in the development of dialogue systems with embodied conversational agents, with the aim of enhancing their performance [10]. However the production of an accurate model of feedback realization is a time-consuming process, which requires extensive and detailed analysis of the feedback phenomena used in real communicative situation by human beings. Previous studies have shown that it is possible to categorize vocal feedback expressions according to their behavioural form, and/or to the function they accomplish [1], a similar categorization can be applied to gestures produced with feedback function [3]. In order to be able to analyse communicative phenomena, it is necessary to dispose of good quality materials, a specific coding scheme and of a dedicated annotation tool. This paper gives a detailed non-formal presentation of a coding scheme developed to analyse feedback phenomena, which has been used to code human-human dialogues [5] and human-computer interactions [6].

The issue of reliability of the coding scheme is also addressed. The results of stability and reproducibility tests run to verify the appropriateness of the coding scheme in different languages and across different modalities are presented.

2. Feedback Coding

In order to be able to code feedback expressions it is necessary to identify them in the first place. To do so it is crucial to take contextual information into account, which means interpreting and categorising feedback expressions in terms of reactions to the previous communicative act. The definition and coding of feedback used in this study is based on the approach proposed by Allwood [1], who defines feedback as "linguistic mechanisms, which enable the participants in a conversation to unobtrusively exchange information about four basic communicative functions: contact, perception, understanding and attitudinal reactions". These functions are related to basic requirements of human communication, in fact in order to carry out a successful communication it is necessary that two participants establish a contact with each other; once the contact is established it is possible to produce a message, which should be perceived by a receiver, who must be able and willing to understand it. Moreover interlocutors might show attitudinal and behavioural reactions towards the meaning conveyed; this includes assent, negation or contradiction, assertion surprise, disappointment and enthusiasm and so on.

Feedback signals are displayed under different forms (by means of verbal and vocal expressions and by means of gestures), transmitted through different channels (visual, auditive, tactile) and produced with several communicative functions (continuation, acceptance, refusal and so on), for this reason the analysis of feedback requires the support of an accurate coding scheme and flexible tools for the

annotation of audio-visual materials. The annotation tool Multitool [2] has been used to analyse materials and carry out the coding. Multitool allows achieving the specific purpose of transcribing, annotating and analysing human behaviours and other visually accessible information in temporal alignment with speech. The synchronisation needs to be done manually by the user.

The annotation is performed on a freely definable multi-layered annotation scheme (consisting of tiers), which can be *ad hoc* defined to label materials at different levels, for instance: speech, gestures and so on.

The coding scheme here presented is specific for the annotation of verbal and gestural feedback expressions, and takes into account the typology of the identified feedback expression, its direction type and its specific communicative function in the given context.

2.1 Typological coding

Feedback expressions are typologically labelled as:
Words (**W**), **P**hrases (**P**), **S**entences (**S**) and **G**estures (**G**).

The label **W** is used to code expressions consisting of one item, such as "yes, no, ok", m-like sounds, and vocalizes.

Some additional labels can be used to specify other characteristics of the feedback expressions, as for instance to indicate whether the feedback expression has been produced in a minimal non-intrusive way, or as part of a longer utterance.

The label **G** (**G**esture) in the typological coding is given to those movements produced to signal feedback without the production of verbal feedback expressions.

When a gestural and a verbal feedback expression are produced simultaneously the annotation specifies also the typology and function of the gesture [3]. Gestures can be:
Head movements (nods, jerks, waggles,etc);
Facial expressions (eyebrow movements, gaze, smile);
Other gestures (shoulder, hand and trunk movements).

Additional labels can be used to signal whether the gesture is repeated, or produced in time sequence with another gesture.

2.2 Direction of feedback

It is possible to distinguish between strategies for giving feedback and strategies for eliciting feedback. This is what Allwood [1] defines as "directional function type" or "orientation". Participants in a conversation give feedback when they wish to show their interlocutor that they are willing to continue the communication and that they are listening, paying attention, understanding/not understanding or agreeing/disagreeing with the message being conveyed. They elicit feedback when they wish to know whether the interlocutor is listening, paying attention, understanding, or agreeing, disagreeing with what they are saying. When feedback is **Giv**en, it is labelled as (**Giv**), when feedback is **Eli**cited, it is labelled as (**Eli**).

2.3 Functional coding

The functional coding of feedback expressions provides a fine-grained categorization of the expressions according to their specific, explicit functions in the given context. It is assumed that feedback expression (both verbal and gestural) always carry out a communicative function.

2.3.1 Functions of verbal feedback expressions

Verbal expressions produced to give feedback can have the following functions:

Continuation (**C**), indicates that the interlocutor has **p**erceived (**p**) and possibly understood the message, but s/he explicitly shows only his/her willingness to go on in the communication, either by signalling the intention to interrupt the current speaker and take the turn (**I** = ***I*** *want to go on*) or by letting the interlocutor continuing to speak (**Y**= ***Y****ou go on*). This is usually done by producing a minimal non-intrusive feedback expression (usually a short verbal expression and/or a head nod). In order to be able to produce a continuation feedback it is necessary to have perceived the message, but it is not necessary to have understood it.

The notion of continuation feedback here includes also the notion of signalling turn-taking [8]. What is annotated with this coding is the explicit function of feedback, which includes an indication of whether the speaker who gives feedback intends explicitly to signal the intention to take the turn or not.

Acceptance (**A**) indicates that the interlocutor has perceived and understood the message and wishes to show acceptance. This implies contact perception and understanding in Allwood's terms and includes Clark and Schaefer's acknowledgement [7], which describes a hierarchy of methods used by interlocutors to signal that a contribution has been understood well enough to allow the conversation to proceed.

Refusal (**R**) indicates that the interlocutor wishes to show refusal, non-acceptance of the information received. This does not always imply contact, perception and understanding, since the information can be refused because of misperception, misunderstanding and disagreement.

Expressive (**Ex**), specifies that the interlocutor wishes to colour his feedback with some attitudinal/emotional reactions towards the meaning conveyed; this includes surprise, disappointment, frustration, enthusiasm and so on, and implies contact, perception and understanding. The functions and the relative labels that expressions produced to give feedback can have are shown in table 1.

When feedback is elicited, the communicative functions it can carry out are: **Req**uire **C**onfirmation of understanding (**Req-C**), ensure interlocutor's attention, agreement and understanding, in other words check that the interlocutor is **Fol**lowing (**Fol**), show the desire for **Mo**re information (**Mo**). The functions and the relative labels of expressions produced to elicit feedback are given in table 2.

Function	Label	Comment
Continuation	**CpI**	I want to go on
	CpY	you go on
Acceptance	**A**	Understanding, agreement, acceptance
Refusal	**R**	Understanding/misunderstanding, refusal
Expressive	**Ex**	expression of an attitude, emotion, point of view

Table 1 Labels used to code the communicative function of expressions that elicit feedback.

Function	Label	Example
Require **C**onfirmation	**Req-C**	"isn't it?"
Check that the interlocutor is **fol**lowing	**Fol**	"are you following"
Desire to receive **Mo**re information	**Mo**	"ok, then?"

Table 2 Labels used to code the communicative function of expressions that elicit feedback.

2.3.2 Functions of the gestural feedback expressions

In human communication a gestural feedback can be produced to accompany a verbal feedback expression. In this case it is assumed that gestural feedback carries out a specific function considered in relationship to the function of the accompanied verbal feedback [9,12].

Function	Label	Comment
Addition	**Ad**	the gesture adds info to the verbal feedback
Emphasis	**Em**	the gesture positively reinforces the verbal feedback
De-Emphasis	**D**	the gesture weakens the verbal feedback
Contradiction	**Con**	the gesture contradicts the verbal feedback

Table 3 Labels used to code the relationship between the function of gestural and verbal feedback expressions.

The function of a gestural feedback can either be **N**eutral (**N**), which means that the gesture simply accompanies the vocal/verbal expression, without modifying its meaning, or can have the functions and labels shown in table 3.

3. Reliability Of The Coding Scheme

A reliability test has been run in order to test whether the coding scheme is appropriate to code feedback phenomena in different languages (i.e. Italian and Swedish) and across different modalities (i.e. audio and visual).

According to [5,11] there are three ways of testing the reliability of a coding scheme:

1) **Stability test**, or inter-variance test, which checks whether the same coder varies his/her judgments over time.

2) **Reproducibility test**, or intercoder-variance, which checks the agreement in the coding of two coders.

3) **Accuracy test**, which compares the codings produced by these two coders to the standard, if the standard is available.

The reliability of the coding scheme proposed in this paper was tested by running stability and reproducibility tests using the following materials:

one map-task dialogue in Italian, available only in audio format (from now onward referred to as IT1.MPA).

one map-task dialogue in Swedish, available only in audio format (from now onward referred to as SW1. MPA).

one dialogue recorded in lab-environment, in audio-video format (from now onward referred to as SWL3.G)[1].

The stability test was performed on the coding made by an expert coder (i.e.author), who first coded all the materials and after about six months repeated the coding on 22 feedback phenomena in each dialogue. The percentage of feedback identification was 100% for the two Map Task dialogues (IT1.MPA, SW1. MPA) and 91% for the SWL3.G dialogue. The overall agreement on the entire coding scheme has been calculated using the kappa coefficient[2].

Reults between 0.41 and 0.6 indicate moderate agreement, between 0.61 and 0.8 indicate substantial agreement and between 0.8 and 1 nearly perfect. Following this interpretation, the results of the stability test, shown in table 4, indicate that the coding is stable over time and the agreement for categories assignment is substantial. These good results should not be surprising, since the expert coder is also the developer of the coding scheme.

Materials	K coefficient
IT1.MPA	0,77
SW1. MPA	0,94
SWL3.G	0,68

Table 4 Result of the stability test, consistency among successive codings of the same coder.

For the reproducibility test two linguists, one native speaker of Swedish with good fluency in Italian (SW-Co), and one Italian native speaker with good fluency in Swedish (IT-Co), were asked to code verbal and gestural feedback in the same materials used by the expert coder to perform the stability test. The two linguists received both an oral explanation and written instruction on the task and

1 More details on the materials are available in previous publications by the author.

2 K= P(A) –P(E)/(1-P(E))

where P(A) is the portion of times that the coders agree and P(E) is portion of times that we would expect the coders to agree by chance.

the details about the coding scheme. Before starting their task they listened and looked at some examples of feedback to get accustomed with their task. The test took about 5 hours. The first task of the test consisted in the identification of feedback expressions. It would have not been possible to test the reliability of the coding scheme if the coders would not agree on identification of the units.
The agreement on identification of feedback expressions for the coders is shown in table 5, listed per dialogue. The identification of the verbal and gestural feedback expressions presented few differences among codings, also the assignment of typological labels showed few disagreements. Most of the disagreements occurred, in fact, in the assignment of the labels for the function of verbal and gestural feedback expressions. For the function of the verbal feedback expressions the categories **CpI** and **CpY** were often confused in Swedish, while in Italian the label **A** was the most difficult to assign. As concerning the function of gestural feedback, most confusion occurred in the assignment of the labels **Ad** and **Em**.

Materials	Inter-coder agreement on feedback identification	
IT1.MPA	19 of 22	86%
SW1. MPA	20 of 22	91%
SwL3.G	19 of 22	86%

Table 5 Inter-coder agreement for feedback identification.
The overall agreement on the entire coding scheme resulted to be quite good. The result of the K test are shown in table 6.

Materials	(K coefficient)
IT1.MPA	0,6
SW1. MPA	0,69
SwL3.G	0,67

Table 6 results of the reproducibility test: overall inter-coder agreement.

It should be mentioned that the identified feedback expressions all have a "Giving" direction. This is due to the fact that giving feedback occurs more often than eliciting feedback in the available materials. As a consequence, only the labels for the function of "given" feedback were used in the test (i.e. the labels in table 2).

4. Discussion

A non-formal presentation of a coding scheme developed to analyse feedback phenomena in speech communication was presented in this paper. The promising results of stability and reproducibility tests run to verify the appropriateness of the coding scheme in different languages and across different modalities can be interpreted in terms of reliability and ease of use of the coding scheme, which has its strength in the fact that is intended to be used for the analysis of a specific communication phenomena: verbal and gestural feedback. Naturally the positive results of the test can be also questioned given the limited amount of test material and the restricted number of coders who take part in the test.

References

1. Allwood J, (ed) 2001, "Dialog Coding - Function and Grammar. Göteborg Coding Schemas". Gothenburg Papers in Theoretical Linguistics, 85. Department of Linguistics, Göteborg University.
2. http://www.ling.gu.se/~mgunnar/multitool/
3. Allwood J, Cerrato L, 2003, A study of gestural feedback expressions. The First Nordic Symposium on Multimodal Communication, Copenhagen (in prees).
4. Carletta, J, et al. 1997, The Reliability of a Dialogue Structure Coding Scheme. Comp. Linguistics, 23(1), 13-31.
5. Cerrato L, Skhiri M, 2003, A method for the analysis and measurement of communicative head movements in human dialogues. *Proc of AVSP '03,* 251-256
6. Cerrato L, 2002, A comparison between feedback strategies in Human-Human and Human-Machine communication. *Proc of ICSLP '02*, 557-560
7. Clark H, Schaefer E, 1989, "Contributing to Discourse" In *Cognitive Science* 13, 259-294
8. Duncan S, Fiske D. 1977, Face-to-face interactions: research, methods and theory Lawrence Erlbaum Ass. Hillsdale N.J.
9. Ekman P. 1979, About Brows: Emotional and Conversational Signals. In M. von Cranach, K. Foppa, W. Lepenies, & D. Ploog (eds.), Human Ethology, 169-249.
10. Granström B, House D, Beskow J, 2002, "Speech and gestures for talking faces". In Granström B; House D; Karlsson (eds), 2002, Multimodality in Language and Speech Systems Kluwer Academic Publishers, 209-241.
11. Krippendorff K, 1980, Content Analysis: an introduction to its analysis. Sage Publications.
12. Poyatos F., 2002 Non-verbal communication across disciplines, J, Benjamin Publishing Company

Multimodal Score: an ANVIL™ Based Annotation Scheme for Multimodal Audio-Video Analysis

Emanuela Magno Caldognetto*, Isabella Poggi°, Piero Cosi*, Federica Cavicchio*, *G. Merola*°

*Istituto di Scienze e Tecnologie della Cognizione
Consiglio nazionale delle Ricerche – Italy[1]
°Dipartimento di Scienze dell'Educazione
Università di Roma Tre - Italy

Introduction

Face-to-face communication is multimodal: we communicate with voice, face, eyes, hand, body. But just a little part of these communicative instruments have been studied thoroughly: while Linguistics, since 2000 years, has studied the rules that govern verbal behaviour, not so much has been done for all other modalities.

Instead, to fully understand human multimodal communication, our task would be to write down the "lexicon" and the "alphabet" of nonverbal signals (Poggi, 2001): that is, on the one hand, to find out the systematic correspondences between signals and meanings in each mode-specific communication system; on the other hand, to single out all the minimal elements that compose all signals in each communication system.

To discover the elements and the rules that make up communication systems, as well as the rules of their simultaneous and sequential combination with each other, is a useful thing for both theoretical purposes and practical applications, such as, among others, the construction of Embodied Agents (Cassell et. al. 2000). But to do so, it is necessary to analyse corpora of Multimodal Communication by using precise methods of segmentation, transcription, and annotation of signals in the different modalities.

In a sense, this is a somehow circular endeavour. Our first task is to construct the alphabet and the lexicon of a communication system – for instance, to find out the correspondences between particular patterns of gaze signals and their particular meanings; to discover these correspondences it is necessary to analyse numerous items of gaze, and to this goal you must use a procedure for this analysis. But once an alphabet or a lexicon is singled out, it would be much easier and clearer how to analyse further corpora, to such an extent that it could sometime be possible to provide a tool for automatic analysis of Multimodal Communication. This is why the construction of tools for the analysis and annotation of multimodal data is an endless job.

In the last ten years, several tools have been proposed for this task: for example, Martin et al. (2001), Kipp (2001), Ingenhoff and Schmitz (2003) and Annual Reports of ISLE and NITE EU Projects. In this paper we present the multimodale Score implemented in ANVIL, a system for the annotation of multimodal data, which is characterized by the attention to the semantic aspect of corpora annotation and by an effort to find out, in the long run, a systematic correspondence between (aspect of the) signals and (aspect of the) meanings in order to exploit these results in future automatic annotation. To better illustrate the potentialities of Multimodal Score implemented in ANVIL we present some examples that show how it can be use in assessing the relationship of speech with gesture and *visual prosody*, of mouth movements with Sign language and the laying out of emotions and attitudes over speech.

1. The Multimodal Score

To describe Multimodality, Magno Caldognetto and Poggi (2001) suggest a method: the Multimodal Score, a procedure to transcribe and analyse the multimodal signals classified separately and in their mutual interaction. This method allows us to transcribe on five parallel lines, like in a musical score, the communication items transmitted at the same time in five modalities: *speech, prosody, gesture, facial*

[1] Part of this work has been sponsored by COMMEDIA (COMunicazione Multimodale di Emozioni e Discorso in Italiano con Agente animato virtuale, CNR Project C00AA71), PF-STAR (Preparing Future multiSensorial inTerAction Research, European Project IST- 2001-37599, http://pfstar.itc.it) and MIUR-FIRB *Nuove tecnologie per la formazione permanente e reti nel sistema socioeconomico italiano* project (MIUR-FIRB RBNE01MRXA_003).

(mouth, gaze, eyes, eyebrows), head and body posture, by labelling signals on five different levels.
In the Score, each signal of each modality goes through five levels of analysis:

- *description*: the gesture or movement is described on the basis of its perceptual characteristics. For example, in the gesture line it is possible to describe gesture both in a word transcription ("right hand draw an arch with the index") or in a codified transcription system by cheremes (the minimal unit of gestural communication, the "phonemes" of gesture, Stokoe 1980). Descriptions of facial and body movements are, for example: "eyebrow raising", "wide shut eyes" etc.
- *descriptive typology*: the gesture or movement is classified on the basis of a typology of gestures, including also self touch and not communicative case such as "hands at rest".
- *meaning*: the movement analysis is paraphrased with words or phrases;
- *meaning typology*: the meaning of each movement or gesture is classified on the basis of a semantic taxonomy that distinguishes Information about the World, the Speaker' Identity and the Speaker's Mind;
- *semantic function:* by comparing the gesture or movement with the coproduced speech signal, five different "functions" are distinguished, that is five kinds of relationships between them: **repetition**, if it bears the same meaning, **addition** if it adds information to word meaning, **substitution** if it replaces a word that is not uttered at all, **contradiction** if it communicates something opposite to what said by words, or **no relationship**, if it makes part of a different communicative plan.

The most distinctive characteristic of this annotation system is that it aims at identifying the meaning of each movement or gesture and translating it into words or sentences. For example a raising intonation contour at clause or phrase end could stand for "I did not finish my talk yet" or " I'm saying something important", the index finger stretched up could mean "attention please", and an eyebrow raising with wide shut eyes is paraphrased as "attention, what I'm saying is really important"; a posture shift as "I am changing the topic of my discourse".

2. The Multimodal Score implemented in ANVIL

In the following pages it is introduced the Multimodal Score implemented in M. Kipp's ANVIL (ANnotation of VIdeo and Language, 2001). This application, here presented in 3.6 version, is used by ISTC-CNR of Padua. At the Institute we are developing, with respect to the visual display planned by M. Kipp, more analytic evaluations linked to the acoustic analysis through PRAAT (Boersma, 1996). For example we insert the phones and syllable transcription level, beyond the phrases and sentences transcription and segmentation provided for ANVIL already. Furthermore we label the pitch and intensity contours on a qualitative basis. At the moment the main problem with ANVIL, is the lack of a quantitative scale for F0 and energy, so it is still impossible to quantify pitch contours and intensity. This blank will be overcome with a specific application.
In order to exploit our example of analysis of Multimodal Score in ANVIL, the first step is to define a menu of all the communicative modalities. Further for each modality we insert the Multimodal Score, i.e. the five levels of analysis previously presented (Poggi and Magno Caldognetto, 1996). In some cases (for example gesture), on the basis of previous researches, it is possible to label the movements with the help of pop up menus. As an example, in the following is reported a list of the pop up menus presently inserted in the Multimodal Score implemented in ANVIL 3.6, for gesture modality:

- *Type of Gesture*→ Batonic, Pantomimic, Pictographic, Symbolic, Deictic, Other, None;
- *Type of meaning*→ CI (Content Information, or Information on the World), SMI (Information on the Speaker's Mind), SP (Self-Presentation, or Information on the Speaker's Identity), Other, None;
- *Function of gesture movement with respect to speech*→ Repetitive, Additional, Substitutive, Contradictory, No relationship, Other;
- *Gesture/movement Segmentation*→ Preparation (start), Stroke, Retraction, End, None;
- *Relationship between hands*→ Mirror, Asymmetric, Independent, Other, None;

Similar pop up menus are already (or about to be) implemented for other modalities like facial expression, gaze, touch and self-touch, etc. To explain the multimodal Score implemented in ANVIL potentiality and the importance of Multimodal Communication we analyse five multimodal communication different typologies. In the following we briefly present the analysis of five examples showing how the ANVIL multimodal Score can also be usefully adopted to analyse, respectively: coverbal gesture (3a.); the strict synchronisation between speech and gaze (in particular eyebrow movements) in marking topic and comment of sentences (visual prosody, 3b.); the relationship between signing hand movements in Italian Sign Language and the concomitant nonmanual components (lip and mouth movements, 3c.); the overlapping of speech and emotion (3d.); facial movements and the expression of emotions and attitudes, and their relation to the vocal signal (3e.). The coproduction of these different not synchronised modalities is fully appreciable only in a multimedia presentation with ANVIL, thanks to whom is possible to outweigh numerousness and not synchronisation of units. In the following we will focus with particular attention to the description of the multimodal communication example concerning with emotions and attitudes. In fact, with respect to a short verbal message, the example 3e (see fig. 1) allows a clearer confrontation of different modalities of communication during the time sequence. Further, in

spite of its time shortness, it is really rich on the multimodal communication point of view[3].

3. Five examples of Multimodal score implemented in ANVIL

3a. Speech and Gestures

In this example, taken from an interview, we analyse:
- acoustic analysis segmentation and labelling: a large number of pauses during the speech, synchronised with the gesture;
- coverbal gestures: the Speaker with his right hand performs a gesture ("the ring"), connected specifically with speech focus, showing an additional relationship with respect to speech;
- gesture coordination to the syllable level;
- mouth and lips linguistic movements (repetitive function);
- head movements: nodding movements toward the Listener (probably a related to feedback);
- gaze: directed to the one who is putting the question, and fixed on him.

3b. Speech and Visual Prosody

In this sample, taken from a TV news, we analyse the interaction between speech and visual prosody, particularly:
- acoustic analysis, segmentation and labelling: slightly descendant pitch contour with final raising (due to the interrogative intonation) and intensity with focus on the word *può* ("could"). rapid eyebrow movements, connected with the topic-comment distinction;
- rapid eyes opening and closing connected, respectively, with comment and topic
- rapid head movements linked to speech production and prosody;
- mouth and lips linguistic movements;
- stressed syllable coordination with the sentence topic.

3c. Italian Sign Language and Visible Articulatory Movements

In this example we analyse a particular sign of LIS (Italian Sign Language) standing for the word bombardare ("to bomb"), keeping our analysis on the visual cues:
- the hands produce a LIS sign standing for "*bombardare*";
- sign language transcription;
- eyes directed to the interlocutor in order to open a communicative channel with the interlocutor;
- emphasised labial movements, probably caused by lip reading needs.
- viseme and phonemes transcription with IPA.
- LIS sign and mouth movements are strictly connected to each other, and lip aperture and closure are emphasised;

3d. Speech and Emotions

In this example, taken from a laboratory experiment, we analyse the facial movements driven by speech and emotions (specifically anger):
- acoustic analysis, segmentation and labelling characterized by high intensity and short duration;
- eyebrows present a frowning movement starting before speech production;
- mouth and lips movements has paralinguistic function (the mouth is wide open because of the expression of emotion).

3e. Speech and Attitudes

In this example of 7 seconds (175 frames), taken from a commercial spot, it is displayed a particular facial expression pattern, linked to the acoustic characteristics of oral production. After phones and syllable segmentation of the word "buonasera", we analyse the F0 contour, that is characterized by a flat course in all the phrase, a narrow range and an extremely long duration.

Concerning with the visual cues, it is shown that the head is directed to the interlocutor from the frame 43, and slightly bow to the left from frame 77 and a right at frame 100. The head lowering movement has maybe a submission-flirting meaning.

The eyebrows are raised 17 frame before the actress starts speaking, the maximum of raising correspond to the syllable [buo] and held until the end of the syllable [se]. This movement is maybe due to surprise.

The eyes are half-closed for 25 frames and wide opened 23 frame before the first syllable [buo], in order to create a visual contact with the interlocutor. Successively there is a short blinking (4 frames) synchronised with the start of the first syllable, then the eyes are opened and they are wide open from the vowel [e] production until the end of the video. This long eye opening is maybe due to joy and surprise mixed together but also interest toward the interlocutor.

The gaze, subsequently to the eyes movements, is directed to the low and from frame 72, is turned to the interlocutor.

The mouth presents open smile from frame 40 until the end of the video, while linguistic movements of opening and closure, related to the speech production, starts at frame 93 and end at frame 142.

Left hand moves to tidy up hair from frame 61 and stops to the neck base at frame 138; it could be labelled as a self touching or embellishment movement.

The synchronic analysis of all the multimodal items based on the frame sequence is useful to confront the different signals, comparing the meaning of each signal and its relationship with speech.

In analysed sample speech, prosody, gesture, facial (mouth, gaze, eyes, eyebrows) movements, head and body posture start before speech production and carry on until the end of speech. It can be noted that the

[3]We will present at the workshop the whole analysis implemented in ANVIL for all the multimodal analysis.

speech production is preceded by the eyes opening, the eyebrow raising, transmitting surprise, and by the smile, indication of happiness. Further the synchronic vision it s useful to underline the co-occurrence speech and movements; the peaks of head, eyes, eyebrows, gaze and smile, correspond to tonic syllable.

4. Software Shortcomings

Actually, we find ANVIL a flexible and adaptable annotation system for the multimodal Score, but it will be developed for what concerns a better accurate sub-lexical acoustic analysis and labelling and improved for an accurate prosodic and intonation acoustic analysis and labelling (i. e. with *ToBI* system). We will also introduce different kinds of transcription systems (i.e. grapheme to phonemes IPA, phonemes to visemes system developed by ISTC, LIS sign Vocabulary). Unlike the qualitative descriptions now available, we want to introduce a quantitative description based, for example, on an analysis of the acoustic signal by PRAAT with visual signal captured and quantified by ELITE (for example Magno Caldognetto et al. 1998).

5. Theoretical Issues and Applications

The Multimodal Communication Score implemented in ANVIL is useful to examine, segment and label a large number of audio-visual token sets contributing to the creation of Multimodal Corpora: this is very important, on the theoretical research side, to find out the units and structure of communication system different from the verbal languages. On the other hand multimodal Score in ANVIL could be useful to outline production models applicable to the construction of Talking Heads and Virtual Agents.

References

Avesani C., ToBIt. "Un sistema di trascrizione per l'intonazione italiana" in *Atti delle V Giornate di Studio del Gruppo di Fonetica Sperimentale*, Trento, 1995, pp. 85-98.

Boersma P., (1996) "PRAAT, a System for Doing Phonetics by Computer", *Glot. International* 5 (9/10), pp. 341-345 (PRAAT web site: http://www.fon.hum.uva.nl/praat/).

Cassell J., Sullivan J., Prevost S. and Churchill E., "Embodied Conversational Agents", Cambridge (Mass.), The MIT Press, 2000.

Ingenhoff D. and Schmitz H., "Comtrans: a Multimedia Tool for Scientific Transcription and Analysis of Communication" in Rector M., Poggi I. and Trigo N. (Eds.) *Gestures. Meaning and use,* Universidad Fernando Pessoa Press, Porto 2003, pp. 389-393.

Kipp M. "From Human Gesture to Synthetic Action", in Pelachaud C. and Poggi I. (Eds.), *Multimodal Communication and Context in Embodied Agents*. Proceedings of the Workshop W7 at the 5th International Conference on Autonomous Agents, Montreal, Canada, 29 May 2001, pp .9-14.

Magno Caldognetto E. e Poggi I., "Dall'analisi della multimodalità quotidiana alla costruzione di agenti animati con facce parlanti ed espressive", in: Cosi P. e Magno Caldognetto E., (Eds.), *Multimodalità e multimedialità della comunicazione*, Atti delle XI Giornate di Studio del G.F.S., Padova 29-30 novembre, 1 dicembre 2000, Padova, Unipress, 2001, pp. 47-53.

Magno Caldognetto E., Zmarich C. and Cosi P., "Statistical Definition of Visual Information for Italian Vowels and Consonants", in *Proc. of AVSP '98*, D. Burnham, J. Robert-Ribes and Vatikiotis-Bateson E. (Eds.), Terrigal (Aus), 1998, pp. 135-140.

Martin, J.C., Grimard, S. and Alexandri, K. "On the Annotation of Multimodal Behaviour and Computation of Cooperation between Modalities". In Pelachaud C. and Poggi I. (Eds.), *Multimodal Communication and Context in Embodied Agents*. Proceedings of the Workshop W7 at the 5th International Conference on Autonomous Agents, Montreal, Canada, 29 May 2001, pp.1-8.

Poggi, I., "Mind Markers" in *Gestures. Meaning and Use* in Rector, M., Poggi, I. & Trigo, N. (Eds), *Gestures. Meaning and use,* Universidad Fernando Pessoa, Porto 2003, pp. 119-132.

Poggi I., "Toward the Lexicon and Alphabet of Gesture, Gaze and Touch", www.semioticon.com 2001.

Poggi, I. and Magno Caldognetto, E. "A Score for the Analysis of Gesture in Multimodal Communication", in *Proceedings of the Workshop on the Integration of Gesture in Speech Newark and Wilmington, Delaware USA,* October 7-8, 1996 (Messing, L. Ed.,) Applied Science and Engineering Labs, pp. 235-244..

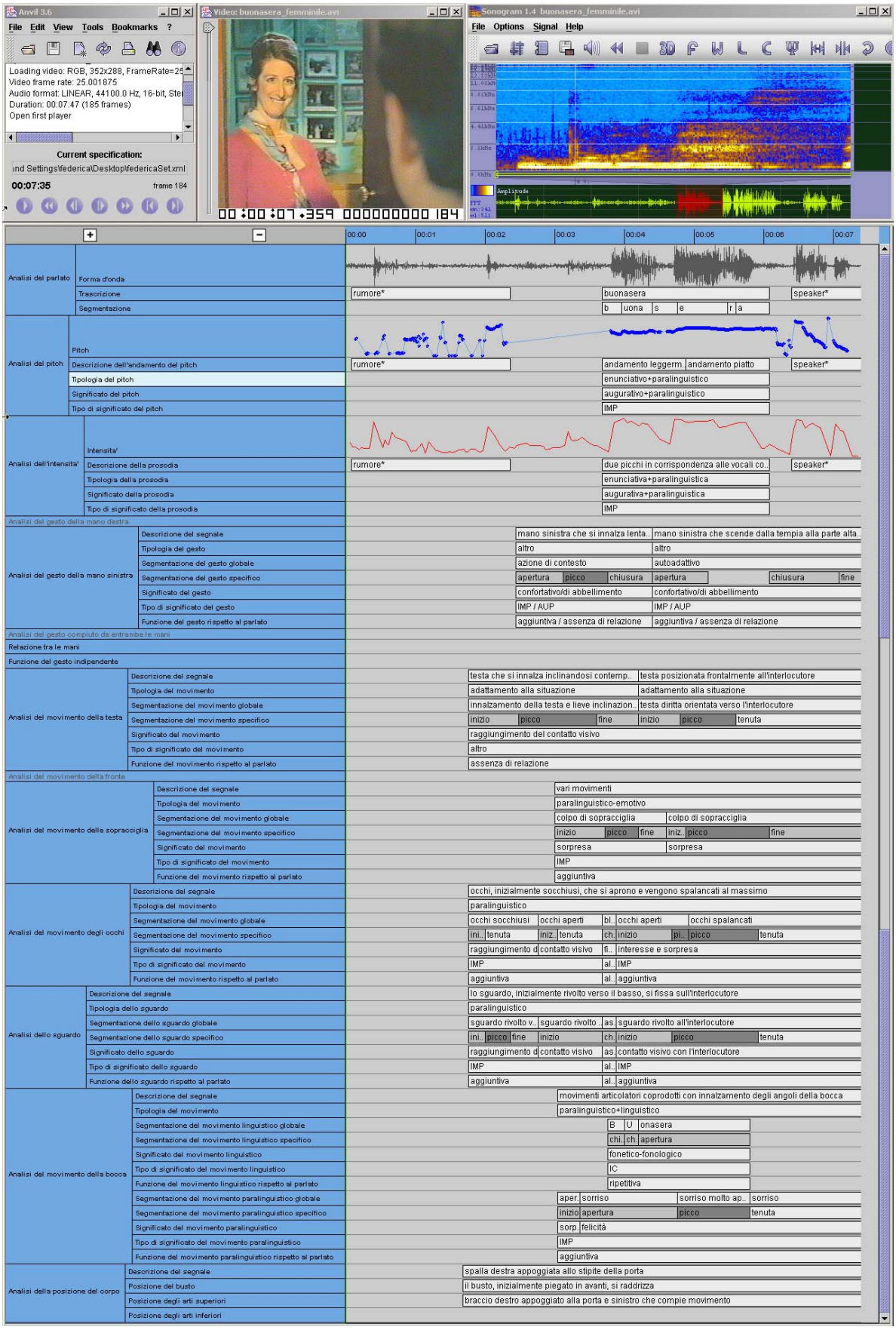
Silverman K., Beckman M., Pitrelli J., Ostendorf M., Wightman C., Price P., Pierrehumbert J. and Hirschberg J.,. "ToBi: a Standard for Labelling English Prosody". In *Proceedings of ICSLP92*, vol. 2 pp. 867-870, 1992.

Stokoe, W. C., (1980) Sign Language Structure, *Annual Review of Anthropology*, 9, pp. 365-390.

NITE Project, Annual Reports, http://nite.nis.sdu.dk/.

ISLE Project International Standards for Language Engineering, Annual Reports, http://isle.nis.sdu.dk.

Fig. 1 the Multimodal Score in Anvil display



The Swedish PF-Star Multimodal Corpora

Jonas Beskow, Loredana Cerrato, Björn Granström, David House
Magnus Nordstrand, Gunilla Svanfeldt[4]
KTH, Speech Music and Hearing, 10044 Stockholm - Sweden
+46 8 79008965
{ beskow, loce, bjorn, davidh, magnusn, gunillas }@speech.kth.se

Abstract

The aim of this paper is to present the multimodal speech corpora collected at KTH, in the framework of the European project PF-Star, and discuss some of the issues related to the analysis and implementation of human communicative and emotional visual correlates of speech in synthetic conversational agents. Two multimodal speech corpora have been collected by means of an opto-electronic system, which allows capturing the dynamics of emotional facial expressions with very high precision. The data has been evaluated through a classification test and the results show promising identification rates for the different acted emotions. These multimodal speech corpora will truly represent a valuable source to get more knowledge about how speech articulation and communicative gestures are affected by the expression of emotions.

Keywords

Multimodal corpora collection and analysis, visual correlates of emotional speech, facial animation.

Introduction

Analysis and synthesis of human-like gestures, in particular synchronisations of synthetic gestures with speech output, is achieving growing attention in the development of embodied conversational agents [9,5]. One of the greatest challenges is to implement believable, trustworthy, pleasant and human-like synthetic agents. This involves, amongst other aspects, having the agents display appropriate conversational behaviour and suitable visual correlates of expressive speech.

Analysis and visual synthesis of emotional expressions is one of the main areas of interest of the European project PF-Star [10]. The project aims at establishing future activities in the field of multisensorial and multilingual communication (Interface Technologies) by providing technological baselines, comparative evaluations, and assessment of prospects of core technologies, which future research and development efforts can build from.

One of the main activities of the first phase of the project has been the collection of audio-visual speech corpora and the definition of annotation format. These multimodal corpora are intended to provide materials for the analysis and modelling of human behaviour to be implemented in synthetic animated agents.

The animated synthetic talking heads that have been developed in our group are based on parameterised, deformable 3D facial models, controlled by rules within a text-to-speech framework. The rules generate the parameter tracks for the face from a representation of the text, taking coarticulation into account. A generalised parameterisation technique to adapt a static 3D-wireframe of a face for visual speech animation is applied [1]. This approach gives great freedom when it comes to making the synthetic faces expressive and having them perform gestures. However manual tailoring of facial gestures and emotional expressions can lead to unnaturalness of the synthesis and result in cartoon-like expressions. One way to avoid this is to obtain data that capture the dynamics of communicative and emotional facial expressions with very high precision. By capturing the facial movement of humans we can gain valuable insight into how to control the synthetic agent's facial gestures. To this end, multimodal speech corpora have been collected, and the aim of this paper is to present the different approaches of this acquisition as well as the content of the corpora and further discuss some of the issues related to the analysis and implementation of communicative and emotional visual correlates of human behaviour in synthetic conversational agents.

Data Collection

In order to be able to automatically extract relevant facial movements a motion capture procedure was employed. The data acquisition was carried out using an opto-electronic system - Qualysis MacReflex Motion Tracking System – [11] which allows capturing the dynamics of emotional facial expressions with very high precision.

Both articulatory data as well as other data related to facial movements can be recorded simultaneously, and the accuracy in the measurements is good enough for re-

[4] Authors in alphabetic order.

synthesis of an animated head (estimated mean error below 0.1 mm).

The data acquisition and processing is similar to earlier facial measurements carried out by [3,4]. Attaching infrared reflecting markers to the subject's face (as shown in figure 2) enables the system to register the 3D-coordinates for each marker at a frame-rate of 60Hz, i.e. every 17ms, by using four infrared cameras.

The utterances to be read and acted were displayed on a screen and recorded in one-minute chunks. Audio data was recorded on DAT-tape and visual data was recorded using a standard digital video camera and the optical motion tracking system.

Two corpora with two different non-professional actors have been collected with this set up:

- **corpus 1** consists of sample recordings aimed at evaluating the feasibility of different elicitation techniques such as reading prompts and interactive dialogue,
- **corpus 2** consists of non-sense words and short sentences, providing good phonetic coverage.

Corpus 1 consists of two sub-corpora, one of prompted speech and one of naturally elicited dialogues.

A total of 33 markers were used to record lip, eyebrow, cheek, chin and eyelid movements. Five markers attached to a pair of spectacles and three on the chest were used as a reference to be able to factor out head and torso movements.

The audio and visual data for the first sub-corpus was collected by having the speaker read prompted utterances, consisting of digit sequences and semantic neutral utterances, such as "*Linköpin*g" and "*ja*".

Besides the seven universal prototypes for emotions: *anger*, *fear*, *surprise*, *sadness*, *disgust*, *happiness* and *neutral* [7], we asked the subject to act *worried*, *satisfied*, *insecure*, *confident*, *questioning*, *encouraging*, *doubtful* and *confirming.* These particular expressions were chosen since, in our opinion, they might be relevant in a future spoken dialogue system scenario. Some of these expressions were previously employed in the dialogue system Adapt [6].

The second sub-corpus consists of natural dialogues which were elicited using an information-seeking scenario. This communicative scenario is similar to one that might arise between a user and an embodied conversational agent in a dialogue system: there are two dialogue participants: A, who has the role of "information seeker" and B, who has the role of "information giver". The domains of the dialogues were movie information (plots, schedules), and direction giving. The focus of the recording is on participant B, the "information giver", and only his movements were recorded (see Figure 1). However the audio recordings included the production of both subjects.

Figure 4 Data collection setup in corpus 1 with video and IR-cameras, microphone and a screen for prompts.

Corpus 2 consists of nonsense words and short sentences, providing good phonetic coverage. An actor was prompted with series of VCV, VCCV and CVC[5] nonsense words and short sentences, such as: "*grannen knackade på dörren*" (*the neighbour knocked on the door)*. The sentences were kept content-neutral in order not to affect the acted expression. The actor was asked to produce them in six different emotional states, consisting of a sub-set of the expressions used in corpus 1, that is: *confident, confirming, questioning, insecure, happy,* and *neutral*. These particular expressions were selected since they are likely to be employed in dialogue systems. Some of these expressions can be interpreted pair wise on a positive-negative scale: confident versus insecure, confirming versus questioning. We did not include sad as opposed to happy and we did not include negative expressions such as anger, fear and disgust since they might not be appropriate expressions to be employed in a dialogue system.

A total of 35 markers were used to record lip, eyebrow, cheek, chin, and eyelid movements. Five markers attached to a pair of spectacles served as reference to factor out head moments (See figure 2).

Besides the natural dialogues in corpus 1, a total of 1700 items (i.e. words and sentences) were recorded. This material will provide the data for deriving statistically based models of the articulatory movements associated with expressive speech in Swedish. Part of this corpus has been used for a cross-evaluation test with the Italian partner of the PF-Star project. The test aims at comparing emotion recognition rates for Italian and Swedish natural (actor) video sequences with those for Italian and Swedish synthetic faces [2].

[5] V= Vowel; C= Consonant

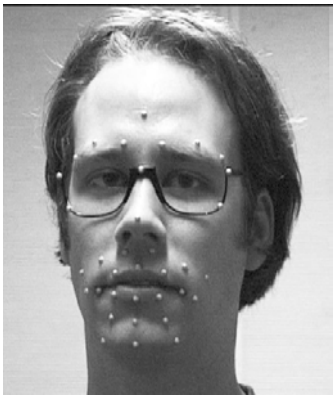


Figure 5 Test subject in corpus 2, with IR-reflecting markers glued to the face.

acted	judged							
	Expression	**hap**	**conf**	**cer**	**neu**	**unc**	**que**	**other**
	Happy	**85%**	2%	1%	1%	2%	8%	1%
	Confirming	12%	**50%**	12%	22%	1%	0%	4%
	Certain	1%	12%	**37%**	24%	3%	7%	16%
	Neutral	1%	3%	13%	**70%**	3%	2%	8%
	Uncertain	0%	3%	2%	2%	**46%**	41%	7%
	Questioning	7%	13%	15%	22%	8%	**29%**	7%

Table 1 Confusion matrix for the identification test

Data Evaluation

A test was conducted to classify the data collected in corpus 2. A group of 13 volunteer Swedish students from KTH (6 female and 7 male) was presented with a total of 90 stimuli, consisting of digitised video-sequences of the Swedish actor uttering a random selection of the sentences in corpus 2 with the six expressions. The test was run in a plenary session, the stimuli were presented using a projected image on a wide screen, in random order, without the audio. Before the experimental session the participants were instructed to look at the video files and after each video-file select one of the seven options on the answering sheet, consisting of the six expressions and an extra category for "other". The latter was inserted to avoid forced choice and a possible over-representation of neutral.
The percentages have been calculated on 78 stimuli, the first and last six stimuli responses were "dummies".
The results are shown in the confusion matrix in Table 1, where the responses for other and no response have been collapsed in one column. On average 7% of each subject's responses fell into these two categories.
All the expressions are identified above chance level, which means that the proportion of times that the subjects correctly identify the emotions is higher than the proportion of times one would expect identification by chance. No significant differences between the responses given by female and male subjects were found.
Happy and *neutral* (which are two of the basic emotions according to Ekman [7]) show much higher identification rates compared to the other expressions. *Confirming* gets 50% identification rate, and this is probably due to the fact that the actor typically produces head nods when acting this expression.
The main confusion seems to occur for *uncertain*, which has been misidentified 41% of the times as *questioning*. However, *questioning* has been misjudged as *uncertain* only 8% of the times. In fact the misjudgements for *questioning* appear to be more evenly distributed across all the other classes.
The confusion between *uncertain* and *questioning* might be due to the fact that it is not easy to discriminate between them on the basis of the visual cues only. These two expressions are quite similar in their meaning (an unsure person might appear questioning at the same time) and the actor's visual interpretation of these two expressions is similar: the typical gesture he uses is in both cases: eyebrow frowning.
Notwithstanding the confusions, and given the fact that the subjects judged the expressions on the basis of the visual cues only (i.e. without the support of the audio information), we believe that these results can be interpreted as an indication that the material collected in our corpus represent a reliable source for the analysis and measurement of different emotional facial expressions.

Data Transcription And Annotation

All nonsense words and short sentences in the two corpora are provided with a phonetic transcription, which was automatically performed by an automatic aligner [12].
For the dialogues it is necessary to perform manual transcriptions and annotation. This can be done by using a dedicated annotation tool, such as ANVIL[6]. The annotation with ANVIL is performed on a freely definable multi-layered (tracks) annotation scheme, which can be *ad hoc* defined to label non-verbal communicative and expressive behaviour. An appropriate coding scheme was created to code the visual correlates of expressive speech and their specific function in the given context [4]. Some effort was spent in transcribing, annotating and analysing human behaviour in the recorded dialogues.
Data annotation is necessary in order to couple the video-data to the 3D-data.

Exploitation Of Data

One of the goals of the analysis of the material in our multimodal corpora is to enable reproduction of trustworthy facial gestures – both emotional and other communicative gestures – in a talking face, to be used in dialogue systems. When trying to transfer the human knowledge in expressing facial gestures to a talking face, several crucial questions arise, such as what are the most

[6] http://www.dfki.de/~kipp/anvil/

appropriate and absolutely necessary expressions to implement? How can facial expressions be measured? How can we capture the complex interactions among articulatory gestures, the labial and facial visual cues related to the expression of communicative and emotional behaviour and the acoustic correlates of emotions, including prosodic features such as fundamental frequency parameters, voice quality and intonation?

Traditionally, visual and speech acoustic cues (both segmental and supra-segmental) conveying emotions have been studied separately. One of the main challenges of the PF-Star project is to understand how speech articulation and communicative gestures are coordinated.

One example is labial movements, which are controlled both by the phonetic-phonological constraints and the configurations required for the encoding of emotions. A preliminary analysis has been carried out to quantify the labial articulatory parameters modifications induced by the different emotions. The results of the investigation have shown how a number of articulatory and facial parameters for some Swedish vowels vary under the influence of expressive speech gestures [8]. Inspired by these results, we aim at building statistical models describing the interactions between articulation and emotional expression, and intend to apply these models to our talking heads.

Discussion And Future Work

The multimodal speech corpora described in this paper are very specifics and even if their dimensions are not so extensive (only two actors and relative few items recorded), they can be valuable sources to get more knowledge about how speech articulation and communicative gestures are affected by the expression of emotions.

In order to extend our corpora, we are going to carry out further data collection with the opto-electronic system. The main focus of the next acquisition will be on dialogic speech. This will give better insight in how speech articulation, facial communicative gestures and emotional expressions interact with each other in a dialogic situation and in a more spontaneous speech style than reading of prompted speech.

Further analysis will be carried out to quantify articulatory parameter modifications induced by the different emotional expressions. Moreover we will study whether certain facial emotional expressions are difficult to produce at the same time as certain communicative gestures and speech articulations. The knowledge acquired by analysing the data can be used to drive our 3D-agents, in terms of non-verbal and verbal emotional behaviour, leading, hopefully to innovative implementation in audiovisual synthesis.

Acknowledgments

Special thanks to Bertil Lyberg for making available the Qualisys Lab at Linkoping University. The PF-Star project is funded by the European Commission, proposal number: IST2001 37599. This research was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organisations.

References

1. Beskow J, 2003, *Talking heads - models and applications for Multimodal speech synthesis*. PhD thesis, TMH/KTH.
2. Beskow J, Cerrato L, Costantini E, Cosi P, Nordstrand M, Pianesi F, Prete M, Svanfeldt G, 2004, Preliminary Cross-cultural Evaluation of Expressiveness in Synthetic Faces, to appear in Proceedings of ADS 04.
3. Beskow J, Engwall O, Granström B, 2003, Re-synthesis of Facial and Intraoral Articulation from Simultaneous Measurements. *Proc. of ICPhS '03*. Barcelona, Spain, 57-60.
4. Cerrato L, Skhiri M, 2003, Analysis and measurement of communicative gestures in human dialogues, *Proc. of AVSP 2003*, St. Jorioz, France, 251-256.
5. DeCarlo D, Revilla C, Stone M, Venditti J, 2002 Making discourse visible: Coding and animating conversational facial displays *Computer Animation 02*, 11-16
6. Edlund J, Nordstrand M. 2002, Turn-taking Gestures and Hour-Glasses in a Multi-modal Dialogue System. *Proc. of ISCA Workshop Multi-Modal Dialogue in Mobile Environments*, Kloster Irsee, Germany, 181-184.
7. Ekman P, 1982, "Emotion in the human face" Cambridge University Press, New York.
8. Nordstrand M, Svanfeldt G, Granström B, and House D, (2003). Measurements of Articulatory Variation and Communicative Signals in Expressive Speech. *Proc. of AVSP'03*, 233-238.
9. Pelachaud C, Badler N, Steedman M., 1996, Generating Facial Expressions for Speech. *Cognitive Science 20*,1-46.
10. PF-STAR: http://pfstar.itc.it/ (Mars 04)
11. Qualysis: http://www.qualisys.se (Mars 04)
12. Sjölander K (2003). An HMM-based system for automatic segmentation and alignment of speech. *Proc. of Fonetik 2003* Umeå University, Department of Philosophy and Linguistics PHONUM 9, 93-96

On the primacy of language in multimodal communication

Jan Peter de Ruiter
Max Planck Institute for Psycholinguistics
Nijmegen, the Netherlands
+31 24 3521541
janpeter.deruiter@mpi.nl

Abstract

In this paper, I will argue that although the study of multimodal interaction offers exciting new prospects for Human Computer Interaction and human-human communication research, language is the primary form of communication, even in multimodal systems. I will support this claim with theoretical and empirical arguments, mainly drawn from human-human communication research, and will discuss the implications for multimodal communication research and Human-Computer Interaction.

Keywords

Multimodality, language, communication, interfaces.

Introduction

In an influential article on multimodal interaction, Oviatt [11] discusses and rejects ten common myths about multimodal interaction. The fourth myth is that *speech is the primary input mode in any multimodal system that includes it* [11, p.77]. I will defend the view that this is not a myth, but rather a deep truth, which multimodal researchers should be aware of, both in Human Computer Interaction (HCI) and in human-human communication research. In what follows, I will defend the *Linguistic Primacy Hypothesis* (LPH). I will formulate the LPH as a generalization of Oviatt's [11] formulation mentioned above, namely: "*Language is the primary input mode in any multimodal system that includes it*". By "language", I mean any modality (or to be more precise, *semiotic channel*, as defined in De Ruiter et al. [6b]) that uses a) arbitrary symbols with conventional meaning (lexical elements), and b) morphosyntactic rules that govern the combination of those lexical elements into larger utterances. In other words, speech, written or typed language, and the sign language of the deaf are all considered to be members of the category language, but for instance speech accompanying gesture is not.

I will defend the LPH by presenting a number of theoretical and empirical arguments to support it. Finally, I will discuss some of the implications of the LPH for multimodal communication research and multimodal HCI.

Arguments Against The Linguistic Primacy Hypothesis

This is not the first time in history that the truth of the LPH is questioned. In the late 1970ies, a number of communication researchers have claimed that nonverbal communication is far more important than language. For instance, Archer & Akert [1], asking their subjects to answer multiple choice questions about video fragments and transcripts, stated that "In fact, the current study provides no indication that verbal transcripts of interactions provide any independent contribution to accurate interpretation".

The claim that communication is mainly determined by nonverbal channels is analogous to the urban myth that we lose 90% of our body heat through our head. If that were actually true, one could safely go skiing naked, dressed only in a warm hat. In fact, it is only true that we lose 90% of our body heat through our head *if* we cover the rest of our body with insulating clothes. The relevance of this analogy becomes clear after realizing that [1] carefully removed verbal expressions from their materials that could have been informative, because they "did not want a simple test of audition". In other words, in their study, language did not get a fair chance.

As Brown [4] persuasively argued, it turned out to be the case in this and similar studies that nonverbal communication was predominant only *in the absence* of relevant linguistic information. When language was included, linguistic content turned out to be the best predictor of subjects' judgments of the emotional quality of the communication [8].

While the studies mentioned above focused mainly on the perceived emotional quality of the communication, more recent studies that have inspired multimodal researchers such as [10], have focused more on the *representational* aspects of communication.

It is obvious that communicating analog information such as spatial configurations can be cumbersome and inefficient in language, and that this is often done more efficiently using analog modalities such as gesture. However, for a fair comparison between language and non-linguistic modalities, it is important to also be aware of the communicative functions that language *can* perform, and the non-linguistic modalities *cannot*.

The Power Of Language

Language can encode and transmit complex information that is very hard, if not impossible, to express in non-linguistic modalities. Some illustrative examples are logical connectives, such as conditionals, and temporal information, such as past and future. Imagine having to express the following simple sentences without using some form of language:

(1) If we don't go now, we'll miss the train.
(2) Last year I finally finished my book.
(3) Although it rains, I will go for a walk.

These examples are by no means an exhaustive demonstration of the expressive powers of language. Anyone who has ever played the family game "pictionary" will realize how hard it is to express certain ideas without resorting to the use of language. A picture may be worth a thousand words, but words are priceless. Oviatt [11] is of course correct in observing that gesture or other 'analog' channels might contain information that can only be expressed in language very inefficiently; my point here is that the reverse, expressing linguistic information in a non-linguistic modality is much harder, often even impossible.

Multimodal Fusion

A strong argument for multimodal input processing is what is generally referred to as multimodal *fusion*. By combining information coming from different modalities it is possible to improve recognition quality and/or confidence. However, multimodal fusion relies fundamentally on different modalities containing redundant information. Since lip movements correlate with speech, they can in principle be used to improve speech recognition. However, many examples of multimodality in human-human communication show the use of what Engle [7] has termed *composite signals*. The information from gesture and the information from speech provide different aspects of a message. For composite signals to work properly, *both* modalities need to be reliable, and because the different components of the composite signal are by definition not correlated at the signal level, multimodal fusion will not improve their respective recognition accuracies. It is important to distinguish between fusion at the *signal* level and fusion at the *semantic* level. In the case of lip movements and speech, fusion is theoretically possible at the signal level, while in the famous "put that there" [3] example of deictic dereferencing, fusion is possible (and necessary) only at the semantic level. For semantic fusion to operate, both modalities need to have their own independent level of accuracy and confidence. In multimodal fusion, we cannot have our cake and eat it at the same time.

In fact, many existing implementations of both signal level and semantic level fusion provide evidence for the LPH because they crucially involve at least one linguistic modality. In many cases, most notably in the case of composite signals involving so-called *iconic* gestures [10], the gestures are generally not even interpretable without access to the affiliated speech [10, 5]. It appears that this often holds for other visual modalities such as facial expression as well [2].

Naturalness

Another strong and often quoted argument for multimodality is to improve the naturalness of the interaction. Just as humans use their face, eyes and hands to transmit messages to one another, machines could do so too, thereby more closely approximating face to face interaction between humans.

While this is a strong case for pursuing multimodal HCI applications, it is worth mentioning that the best way to make a multimodal interface appear *un*natural is by equipping it with slow and unreliable speech processing.

One of the main motivations to use Wizard of Oz (WoZ) technology in human factors experiments is that we suspect that leaving the speech modality to be processed by the machine will prevent us from obtaining interesting results. It is again the primacy of linguistic communication that is the reason for using WoZ procedures primarily for the linguistic modality.

Empirical Evidence From Human-human Communication

First of all, it is truly amazing what humans can accomplish by using only the linguistic modality to communicate. Not only can we satisfy virtually every communicative need by using only speech (e.g. by telephone), but even email and chat, where we don't even have access to paraverbal information such as prosody or voice quality, are highly effective in exchanging information, performing joint tasks, and maintaining social relationships.

In contrast, being in an environment in which we do not speak the language of our communicative partner will seriously hamper our communicative abilities, no matter how eloquently we gesture, draw pictures and faces, and pantomime. It is in these contexts that the lack of the previously mentioned capabilities of language become painfully obvious.

The SLOT experiment

In the COMIC project, we use an experimental paradigm called SLOT (Spatial Logistics Task, see [6b] for details). In this paradigm, two subjects are facing each other, both looking at their own copy of a "map" displayed on a graphical tablet front of them (see Figure 1).

Figure 1. Snapshot from SLOT experiment

The subjects' task is to negotiate a route through the cities on the map while trying to minimize the "cost" of that route for themselves. In order to facilitate the negotiation process, subjects can draw on the map with an electronic pen. The tablet and electronic pen in SLOT implement a "shared whiteboard" metaphor.

One of the prime motivations for the development of SLOT is that we can selectively shut down certain modalities without changing the essential characteristics of the task. We can, for instance, put a screen or a one-way mirror between the subjects to block the transmission of facial expression and eye-gaze. Also, we can enable or disable the use of the electronic pen. During the pilot phase of SLOT, we also considered blocking speech (e.g. by letting the subjects wear headphones). However, even though the subjects could then still use the pen to draw suggested routes, the negotiation process crucially depends on exchanging, attacking, and defending *arguments* (motivations) for or against proposed routes. This is fundamentally impossible without speech, unless subjects use handwriting and write letters on the whiteboard to one another, which would defeat the purpose. We were mainly interested in the composite signals created by the parallel use of speech and pen gesture.

We ran a SLOT experiment with eight dyads that could use the pen, and eight dyads that could not. The latter group therefore had no choice but to describe proposed routes through the map using speech, whereas the former could (and did) draw them directly on the map. We expected that the total negotiation times would increase significantly for the dyads that could not use the pen. To our surprise, this was not the case at all. In Figure 2, the average negotiation times for the with-pen and without-pen conditions are shown.

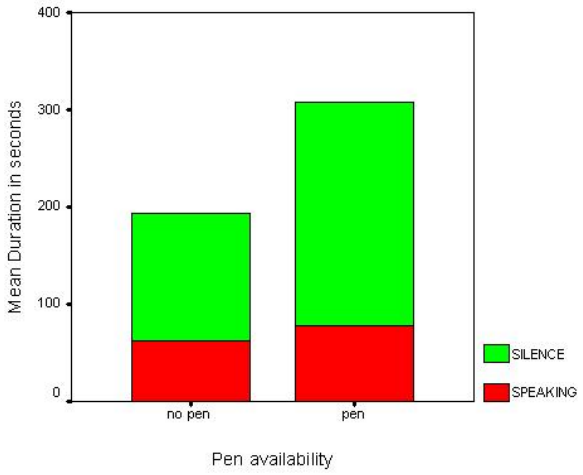


Figure 2. Average negotiation times

As can be seen from the graph, the without-pen dyads were even *faster* than the with-pen dyads, and there were no significant differences between the average speech durations. Importantly, the *quality* of the negotiated solution (measured as the sum of the incurred cost for the negotiated route for both subjects) was the same for both groups.

The main point to be made here is that the absence of the - in this context very natural and popular - pen modality did not lead to noticeable problems, neither in the efficiency of the negotiations nor in the quality of the negotiation outcome. Our subjects may have *preferred* to have used the pen, but they certainly didn't *need* it. Without speech, however, they could have drawn routes and perhaps used facial expression to display their evaluations of the routes, but they could not have *discussed* them.

Implications

So if indeed the LPH is correct, what are the consequences for multimodal communication research and HCI?

Let me emphasize that my arguments for the truth of the LPH are not in any way intended to discourage or discredit research efforts into multimodal communication, the use of multimodal fusion, or efforts to build maximally ergonomic, natural and efficient multimodal interfaces. On the contrary, I believe that an appreciation of the incredible flexibility and expressiveness of language can actually help us realize the goals of multimodal communication research.

First of all, by acknowledging the central role of language, we acknowledge the urgent need to improve language processing, especially at the input side. Speech recognition is often a serious bottleneck for the efficiency and naturalness of multimodal interfaces.

Second, as Levinson [9] has argued, speech is a very slow communication medium in terms of bits per second. The reason we can nevertheless communicate so efficiently in speech is that we can, in Levinson's words "piggyback meaning on top of meaning" [9, p.6]. In other words, not all

relevant information needs to be contained in the signal. Verbal utterances are interpreted within a cognitive context. To model this remarkable human capacity in machines, it is necessary to interpret utterances against a background of contextually relevant knowledge. To model this functionality in machines, we need to have a) detailed, implementable knowledge about the implicatures, inferences and pragmatic conventions that are used by human language users, integrated with b) symbolic representations of the contextually relevant knowledge for the domain at hand. Multidisciplinary efforts involving both linguistics and Artificial Intelligence are essential for making our interfaces truly communicative.

Most importantly, for both human-human multimodal research and for multimodal systems it is essential that we develop annotation schemes and representational frameworks that enable us to represent the meaning of both linguistic and non-linguistic signals at the same representational level (see e.g. [6a]). This is especially challenging for those signals that do not carry representational meaning but are related to socio-emotional communication.

Acknowledgments

I would like to thank Theo Vosse for his thoughtful comments on an earlier draft of this paper. Also, I am indebted to my colleagues in COMIC and in the Multimodal Interaction Project at MPI for many fruitful and inspiring discussions on multimodal communication.

References

1. Archer, D. and R.M. Akert, *Words and everything else: Verbal and nonverbal cues to social interpretation.* Journal of Personality and Social Psychology, 1977. 35: p. 443-449.
2. Bavelas, J.B. and N. Chovil, *Visible acts of meaning - an integrated message model of language in face-to-face dialogue.* Journal of Language and Social Psychology, 2000. 19: p. 163-194.
3. Bolt, R.A., *Put that there: Voice and gesture at the graphics interface.* ACM Computer Graphics, 1980. 14(3): p. 262-270.
4. Brown, R., *Social Psychology*. 2nd ed. 1986, New York: The Free Press.
5. De Ruiter, J.P., *Gesture and Speech Production*. 1998, Doctoral Dissertation: University of Nijmegen.

6a. De Ruiter, J.P., 2003. A quantitative model of *Störung*. In: Kümmel, A. & Schüttpelz, E. (eds) *Signale der Störung*. Wilhelm Fink Verlag, München.

6b. De Ruiter, J.P., et al., *SLOT; a Research Platform for Investigating Multimodal Communication.* Behavior Research Methods, Instruments and Computers, 2003. 35(3): p. 408-419.

7. Engle, R.A. *Not channels but composite signals: Speech, gesture, diagrams, and object demonstrations in explanations of mechanical devices*. in *Twentieth Annual Conference of the Cognitive Science Society*. 1998. Madison, Wisconsin.
8. Krauss, R.M., et al., *Verbal, vocal, and visible factors in judgments of another's affect.* Journal of Personality and Social Psychology, 1981. 40: p. 312-319.
9. Levinson, S.C., *Presumptive Meanings; The Theory of Generalized Conversational Implicature*. 2000, Cambridge, Massachusetts: The MIT Press.
10. McNeill, D., *Hand and Mind*. 1992, Chicago, London: The Chicago University Press.
11. Oviatt, S., *Ten myths of multimodal interaction.* Communications of the ACM, 1999. 42(11): p. 75-81.