Design Guidelines and Tools
for the Design of Multimodal Interfaces

Report for COST278, “Spoken Language Interaction for Telecommunication”

August 28, 2003
\cost278-guidelines-V1.pdf

Table of contents

1 Introduction

Now that advances in computer technology and in enabling technologies for interactive systems have led to an increase in the expressive and receptive powers of interactive systems, bringing closer the objective of *natural interaction*, designers of such systems are facing more complex choices during the design process. For instance, if information may be conveyed both through speech and by pointing at or showing on a display, what is the best choice? In essence, this is the type of question that designers of multimodal interactive systems have to answer. Ideally, the design process involves the following steps:

- Task analysis: What are the actions that need to be performed?
- Task allocation: What party is the most suitable candidate for performing particular actions?
- Modality allocation: What modality or combination of modalities is most suited to perform particular actions?

In this report, we will focus on the last question, concerning the suitability of modalities for particular actions. Obviously, a sound answer to this question needs to address the following aspects:

- Nature of the information to be exchanged
  Some kinds of information are by their intrinsic properties more suited for display on a screen whereas others lend themselves more to expression in the speech modality
- Interaction paradigm
  Command&Control-type interaction may be better tuned to the graphical domain whereas natural language dialogue may be better tuned to the speech domain
- Physical Context, Discourse Context
  The sound and illumination characteristics of particular environments may impose constraints on whether information can be conveyed through sound or on a visual display. Previous actions in the interaction may influence the choice of modality.
- Platform
  The size of a display may affect the suitability of the display for conveying particular information
- Accessibility
  The characteristics of the user may impose constraints on the way information may be offered
- Multitasking
  Often, a particular activity or task is performed along with other activities (e.g. managing the controls of the car stereo while driving. Clearly, this will affect the choice of particular modalities for conveying information.

While in the end design guidelines capturing all these aspects should emerge from research on the design of interactive systems, for the moment the emphasis is on the nature of the information to be exchanged. One might say that this is the primary question: first we address mappings between the nature of the information to be conveyed and the expressive medium, and later on we look for compensatory mechanisms if the optimal mappings cannot be applied for whatever reason.

Before proceeding, it will be convenient to define the notion of ‘modality’. Several definitions have been proposed in the literature. The simplest one links the notion of modality to the sensory channels (vision, hearing, sight, taste, smell, touch) and lets modality refer to the “type of communication channel used to convey or acquire information (Nigay & Coutaz, 1993). Often, speech is added as a separate modality, since speech poses special problems that are not well covered by the cover term hearing. Other definitions have focused more on the representational aspect, because, as was expressed above, the basic question is about the mapping and the nature of the information to be exchanged. Obviously, stating that particular numeric information should be presented in the visual modality provides little grip. Here, additional information is needed to determine whether a table or a graph is preferred. In order to distinguish the former use of modalities from the latter use, the notion of “representational modalities” has been introduced for the latter notion (Bernsen, 2001). Maybury (2001) distinguishes between media, mode and code: “media (video, audio, text) are used to capture, store, and/or transmit encoded information, modalities (vision, audition, gesture) are human sensory facilities used to interact with our environment, and codes (language, sign language, pictorial language) are systems of symbols, including their representation and associated constraints. For our current purposes, we will use modality in the more restricted sense of a sensory channel, and look for insights and guidelines concerning natural mappings between information and (combinations of) modalities that may guide design.

2 Rules of Thumb

A first step towards the formulation of guidelines concerning the application of speech in multimodal interfaces were suggestions by Michaelis and Wiggins (1982) concerning when to use speech output: speech generation is preferable when the

- message is simple.
- message is short.
- message will not be referred to later.
- messages deal with events in time.
- message requires an immediate response.
- visual channels of communication are overloaded.
- environment is too brightly lit, too poorly lit, subject to severe vibration, or otherwise unsuitable for transmission of visual information.
- user must be free to move around.
- user is subjected to high G forces or anoxia.

Tentative guidelines for when NOT to use speech may be derived from these suggestions through negation.

As can be seen, the suggestions take into consideration both properties of the messages to be transmitted and of the modality (e.g. the volatile nature of speech, its omni-directionality and the fact that it extends in time but not in space), as well as a number of additional dimensions (context, task and so forth, e.g. the illumination conditions of the visual context).

A partially overlapping, partially complementary set of suggestions is proposed by Cohen and Oviatt (1994): “A number of situations have been identified in which spoken communication with machines may be advantageous:

- when the user’s hands or eyes are busy
- when only limited keyboard and/or screen is available
- when the user is disabled
- when pronunciation is the subject matter of computer use
- when natural language interaction is preferred”

The partial complementarity stems from the fact that these suggestions also address the application of speech input.

Again, it is an enumeration of opportunities for the application of speech in the interface, without little internal structure. Cohen and Oviatt are aware of this, witnessing their comment that “As yet, there is no theory or categorization of tasks and environments that would predict, all else being equal, when voice would be a preferred modality of human-computer communication.”

Coming up with such a theory of when voice would be a preferred – or appropriate - modality of human-computer interaction has been the objective of Bernsen and co-workers. We will go into this in the next section.

3 Modality Theory

The theoretical framework developed by Bernsen c.s., known as Modality Theory, is presented in a number of papers that can be accessed from the website http://www.nis.sdu.dk/~nob/modalitytheory.html. The main sources have been included in the References.

The aim of Modality Theory has been formulated as follows:

> *Given any particular class of task domain information which needs to be exchanged between user and system during task performance, identify the set of input/output modalities which constitute an optimal solution to the representation and exchange of that information (Bernsen, 2001).*

Originally, the following ingredients were foreseen.

1. Analyses of unimodal or multimodal *output* representation in terms of well-founded taxonomies;
2. Analyses of *input* modalities and entire interactive computer interfaces in terms of well-founded taxonomies;
3. A practical methodology for applying the results of (1) and (2) to the problem of information-mapping in information systems design.

Parts 1 and 2 involved characterizing the different input and output modalities in terms of a limited number basic features such as *linguistic/nonlinguistic, analogue/non-analogue, arbitrary/nonarbitrary, static/dynamic.* Once again, it needs to be reminded that that the modalities are *representational modalities*, such as Dynamic analogue graphic language (e.g. gestural language), Analogue spoken language (everyday spoken language), Dynamic non-analogue graphic language (e.g. moving text; digital clocks), Non-analogue spoken language (e.g. spoken letters and words; list orderings), Non-analogue touch language (e.g. Braille), Diagrammatic pictures (e.g. diagrams, maps, cartoons), Non-diagrammatic pictures (e.g. photographs).

The practical methodology involved five steps:

Step 1: Identification of Information and Tasks
Step 2: Selective Task Analysis
Step 3: Information Representation
Step 4: Information-Mapping
Step 5: Trade-Offs

Step 4 would involve the application of a set of Information Mapping Rules such as the following.

If the task is T1 and user group is UGm and the goal is to optimize efficiency of interaction, then use modality or modality combination Mn.

Step 5 would involve high-level filtering to choose between multiple solutions in case Step 4 would result more than one solution.

Taking into consideration the high dimensionality of the design space and the fact that an answer to the question for optimal mappings cannot be given without taking the context into consideration, it became apparent that in fact this approach was infeasible. Instead, a more modest approach was adopted where modalities were characterized in terms of modality properties derived from the taxonomic analyses, and these modality properties would then be applied according to the following procedure:

1. Requirements Specification >
2. Modality Properties + Natural Intelligence >
3. Advice/Insight with respect to modality choice.

An overview of Modality Properties is shown in the Table below.

1	Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information. They are therefore unsuited for conveying high-specificity information including detailed information on spatial manipulation and location.
2	Linguistic input/output modalities, being unsuited for specifying detailed information on spatial manipulation, lack an adequate vocabulary for describing the manipulations.
3	Arbitrary input/output modalities impose a learning overhead which increases with the number of arbitrary items to be learned.

4	Acoustic input/output modalities are omnidirectional.
5	Acoustic input/output modalities do not require limb (including haptic) or visual activity.
6	Acoustic output modalities can be used to achieve saliency in low-acoustic environments. They degrade in proportion to competing noise levels.
7	Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction.
8	Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection.
9	Dynamic acoustic output modalities can be made interactively static (but only small-piece-by-small-piece).
10	Speech input/output modalities, being temporal (serial and transient) and non-spatial, should be presented sequentially rather than in parallel.
11	Speech input/output modalities in native or known languages have very high saliency.
12	Speech output modalities may complement graphic displays for ease of visual inspection.
13	Synthetic speech output modalities, being less intelligible than natural speech output, increase cognitive processing load.
14	Non-spontaneous speech input modalities (isolated words, connected words) are unnatural and add cognitive processing load.
15	Discourse input/output modalities have strong rhetorical potential.
16	Discourse input/output modalities are situation-dependent.
17	Spontaneous spoken labels/keywords and discourse input/ output modalities are natural for humans in the sense that they are learnt from early on (by most people and in a particular tongue and, possibly, accent). (Note that spontaneous keywords and discourse must be distinguished from designer-designed keywords and discourse which are not necessarily natural to the actual users.)
18	Notational input/output modalities impose a learning overhead which increases with the number of items to be learned.
19	Analogue graphics input/output modalities lack interpretational scope, which makes them eminently suited for conveying high-specificity information. They are therefore unsuited for conveying abstract information.
20	Direct manipulation selection input into graphic output space can be lengthy if the user is dealing with deep hierarchies, extended series of links, or the setting of a large number of parameters.
21	Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information.
22	Linguistic text and discourse input/output modalities have very high expressiveness.

23	Images have specificity and are eminently suited for representing high-specificity information on spatio-temporal objects and situations. They are therefore unsuited for conveying abstract information.
24	Text input/output modalities are basically situation-independent.
25	Speech input/output modalities, being physically realised in the acoustic medium, possess a broad range of acoustic information channels for the natural expression of information.

As can be seen, however, the modality properties are often formulated at a quite abstract level, and therefore, the threshold for applying this methodology in the design of multimodal interfaces seems quite high. Among other things, functional and information requirements are often not formulated at these levels of abstraction. For that reason, a conversion is needed into a format that is more accessible to designers. In section 4, we consider two attempts at such a conversion.

4 Design Tools

4.1 SMALTO

SMALTO has been developed by Bernsen c.s. and stands for Speech Modality AuxiLiary Tool. That is, it does not so much deal directly with the design of multimodal interfaces, but rather addresses the "speech functionality problem"), i.e., the question whether or not speech or a combination of speech and other modalities, should be used in the interface, given a number of parameters P1, .., Pn representing dimensions such as Task, User Group, Environment and so on:

> [*Combined speech input/output, speech output, or speech input modalities M1, M2 and/or M3 etc.*] or [*speech modality M1, M2 and/or M3 etc. in combination with non-speech modalities NSM1, NSM2 and/or NSM3 etc*.]
>
> are *[useful or not useful]*
>
> for *[generic task: GT]*
>
> and/or *]speech act type: SA]*
>
> and/or *[user group: UG]*
>
> and/or *[interaction mode: IM]*
>
> and/or *[work environment: WE]*
>
> and/or *[generic system: GS]*
>
> and/or *[performance parameter: PP]*
>
> and/or *[learning parameter: LP]*
>
> and/or *[cognitive property: CP]*
>
> and/or *[preferable or non-preferable]* to *[alternative modalities AM1, AM2 and/or AM3 etc.]*
>
> and/or *[useful on conditions] C1, C2 and/or C3 etc.*

SMALTO has been created by taking a large number of claims or findings from the literature on designing speech or speech-centric interfaces and casting these claims into the structured representation expressing the Speech Functionality Problem: the claims from the literature are expressed as statements that speech or a combination of speech and other modalities is useful or not given (a subset of) parameters P1 … Pn. When designing a concrete interface, the designer can inspect relevant claims to find out whether tentative information-modality mappings are supported or "approved" by findings from the literature.

SMALTO has been evaluated within the framework of projects involving the originators and in the DISC project (see Luz and Bernsen, 2001). We are not aware of reports about applications of SMALTO after this date.

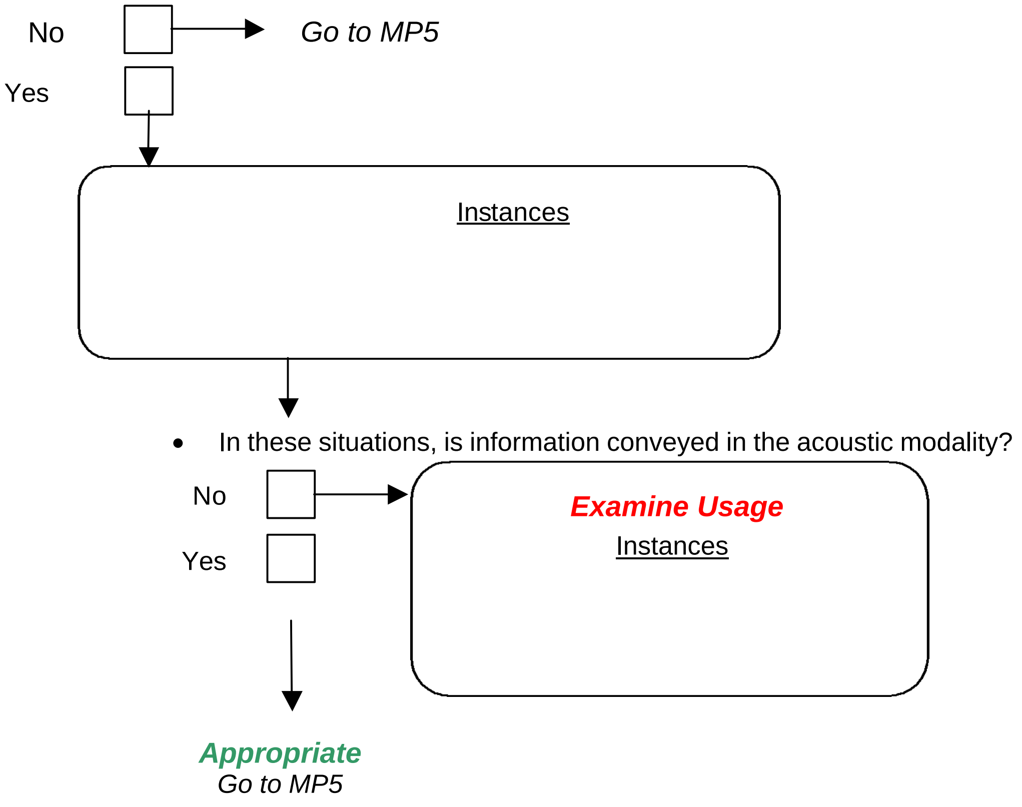
4.2 Multimodal Property Flowchart

One group who has tried to apply Bernsen's Modality theory to the design of a multimodal interface is Sony Corporate Laboratories Europe, in the framework of the Embassi project. In line with more recent views on Modality theory, the aim was not to use Modality theory to come up with information–modality mappings but rather to use modality theory as a checklist for evaluation. This might be done both for evaluating existing interfaces and for evaluating tentative mappings that were proposed in the design process.

It was noted, however, that "the form in which the modality properties are stated is not always amenable to a quick evaluation of an interface. For the purposes of a quick evaluation, [it was advocated] to re-arranging the property statements into a series of questions, which first allow it to be established if the guidelines are relevant, and then focus on the most salient aspect to which the guideline refers (e.g. looking for situations in which a large amount of information needs to be compared as opposed to the use of dynamic modalities). This allows irrelevant guidelines to be quickly overlooked, and allows aspects of the interface which are referred to by relevant guidelines to be identified and evaluated."

An example is shown below:

MP4 – Does Information need to be conveyed in a situation where direct position or gaze cannot or should not be maintained (e.g. might users be in different locations in relation to the system as it is being used)?



No □ → *Go to MP5*

Yes □

Instances

- In these situations, is information conveyed in the acoustic modality?

No □ → ***Examine Usage*** Instances

Yes □

Appropriate
Go to MP5

The results of applying the checklist in combination with guidelines for the design of Graphical User Interfaces (GUIs) to the multimodal interface that was developed in the EMBASSI project were satisfactory. On the whole, multimodality was found to have been applied adequately. At the same time, application of the checklist in combination with guidelines for GUIs resulted in the identification of potential problem areas.

5 Conclusions/Future developments

Some progress has been made going from a set of rules of thumb towards a number of guidelines that are based on a more systematic analysis of modality properties. Furthermore, the guidelines have been incorporated in tools or structured checklist. However, the tools have not been widely applied, for several reasons; seldom, they have been applied by other people than the originators themselves.

For COST278 we see three potential action lines:

1. Evaluating existing multimodal interfaces
   In order to assess the utility of the available tools, a wider range of interfaces might be evaluated against the tools, also by parties who have not been involved in their creation. It would be especially useful if independent evaluations are available. This would make it possible to see whether the tolls are able to make the correct predictions.

2. Using the tools as a guide in the design of new interfaces or new versions of interfaces
   This would enable us to assess the sufficiency and completeness of the guidelines. For instance, as Williams c.s already noticed, guidelines emerging from Modality theory had to be combined with guidelines for the design of Graphical User Interfaces. Obviously, to the extent that Action Line shows that incorrect predictions are made, this would also provide us with clues about the sufficiency and completeness of the guidelines.

3. Validation of guidelines
   A related action line might more directly focus on the validation of the guidelines. To some extent, this is already implied in Action Lines 1 and 2: testing the sufficiency and completeness is directly relevant to the validation of the guidelines, as it provides evidence that application of the guidelines does or does not lead to adequate interfaces. However, the activities under Action Lines 1 and 2 do not automatically lead to validation, so that it makes sense to distinguish a separate action line for this activity.

Disclaimer
The current overview does not claim to be complete. Other sets of guidelines have been proposed (see for instance Rudnicky) and should certainly be mentioned in subsequent versions of this document. Pointers to such guidelines are appreciated.

6 Acknowledgements

COST278

7 References

Bernsen, N.O.: Foundations of multimodal representations. A taxonomy of representational modalities. *Interacting with Computers* Vol. 6 No. 4, 1994, 347-71.

Bernsen, N. O.: Modality Theory in support of multimodal interface design. In *Proceedings of the AAAI Spring Symposium on Intelligent Multi-Media Multi-Modal Systems*. Stanford, March 1994, 37-44.

Bernsen, N.O.: Towards a tool for predicting speech functionality. *Speech Communication* 23, 3, 1997, 181-210.

Bernsen, N. O. and Dybkjær, L.: A theory of speech in multimodal systems. In Dalsgaard, P., Lee, C.-H., Heisterkamp, P. and Cole, R. (Eds.): *Proceedings of the ESCA Workshop on Interactive Dialogue in Multi-Modal Systems,* Irsee, Germany, June 1999. Bonn: European Speech Communication Association, 1999, 105-108.

Bernsen, N. O.: Multimodality in language and speech systems - from theory to design support tool. In Granström, B. (Ed.): *Multimodality in Language and Speech Systems*. Dordrecht: Kluwer Academic Publishers 2001.

Bernsen: SMALTO website: http://disc.nis.sdu.dk/smalto/

Cohen, P. R. & Oviatt, S. L. (1994) The role of voice in human-machine communication, *Voice Communication between Humans and Machines* (ed. by D. Roe and J. Wilpon), National Academy of Sciences Press, Washington, D. C., ch. 3, 34-75.

Ericsson, M., (1996) *Commenting Systems as Design Support*
ftp://ftp.ida.liu.se/pub/labs/aslab/people/miker/contrib/phlic/mikerlic.pdf (Large file - 192 pages)

Luz, S. and Bernsen, N. O. (2001) A tool for interactive advice on the use of speech in multimodal systems. *Journal of VLSI Signal Processing 29*, 129-137

Najjar, L. J., Ockerman, J. J., & Thompson, J. C. (1998). *User interface design guidelines for speech recognition applications.* Unpublished manuscript. Georgia Tech Research Institute, Multimedia Information in Mobile Environments Laboratory
http://mime1.gtri.gatech.edu/mime/papers/Speech%20recognition.htm

Maybury, M.T. (2001) *Intelligent Interfaces For Universal Access: Challenges And Promise*
http://www.mitre.org/work/tech_papers/tech_papers_01/maybury_intelligent/maybury_intelligent.pdf.

Michaelis, Paul Roller, and Wiggins, Richard H., (1982) A human factors engineer's introduction to speech synthesizers. In Badre, A. and Shneiderman, B. (Editors), *Directions in Human-Computer Interaction*, Ablex, Norwood, NJ, p. 149-178

Rudnicky (1996) *Guidelines for speech interfaces*
http://www.speech.cs.cmu.edu/air/papers/SpInGuidelines/SpInGuidelines.html **CHECK LINK!**

Williams, J., Michelitsch, G., Moehler, G., and Rapp, S. (2002) "A methodology for evaluating multimodality in a home entertainment system." *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces 2002*; Oct. 14th-16th - Pittsburgh, USA.