
Deliverable D2.10

# Working Paper on Speech Functionality

April 1999

Esprit Long-Term Research Concerted Action No. 24823

**Spoken Language Dialogue Systems and Components: Best practice in development and evaluation.**

# DISC

| | |
|---|---|
| **TITLE** | **D2.10 Working paper on dialogue management evaluation.** |
| **PROJECT** | DISC (Esprit Long-Term Research Concerted Action No. 24823) |
| **EDITORS** | - |
| **AUTHORS** | Niels Ole Bernsen and Laila Dybkjær, Natural Interactive Systems Laboratory, Odense University, Denmark nob@nis.sdu.dk, laila@nis.sdu.dk |
| **ISSUE DATE** | 22.4.1999 |
| **DOCUMENT ID** | wp2d2.10 |
| **VERSION** | 2 |
| **STATUS** | Restricted |
| **NO OF PAGES** | 74 |
| **WP NUMBER** | 2 |
| **LOCATION** | |
| **KEYWORDS** | WP2, speech functionality, design support tools, best practice, modality theory |

**Document Evolution**

| Version | Date | Status | Notes |
|---|---|---|---|
| 1 | 7/4/99 | Draft | First draft published for review by partners. |
| 2 | 22/4/99 | Final | Final version |

# Working Paper on Speech Functionality


## Abstract

Increasingly, speech input and/or speech output is being used in combination with other modalities for the representation and exchange of information with, or mediated by, computer systems. Therefore, a growing number of developers of systems and interfaces are faced with the question of whether or not to use speech input and/or speech output in multimodal combinations for the applications they are about to build. There is as yet no tool available to support developers in their decisions on whether, and how, to use speech in speech-only systems or in multimodal systems which include speech input and/or speech output.

This DISC Working Paper represents a final step towards developing such a tool. Following a previous study, the paper presents analysis a large corpus of 153 claims on speech functionality drawn from recent literature on speech functionality, i.e. on the question of what speech is good or bad for, or under which conditions to use, or not to use, speech for information representation and exchange - either speech alone or in combination with other modalities. The result of the analysis is that the applied theoretical approach adopted, which is based on Modality Theory, is suffuciently successful to warrant tools development.

# Contents

# 1. Introduction

## 1.1 Earlier work

Increasingly, speech input and/or speech output is being used in combination with other modalities for the representation and exchange of information with, or mediated by, computer systems. Therefore, a growing number of developers of systems and interfaces are faced with the question of whether or not to use speech input and/or speech output in multimodal combinations for the applications they are about to build. There is as yet no tool available to support developers in their decisions on whether, and how, to use speech in speech-only systems or in multimodal systems which include speech input and/or speech output.

This DISC Working Paper represents a final step towards developing such a tool. Following a previous study, the paper presents a large corpus of 153 claims on speech functionality drawn from recent literature on speech functionality, i.e. on the question of what speech is good or bad for, or under which conditions to use, or not to use, speech for information representation and exchange - either speech alone or in combination with other modalities. The result of the analysis is that the applied theoretical approach adopted is suffuciently successful to warrant tools development.

Prior to the work presented here (and prior to DISC), the theoretical foundations of the approach taken to speech functionality have been established. The theory in point is Modality Theory [24, 25, 26]. The potential of Modality Theory to provide support for speech functionality decisions, and the possibility for building an early design support tool for the purpose, were shown in [27]. In [27], 120 claims about speech functionality drawn from the literature up to 1993, were represented semi-formally and analysed from the point of view of the comprehensive knowledge of all possible unimodal modalities in the media of graphics, acoustics and haptics, which has been generated as part of Modality Theory. It turned out that a relatively small set of 18 modality properties were sufficient to justify of support 97% of those 109 among the 120 claims which were not either false of too vague for evaluation as to their truth value. This result was viewed as encouraging because it suggested the possibility that a key to solving the many complex speech functionality issues facing today's developers, might be knowledge about a small set of core properties of the modalities involved. Before proceeding to develop a speech functionality tool, however, it was clear that a control study had to be performed following the strict principles of scientific impartiality which were applied in the first study. The control study should target the literature on speech functionality since 1993 in order to expose the approach to data and views from recent research.

## 1.2 Methodology

This paper presents the data used in the control study as well as their global analysis. To ensure impartiality in the selection of the data, which are actually claims about speech functionality made in the literature by scientists from speech, HCI, multimodal systems development etc., one of the authors, Laila Dybkjær, selected a "brutto" data set of close to 200 speech functionality claims from about 30 papers from recent literature on the subject. In the following analytical process, Laila Dybkjær acted as quality controller of the work done by the first author. She had to accept every data handling move made by Niels Ole Bernsen. All

disagreements were resolved through, sometimes lengthy, argument until consensus was reached. This was done iteratively in three phases as follows.

In the *first phase,* Laila Dybkjær had to confirm any removal of a data point from the original set. Removals were made because of (a) redundancy (several claims turned out to be strictly identical); (b) non-removable ambiguity of a claim; (c) a claim being out of scope of the theory. (c) formed the largest single reason for claims removal. Thus, in many cases, claims were comparing speech with speech, which means that the developer would have chosen to focus uniquely speech already. These cases were removed from the data set. An example is:

> "In telephone applications with conversational dialogue capabilities, the payoff for the caller is an easier to use system but it is still one which is a long way from the power and flexibility of service offered by a human agent."

Although addressing speech functionality, this claim was considered as being out of scope because it does not address the issue of whether or not to use speech at all for the interface.

In a considerable number of other out-of-scope cases, closer inspection of the data point (or claim) and the context from which it derived, showed that speech was not involved at all. An example is:

> "[Non-speech] auditory feedback also helps users perceive changes in the interface based on their input or application events. For example, rising and falling whistling sounds accompany the appearance and disappearance of pop-up windows."

A final out-of-scope category, the "other" category (d), had one - amusing - member:

> "We found that speech when uttered in parallel with deictic gestures, often tends to break into fragments, and is in many cases incomprehensible without the information provided in the gestures."

The reason for removing this claim was that it does not represent a speech functionality problem. Rather, it illustrates the complementarity of speech and gesture: if one goes away, the other suffers, as when the TV image disappears but the sound continues. Interface developers need not worry about such issues during early design or otherwise.

The claims removal process took several iterations as some problems were only discovered in the course of the thorough second-phase analysis. In particular, redundancy among the claims took some time to remove for the obvious reason that the data set is quite large. At the end of the first phase *cum* later iterations, the present set of 153 claims drawn from 23 papers was left (Sections 4 and 6).

In the *second phase,* Niels Ole Bernsen represented each claim semi-formally following a slightly improved representation format compared to the one used in [27]. In particular, it is now possible for the reader to inspect the original claim together with its semi-formal representation. This enables readers to verify the correctness of each transformation of a claim into a semi-formal format. Secondly, "notes" ′have now been included in the semi-formal representation of most claims in order to clarify the reasons for the evaluation of a particular claim. Thirdly, each claim has been given an explicit truth value, such as 'true', 'false', 'moot'. When the truth value could not be assessed, the reason has been indicated by characterising the claim as, e.g., 'too unlimited to justify or support' or 'too vague to justify or support'. Laila Dybkjær verified the semantic equivalence of all semi-formal renderings of the 153 claims. This work included verifying, in addition, the evaluation of the claim done by reference to the modality properties derived from Modality Theory, the claims type (see Section 3), the contents of the '"notes", and the attributed truth value of the claim.

The claims evaluation by reference to Modality Theory was initially done on the basis of the 18 modality properties identified and used in [27]. It must be emphasised here that those 18 modality properties were introduced and used in [27] for the sole purpose of applying Modality Theory to the evaluation of the actual 120 claims. This means that the set of 18 modality properties has no closure in itself: it is there because its 18 individual modality properties were found to be all that Modality Theory had to bring to bear on the 120 claims. In other words, it was expected from the outset that additional modality properties would be needed for the evaluation of a new and rather different claims set, such as the set of 153 claims presented below. The important question was how many new modality properties would be needed to evaluate the new claims set (see Sections 1.3 and 2).

In the *third phase,* Niels Ole Bernsen analysed the results. An overview is shown in Section 1.3. The global analysis is presented in Section 5.

# 1.3 General results

The 153 claims represented semi-formally in Section 4 constitue a 'data mine' the analysis of which will continue for some time. Intermediate results have been published in [28] and [29]. The following brief results analysis focuses on the global results which are of crucial importance to the decision whether or not to develop a speech functionality tool.

Evaluation of a claim from the point of view of Modality Theory is done according to the following procedure. First, the claim is evaluated as to whether it is specific enough to be evaluated from the point of view of Modality Theory. This step is necessary, because some claims are simply too semantically vague for Modality Theory evaluation. An example is Data Point 1 in the present data set:

> "Speech input/output is the fastest ... means for simple exchange of information with computers."

With such claims, we don't know exactly what we are supposed to evaluate because of the lack of precision inherent to the claim. 10 claims were of this or similar nature. These are labelled 'out' below. This left 143 claims for evaluation from the point of view of Modality Theory.

| EVALUATION | COUNT |
|---|---|
| j-hit: true + justified | 102 |
| s-hit: not completely true + supported | 16 |
| c-hit: false, corrected | 2 |
| j failure: true + (only) supported | 14 |
| j/s failure: true + neither justified not supported | 9 |
| out: too vague to support or justify | 10 |
| **Total no. of data points processed:** | **153** |
| Success: j-hit + s-hit + c-hit = 102+16+2 | 120 |
| Partial success: j failure | 14 |
| Failure: j/s failure | 9 |
| **Total no. of data points excluding 'out' category** | **143** |

| | |
|---|---|
| j-hit = 102/143 | 71% |
| j-hit + s-hit + c-hit = 120/143 | 84% |
| **j-hit + s-hit + c-hit + j failure = 134/143** | **94%** |
| **Success + partial success = 134/143** | **94%** |
| Failure = 9/143 | 6% |
| **Earlier result: success + partial success** | **97%** |

**Table 1.** Overall results from evaluating 153 claims using modality properties. The final row shows the result of an earlier study [27].

Modality Theory evaluation is done by identifying modality properties which can justify a claim, support a claim which is not completely true, correct a false claim, or support (but not fully justify) a true claim. The evaluations are classified correspondingly as 'j-hits' (full justifications), 's-hits' (can be supported but not justified), 'c-hits' (false and corrected), and 'j failures' (true but only supported, not justified). The "worst" cases for Modality Theory are the claims which are true but which can be neither justified nor supported by the theory. These are labelled 'j/s failures' (justification and support failures) below.

As remarked earlier, the previous claims study [27] showed that 97% of the analysed claims were either 'j-hits' (full justifications), 's-hits' (can be supported but not justified), 'c-hits' (false and corrected), or 'j failures'. In all these cases, Modality Theory can note partial success, at least, in evaluating the claim. The corresponding results from the present study are shown in Table 1.

The difference shown in Table 1 between the 97% (success + partial success) achieved in the previous study and the 94% (success + partial success) achieved in the present study is rather small. The nine 'j/s failures' which produce the 94% have been analysed. The analysis shows that seven of these claims (see Section 4, Claims 60, 79, 98, 101, 102, 122 and 123) concern input speed which is notorious for being difficult to predict on theoretical grounds. This fact alone would seem to explain the difference between the previous study and the present one. In any case, the two figures of 97% and 94% jointly show that the use of modality properties for the evaluation of problems of speech functionality represents an approach which could contribute significantly to the resolution of speech functionality problems faced by systems developers in early design. This, again, is an encouraging basis for the decision to develop a speech functionality early design support tool incorporating the modality properties shown in Section 2.

However, in order to fully assess the result of the present study, we need a second figure. This is the number of *new* modality properties which needed to be added from Modality Theory in order to achieve the 94% result shown in Table 1. As it turned out, the new set of 153 data points forced the addition of 7 new modality properties, bringing the total number of modality properties used in the evaluation process from 18 to 25. In addition, the exact wording of several of the previously used modality properties were changed to make the properties fully general. This had not been needed in [27]. To judge these figures, one might compare the use of 18 modality properties for the evaluation of 120 data points to the use of 25 modality properties for the evaluation of 273 (120 + 153) data points. This comparison shows that the 18 modality properties constitute 15% of the number of data points evaluated, whereas the 25 modality properties are only 9% of the data points evaluated. It is, of course, impossible to tell if this apparent convergence towards zero is linear in the sense that, e.g., doubling the data set

would only need a few new modality properties to be added. Future research will tell if this is the case. However, on the indication of convergence towards zero provided by comparing the two studies we have done, we have decided to proceed with developing the speech functionality tool based on the 273 data points gathered and the 25 modelity properties used for their evaluation. Whether or not zero convergence will be achieved in the future, it seems clear that the two studies agree in demonstrating the power of using modality properties for addressing realistic problems of speech functionality during early design of speech-only systems as well as multimodal systems including speech as one of their modalities.

## 1.4 Plan

Below, Section 2 shows the modality properties used to evaluate the 153 claims addressed. Section 3 shows the typology used to categorise the data points. Section 4 shows the claims in their original wording and their semi-formal equivalents. Section 5 presents the global analysis of the data.

# 2. Modality Properties

This section shows the modality properties used in the present study (Table 2).

| NO. | MODALITY | MODALITY PROPERTY |
|---|---|---|
| **MP1** | Linguistic input/output | 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. |
| **MP2** | Linguistic input/output | Linguistic input/output modalities, being unsuited for specifying detailed information on spatial manipulation, lack an adequate vocabulary for describing the manipulations. |
| **MP3** | Arbitrary input/output | Arbitrary input/output modalities impose a learning overhead which increases with the number of arbitrary items to be learned. |
| **MP4** | Acoustic input/output | Acoustic input/output modalities are omnidirectional. |
| **MP5** | Acoustic input/output | Acoustic input/output modalities do not require limb (including haptic) or visual activity. |
| **MP6** | Acoustic output | Acoustic output modalities can be used to achieve saliency in low-acoustic environments. They degrade in proportion to competing noise levels. |
| **MP7** | Static graphics/ haptics input/output | Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction. |
| **MP8** | Dynamic input/output | Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection. |
| **MP9** | Dynamic acoustic output | Dynamic acoustic output modalities can be made interactively static (but only small-piece-by-small-piece). |
| **MP 10** | Speech input/output | Speech input/output modalities, being temporal (serial and transient) and non-spatial, should be presented sequentially rather than in parallel. |
| **MP 11** | Speech input/output | Speech input/output modalities in native or known languages have very high saliency. |
| **MP 12** | Speech output | Speech output modalities may complement graphic displays for ease of visual inspection. |
| **MP 13** | Synthetic speech output | Synthetic speech output modalities, being less intelligible than natural speech output, increase cognitive processing load. |
| **MP 14** | Non-spontaneous speech input | Non-spontaneous speech input modalities (isolated words, connected words) are unnatural and add cognitive processing load. |

| | | |
|---|---|---|
| **MP 15** | Discourse input/output | Discourse input/output modalities have strong rhetorical potential. |
| **MP 16** | Discourse input/output | Discourse input/output modalities are situation-dependent. |
| **MP 17** | Spontaneous spoken labels/-keywords and discourse input/-output | 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.) |
| **MP 18** | Notational input/output | Notational input/output modalities impose a learning overhead which increases with the number of items to be learned. |
| **MP 19** | Analogue graphics input/output | 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. |
| **MP 20** | Haptic manipu-lation selection input | 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. |
| **MP 21** | Haptic deixis (pointing) input | Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information. |
| **MP 22** | Linguistic text and discourse input/output | Linguistic text and discourse input/output modalities have very high expressiveness. |
| **MP 23** | Images input/output | 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. |
| **MP 24** | Text input/output | Text input/output modalities are basically situation-independent. |
| **MP 25** | Speech input/output | 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. |

**Table 2.** The 25 modality properties used fror claims evaluation in the present study.

# 3. Claim Types

This section shows the types used to classify claims in the present study. Compared to the claims types used in [27], the only addition is type T14. T14 had to be added because of the large number of claims on multimodal combinations involving speech in the data set. It will be noted that the types (T1-T6 + T8-T13) form a closed set.

**T1**: Claims recommending combined speech input/output.

**T2**: Claims positively comparing combined speech input/output to other modalities.

**T3**: Claims recommending speech output.

**T4**: Claims positively comparing speech output to other modalities.

**T5**: Claims recommending speech input.

**T6**: Claims positively comparing speech input to other modalities.

**T7**: Conditional claims on the use of speech.

**T8**: Recommendations against the use of combined speech input/output.

**T9**: Claims negatively comparing combined speech input/output to other modalities.

**T10**: Recommendations against the use of speech output.

**T11**: Claims negatively comparing speech output to other modalities.

**T12**: Recommendations against the use of speech input.

**T13**: Claims negatively comparing speech input to other modalities.

**T14**: Recommendations of speech in combination with other modalities.

# 4. Claims

The 153 claims that were semi-formally represented and analysed are presented in this section. They have been simply organised according to the papers from which they were collected. The papers are presented in the same numbered order in the list of references. In the square bracketed references, the first numeral refers to the paper, the second refers to the page.

## Article 1

**1.** Speech input/output is the fastest ... means for simple exchange of information with computers. [1, 10]

> Data point 1. **Generic task** [simple exchange of information with computers]: speech input/output is **performance parameter** [fastest]. No justification. Claims type: **T2**
>
> **NOTE:** This claim is too general and too vague to justify or support. As a generalisation it is most certainly false. Mouse clicking or function keys, for instance, can get a lot done quickly. The phrase "simple exchange of information" has no clear meaning. Most people can speak, but most people can also read and use a pointing/selection device for exchanging information with computer systems. Input and/or output speed, although admittedly important to efficiency, is a highly device-, task- and user skill-dependent notion which is difficult or impossible to generalise. This is why Modality Theory has little to say about it except sometimes by implication.
>
> **Too vague to justify or support.**

**2.** Speech input/output is the ... easiest means for simple exchange of information with computers. [1, 10]

> Data point 2. **Generic task** [simple exchange of information with computers]: speech input/output is **performance parameter** [easiest]. No justification. Claims type: **T2**
>
> **NOTE:** This claim is too general and too vague to justify or support. As a generalisation it is most certainly false. Mouse clicking or function keys, for instance, can get a lot done easily. The phrase "simple exchange of information" has no clear meaning. Most people can speak, but most people can also read, move and use a pointing/selection device. MP17, for instance, would both serve to support and criticise the claim: speech is natural, which makes it easy to use for most users, but spoken keywords which users have to remember, can be difficult to use. The present claim is the sort of claim about speech which should be replaced by more precise and well-founded knowledge.
>
> **Too vague to justify or support.**

**3.** Speech output is [slower and more difficult compared] to other means in conveying complex information. A variety of information can be displayed at once by images and text. [1, 10]

> Data point 3. **Generic task** [conveying complex information]**:** speech output is **performance parameters** [slower and more difficult] than graphics (combined images and text) output. Justified by MP1: "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." And MP19: "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." Claims type: **T11**
>
> **NOTE:** "Complex information" is a woolly term which may mean, e.g. highly abstract information as well as high-specificity information, such as that found in a photograph (a static graphic image). What the present claim really says, then, is that a combination of analogue graphics and any natural language modality, such as speech, must be superior in expressiveness to speech-only.
>
> **True.**

**4.** Input by speech (recognition) and output by [static] images and text is considered to be an ideal combination for most interactive systems with computers. [1, 10]

> Data point 4. **Generic system** [most interactive systems]: speech input combined with static graphics (images and text) output is considered **performance parameter** [ideal]. Supported by MP1: "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." And MP19: "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." Claims type: **T14**
>
> **NOTE:** The claim is unjustifiable because nobody will be able to count the systems (and their relative import etc.) which support the claim vs. the systems which counter the claim. MPs 1 and 19 go a long way towards explaining why there are systems which support the claim, and MP1 goes some way towards explaining why there are systems which counter the claim.
>
> **Too unlimited to justify.**

**5.** Displaying recognition results is very useful to enable users to detect recognition errors and to reduce redundancy in conversation with computers. [1, 10-11]

Data point 5. Combined speech input and static graphics (text) output showing recognition results is **performance parameters** [very useful to detect recognition errors and reduce redundancy in conversation with computers]. Supported by MP7: ”Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction.” Claims type: **Rsc.**

**NOTE:** Despite the (modest) support from MP7, it seems fair to say that, at the moment, there are mixed feelings about text displays of recognition results. The display distracts the user from the task and spoken feedback in some form might be just as useful or better.

**Moot.**

# Article 2

**6.** Rather than simply invoking commands a step at a time, users will specify to the computer the overall goals of a task and delegate to it responsibility for working out the details. The most natural and convenient way to interact with such systems will be by means of a natural spoken dialogue. [2, 109]

Data point 6. **Generic system** [personal intelligent assistant] + **generic task** [specifying overall task goals]**:** natural speech input is **performance parameter** [more convenient] + **cognitive property** [more natural] than other input modalities. Justified by MP1**:** “Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information.“ And MP5: “Acoustic input/output modalities do not require limb (including haptic) or visual activity.” And MP17: ”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).” Claims type: **T6**

**NOTE:** It is hard to justify a comparison between one modality and all other modalities. Arguably, this has been done in the present case: MP1 focuses on linguistic modalities, whilst MP5 and MP17 point out the senso-motoric ease and naturalness of using speech input.

**True.**

**7.** Users will likely be unwilling to speak to the computer in restricted command languages, so conversational interaction will only become popular when the assistant can understand a broad range of English paraphrases of the user’s intent. [2, 109]

Data point 7. **Generic system** [personal intelligent agent] + **generic task** [specifying overall task goals]**:** natural speech input rather than restricted command language will be needed for **cognitive property** [popularity]. Justified by MP17: ”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.)" And MP18: "Notational input/output modalities impose a learning overhead which increases with the number of items to be learned." Claims type: **T7**

**NOTE:** Designer-designed keywords and discourse are, in a sense, worse than technical notation because the user feels that these sub-sections of free-style natural language should be easy, but definitely aren't. This is why we are often annoyed of not being able to find the items we want in phone books.

**True.**

**8.** It is precisely the flexibility ... of spoken language that makes it such an attractive interface. [2, 110]

Data point 8. Speech discourse input/output, being **performance parameter** [flexible], is **cognitive property** [attractive] as interface. Justified by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP15: "Discourse input/output modalities have strong rhetorical potential." And MP16: "Discourse input/output modalities are situation-dependent." Claims type: **T1**

**NOTE:** The somewhat woolly term "flexibility" can be interpreted as having "high expressiveness" relative to other modalities. Arguably, MPs 1, 15 and 16 express the cash contents of claim N8. Note that MPs 15 and 16 do not apply to written text.

**True.**

**9.** Attempts to define specialised English subsets as command languages can be frustrating for users who discover that natural (to them, if not the designer) paraphrases of their requests cannot be understood. [2, 110]

Data point 9. Designer-designed speech input command languages can **cognitive property** [cause frustration]. Justified by MP17: "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.)" And MP18: "Notational input/output modalities impose a learning overhead which increases with the number of items to be learned." Claims type: **T12**

**True.**

# Article 3

**10.** Speech input is potentially more direct [than mouse clicking]. [3, 165]

Data point 10. **Generic task** [navigating hypermedia information in graphic output web space]**:** speech input is **performance parameter** [potentially more direct] than haptic (mouse clicking) input. Justified by MP1**:** "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP20: "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." Claims type: **T6**

**NOTE:** The directness attributed to speech input is due to the fact that, potentially, linguistic input can get the user directly to the target information.

**True.**

**11.** Speech input ... may be regarded as more natural [than mouse clicking] by some users. [3, 165]

Data point 11. **Generic task** [navigating hypermedia information in graphic output web space]: speech input may **cognitive property** [be regarded as more natural] than haptic (mouse clicking) input by **user group** [some users]. Justified by MP17: "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.)" And MP20: "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." Claims type: **T6**

**NOTE:** The modesty of the claim ("may be" ... "some users").

**True.**

**12.** Natural language processing allows the user to access relevant information immediately, rather than having to navigate through a hierarchy of WWW pages. [3, 165]

Data point 12. **Generic task** [navigating hypermedia information in graphic output web space]**:** natural language input provides **performance parameter** [immediate access to relevant information] whereas haptic (mouse clicking) input does not. Justified by MP1**:** "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP20: "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." Claims type: **T6**

**True.**

**13.** Speech is also a useful additional output medium, to introduce a large amount of text. [3, 165]

Data point 13. **Generic task** [web interaction: being introduced to a large amount of text]: speech output is a useful additional modality to static graphic text output. Justified by MP12: "Speech output modalities may complement graphic displays for ease of visual inspection." Claims type: **T14**

**True.**

**14.** Speech is also a useful additional output medium, to draw attention to key information. [3, 165]

Data point 14. **Generic task** [web interaction: highlighting key information]: speech output is a useful additional modality to static graphics output. Justified by MP12: "Speech output modalities may complement graphic displays for ease of visual inspection." Claims type: **T14**

**True.**

**15.** We do not see speech as a replacement for graphical user interfaces. Instead, we believe that the best solution is an integrated interface allowing the different input and output modes to complement and enhance each other. [3, 165]

Data point 15. Speech input/output is believed to complement and enhance, rather than replace, graphical user interfaces. Justified by MP1: "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." And MP7: "Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction." And MP19: "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." Claims type: **T14**

**True.**

**16.** Some users may be uncomfortable with the idea of talking to a computer ... In such circumstances, text input is an attractive alternative. [3, 165]

Data point 16. Speech input may be **cognitive property** [less comforting] than haptic (text) input for **user group** [some users]. Supported by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." Claims type: **T13**

**NOTE:** The weak epistemic modifier "may be" and the weak quantifier "some (users)". Such claims are very likely to be true but extremely hard to justify on principled grounds. Some users will always be hesitant wrt. new technology. The support concerns the basic similarities between speech and text. If the issue is important, user attitudes may need empirical investigation.

True.

**17.** The system may have difficulty understanding [users] if they have a strong accent. In such circumstances, text input is an attractive alternative. [3, 165]

Data point 17. If strongly accented, speech input may be **performance parameter** [more difficult to make the system understand] than haptic (text) input. Supported by MP17: "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)." Claims type: **T13**

**NOTE:** For this claim to have been fully justified, an additional MP might have pointed to the well-known fact that current systems have difficulty understanding strong input accents and dialects. Alternatively, it might have been pointed out that speech naturalness does not carry from one tongue to another. People with strong accents may or may not be good at spelling, though.

**True.**

# Article 4

**18.** It has been hypothesized that deictic gestures will have a favorable influence upon spoken language interpreting systems because they will reduce the speech recognition workload. (Background for hypothesis: In person to person speech, deictic gestures eliminate the need for a lengthy definite description and simplify the dialogues.) [4, 281]

Data point 18. Speech input combined with haptic (deictic gesture) input, may reduce the system's speech recognition workload because deictic input gesture eliminates the need for lengthy definite description and simplifies the dialogue. Justified by MP1: "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." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T14**

**True.**

# Article 5

**19.** Hands-free operation is a unique property of voice-control. Speech input allows hands- and eyes-free operation which is very important in hands- or eyes-busy situations, e.g. while driving a car. [5, 1453]

Data point 19. **Generic task** [hands- or eyes-busy situations, e.g. driving a car]**:** speech input is **performance parameter** [unique] as it allows hands- and eyes-free operation. Justified by MP5: "Acoustic input/output modalities do not require limb (including haptic) or visual activity." Claims type: **T5**

**NOTE:** This claim is equivalent to the MP which justifies the claim. This is the the best that can be hoped for.

**True.**

**20.** Remote control operation is a unique property of voice-control. Speech input can be used for remote control, e.g. via telephone, in order to control a system which is out of manual reach. [5, 1453]

Data point 20. **Generic task** [remote system control] + **interaction mode** [telephone and similar devices]: speech input is **performance parameter** [uniquely suited]. No justification. Claims type: **T5**

**NOTE:** It is not obvious why speech input is uniquely suited for the purpose by contrast with, e.g., a graphical user interface through which remote control is being done, unless use of a telephone or similar device is mandatory. In that case, the claim would seem trivially true, i.e. not to need justification. If false, the claim's falsehood is at the device level and therefore not within the scope of modality properties.

**False.**

**21.** Direct access operation is a unique property of voice-control. Voice control often avoids the translation of a function into a code. Consider name dialling as an example: in order to place a telephone call the name of the person to be called is just spoken instead of translating the name into a code (the telephone number) and keying in that code. [5, 1453]

Data point 21. **Generic task** [name dialling] + **interaction mode** [telephone]: speech input avoids the translation of a person name into haptic (telephone keyboard) input code (the telephone number). Justified by MP3: "Arbitrary input/output modalities impose a learning overhead which increases with the number of arbitrary items to be learned." And MP17: "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).” Claims type: **T6**

**True.**

# Article 6

**22.** Speech is a natural and easy-to-use modality for humans. [6, 671]

Data point 22. Speech input/output is **performance parameter** [easy-to-use] and **cognitive property** [natural] for **user group** [humans]. Supported by MP17, Corrected by MP17: ”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.)” Claims type: **T1**

**NOTE:** The corrections to the present claim by MP17 is that MP17 contains the qualifier “by most [but not all] people” and the warning about designer-designed keywords and discourse.

**Partly true.**

**23.** Human-computer interaction with the combination modes of GUI and speech interface is expected to be the largest potential area of speech recognition applications. [6, 672]

Data point 23. **Generic system** [speech recognition applications]: speech input/output combined with multimodal combination (graphical user interfaces) is expected to be the largest potential application area for speech recognition. Supported by MP1: “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.” And MP19: ”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.” Claims type: **T14**

**NOTE:** Nobody can fully justify such complex quantitative comparisons. The support provided is not particularly strong.

**Too unlimited to justify.**

# Article 7

**24.** The acceptance by the public of automated telephone information services and other applications is likely to be limited until more natural spoken dialogues are possible. [7, 1947]

Data point 24. **Generic systems** [automated telephone information services]: **cognitive property** [broad public acceptance] is likely to require more natural speech input/output discourse. Supported by MP14: "Non-spontaneous speech input modalities (isolated words, connected words) are unnatural and add cognitive processing load." And MP15: "Discourse input/output modalities have strong rhetorical potential." And MP16: "Discourse input/output modalities are situation-dependent." And MP17: "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.)" Claims type: **T7**

**NOTE:** Still, the claim may be legitimately doubted. It is possible to design natural discourse for a certain task even though the discourse is only a tiny fragment of the language needed for natural conversation.

**Too strong.**

# Article 8

**25.** In the opinion of the bank we partnered with, the overall automation rate achieved by the speech interfaces in Money Talks is competitive with—but not clearly superior to—automation rates they have achieved with their Touch-Tone interface. It must be emphasized that this is a comparison for a particular calling population, and compares a long-familiar interface against a new one. [8, 1941]

Data point 25. **Generic system** [financial information]: speech input/output interfaces have **business parameter** [automation rate] comparable to combined telephone key input + speech output for **user group** [familiar with touch-tone but not with spoken dialogue systems]. Supported by MP17: "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)." Claims type: **Claims equalling speech with other modalities.**

**NOTE:** Assuming that the Money Talks interfaces have state-of-the-art usability, it seems likely that the tasks addressed are ones for which touch-tone interfaces are suitable and speech input/output interfaces do not add significant value. These are tasks where the touch-tone system does not overload the listener. Modality Theory supports the use of both systems but does not imply that they are equally suitable. That would require the mentioned task analysis and would seem too fine-grained for inclusion in a Modality Property.

**True.**

# Article 9

**26.** ... noise (which will degrade speech input and output). [9, 1671]

Data point 26. Speech input/output is degraded by **work environment** [noise]. Justified by MP6: ”Acoustic output modalities can be used to achieve saliency in low-acoustic environments. They degrade in proportion to competing noise levels.” Claims type: **T8**

**True.**

# Article 10

**27.** The [Wizard of Oz] study generated behavioral, timing, and error data for UI design (e.g., integrated use of speech and touch; use of visual displays such as maps and browsers; and changing patterns of interaction with experience). These experiments indicated ... that vocal output resulted in slowed performance, as users waited until the end before continuing the transaction. [10, 1673]

Data point 27. **Generic system** [simulated train information kiosk combining speech input/output, haptic touch-screen input, graphics output showing complex train information]: speech output produced **performance parameter** [slowed performance as users waited until the end before continuing the transaction]. Justified by MP8: ”Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection.” And MP7: ”Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction.” Claims type: **T10**

*Assumption1:* the messages the users waited to hear until the end were relatively insignificant messages about what they could see on the screen.

*Assumption2:* the system was completely new to the users.

**NOTE:** Under the stated assumptions, what happened was that essentially redundant spoken output was listened to rather than being ignored by the naive users. Arguably, the graphics gave the users all the feedback they needed to their queries and the spoken feedback only slowed them down.

**True.**

# Article 11

**28.** Speech recognition technology is necessary to automate services where the number of service options is large. For example, a restaurant selector service that asks callers which cuisine they would like would be manageable as a speech automated service (“What kind of cuisine would you like?”) but unwieldy as a Touch-Tone service (“For Chinese food, press 11; for Italian food, press 12 ...”) [11, 1681]

Data point 28. **Generic task** [large number of service options, e.g. restaurant cuisine options]: speech input/output is **performance parameter** [manageable] whereas menu style touch-tone interaction, i.e. haptic (telephone keys) input/speech output, is not. Justified by MP8: ”Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection.” Claims type: **T6**

**NOTE:** Because users cannot freely inspect the offered cuisine options, they may have to wait till the end of the output speech before making their selection. This makes it likely that they will tend to forget some of the options announced as well as the corresponding digits. With input speech, on the other hand, they can say what they want from the outset – if they know what they want. The justification, incidentally, implies that the output task might be done by static graphics (text possibly supplemented with images for illustration), cf. MP7.

**True.**

**29.** The circumstances that are most conducive for speech recognition automation are cases where Touch-Tone is not in place ... [11, 1682]

Data point 29. **Generic system** [telephone services]: consider using speech input when menu style touch-tone interaction, i.e. haptic (telephone keys) input + speech output, is **circumstantial parameter** [not in place]. Supported by MP17: ”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).” Claims type: **T6**

*Assumption:* the claim only deals with tasks for which touch-tone is adequate.

**NOTE:** Touch-tone systems are useful. It is not known whether they have come to stay. Speech input/output systems can do everything that touch-tone systems can do, and more. Modality Theory focuses of single modalities and therefore does not include does not include combined modality properties, such as “telephone key input/speech output is unnatural”. Similarly, Modality Theory does not include complex comparative claims, such as “for some tasks, touch tone and speech input/output” are equally useful. Including such combinations and comparisons would lead to an unlimited number of modality propserties. The claim appears true because, if touch-tone is in place and works, why bother to replace it by speech input/output?

**True.**

**30.** The circumstances that are most conducive for speech recognition automation are cases where Touch-Tone is ... not extremely popular. [11, 1682]

Data point 30. **Generic system** [telephone services]: consider using speech input when menu style touch-tone interaction, i.e. haptic (telephone keys) input + speech output, is **cognitive property** [not popular]. Justified by MP8: ”Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection.” Claims type: **T6**

**NOTE:** MP8 points to the core problem with touch-tone interaction for certain tasks, i.e. the core reason why touch-tone interaction is not popular. Users tend to forget some of the options announced as well as the corresponding digits to press.

**True.**

# Article 12

**31.** Many auditory interfaces are based on hierarchical models. For example, interfaces for voice mail allow the user to navigate through a hierarchy of choices for listening to and deleting messages. Hierarchical models are used because they can abstractly represent groups. It is also relatively easy to navigate these auditory interfaces using ... voice input, although the requisite path from one object to another may be lengthy. [12, 22]

Data point 31. **Generic task** [navigating limited-size hierarchical auditory output interfaces, e.g. for voice mail] + **user group** [the blind]: consider using speech input for **performance parameter** [relative ease]. Justified by MP5: “Acoustic input/output modalities do not require limb (including haptic) or visual activity.” And MP17: ”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).” Claims type: **T5**

**True.**

**32.** The inherent disadvantage of all auditory interfaces is they are largely invisible. [12, 24]

Data point 32. Auditory output interfaces, e.g. for voice mail, being largely invisible, require exposition of their contents for **performance parameter** [usability]. Justified by MP5: “Acoustic input/output modalities do not require limb (including haptic) or visual activity.” And MP8: ”Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection.” Claims type: **T10**

**NOTE:** Dynamic acoustics cannot give us a static menu which can tell us about the options we have.

**True.**

**33.** Numerous strategies for conveying objects in auditory interfaces have already been suggested by previous work. Possible strategies include using speech ... For example, an auditory cue to convey a text-entry field could be a synthesized voice saying "text-entry". [12, 25]

Data point 33. **Generic task** [conveying objects in auditory output interfaces, e.g. text-entry fields]: consider using speech output keywords. Justified by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP5: "Acoustic input/output modalities do not require limb (including haptic) or visual activity." And MP17: "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.)" Claims type: **T3**

**NOTE:** What needs to be justified only is that speech output keywords is a possible solution, not that this solution is particularly good or without problems.

**True.**

**34.** This design is based on the premise that auditory icons [i.e. auditory images of something] offer the most promise for producing discriminable, intuitive mappings. In the previous example the sound of an old-fashioned typewriter maps easily to a text-entry field. ... Two alternate design strategies that were considered and discarded were using speech or earcons [i.e. little tunes]. Synthesized speech is required for presenting textual information in the graphical interface. This information (e.g. the text in an electronic mail message or the labels on a pull-down menu) is domain dependent. By relegating speech to domain-dependent information and respectively relegating nonspeech cues to domain-independent information, the user can more easily separate these classes of information. [12, 25]

Data point 34. **Generic task** [conveying, in auditory output interfaces, text-entry fields and the text entered into them]: prefer speech output for entered text (domain-dependent) and some non-speech acoustic modality for the text-entry field (domain-independent): for **performance parameter** [ease of separation] of information types. Supported by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP11: "Speech input/output modalities in native or known languages have very high saliency." Claims type: **T14**

**NOTE:** The point made in this claim is that speech output should not be used for both of the purposes mentioned, and, being necessary for rendering the entered text, therefore should not be used for rendering the nature of the text-entry field

itself. This is clever and plausible design proposal and the problem addressed is an important one. The claim is impossible to verify at this point given the potentially large number of alternatives.

**Too unlimited to justify. Plausible.**

**35.** From our model of the graphical interface, we know there are many characteristics of the interface objects that need to be conveyed to the user. The use of auditory icons often serves to convey the affordances of the object as well. … But there are other attributes of objects we need to convey, such as its label, whether it is grayed out, and its relative size. Text-based attributes can be presented via synthesized speech. For example, the auditory icon [i.e. auditory image of something] for a push button can be presented simultaneously with its text label. [12, 26]

Data point 35. **Generic system** [auditory output interfaces]: speech output can be used to label auditory images. Supported by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP11: "Speech input/output modalities in native or known languages have very high saliency." Claims type: **T14**

**NOTE:** This is close to justification. The missing piece is an emplicit statement to the effect that speech modalities have higher saliency than auditory images and hence can co-exist with them. Including such comparative claims would probably lead to an explosion in the number of Modality Properties.

**True.**

**36.** One limitation of auditory interfaces is the difficulty in presenting an overview of the interface contents. [12, 30]

Data point 36. **Generic task** [presenting overview of output interface contents]: auditory output interfaces have **performance parameter** [difficulty of presentation]. Justified by MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." Claims type: **T10**

**True.**

**37.** Sighted computer users could also benefit from auditory representations of graphical interfaces while performing eyes-busy tasks such as driving, performing maintenance of an airplane, or inspecting a manufacturing plant. However, the needs of these users are different. For instance, supporting mobility is a key requirement. In these cases, improving the flexibility of conversational interfaces may provide the most promise. [12, 43]

Data point 37. **Generic task** [eyes-busy, mobility/away from the desk, e.g. driving, airplane maintenance, manufacturing plant inspection]: conversational speech input/output may be **performance parameter** [the more promising]

replacement of multimodal combination (graphical user interfaces). Justified by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP5: "Acoustic input/output modalities do not require limb (including haptic) or visual activity." And MP17: "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)." Claims type: **T2**

*Assumption 1:* Work is being done in (relatively) low-noise environments (cf. MP6).

*Assumption 2:* The tasks do not involve specifying detailed information on spatial manipulation and location (cf. MP1).

**NOTE:** Given the stated assumptions, this seems to be close enough to constitute a justification (rather than support).

**True.**

# Article 13

**38.** What is spoken has to either be retained in the listener's internal memory or it is lost. A listener can retain the gist of an utterance, the surface structure being lost. This is usually acceptable for everyday conversation and listening to plain text in synthetic speech. However, as algebra notation is concise and lacks redundancy, loss of any of this information can be catastrophic. [13, 55]

Data point 38. **Generic task** [accessing spoken output algebraic notation]: speech output is inferior to static graphic (text) output with respect to **cognitive property** [grasping and processing of the information represented]. Justified by MP7: "Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction." And MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." Claims type: **T11**

**True.**

**39.** The listener does not control which part of the [algebraic] information is to be heard. When there is control (e.g., with a tape recorder), the control is so slow and inaccurate that it is almost useless. This lack of control makes the listener passive and this passivity often leads to lapses of concentration, which leads to greater need for control over the flow of information. Short-term memory is easily overloaded leading to loss of information, increased

mental workload, and a lack of cognitive resources to be focused on the comprehension task itself. [13, 55]

Data point 39. **Generic task** [accessing spoken output algebraic notation]: lack of **performance parameter** [adequate control] of output speech produces **cognitive properties** [passive listening, lapse of concentration, need for control, short-term memory overload, loss of information, increased mental workload, lack of comprehension resources for the task]. Justified by MP8: ”Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection.” And MP9: ”Dynamic acoustic output modalities can be made interactively static (but only small-piece-by-small-piece).” Claims type: **T10**

**NOTE:** How this justification relies on our intuitive understanding of the syntactic complexity of algebraic notation.

**True.**

**40.** To avoid indicating deeper mathematical meaning in speech is easy. To only display grouping of symbols, the manner of grouping—without indicating some of the syntactic meaning of that grouping—presents some problems. The symbols $2x^2$ may be spoken in a variety of ways (e.g., as “two x squared” or “two x superscript two”). These renderings span a range of added meaning. ... However, the English language often lacks neutral words for constructs and use of unfamiliar words will have a concomitant effect on usability. Thus compromise is sometimes needed so that the best form of presentation is used. [13, 57]

Data point 40. **Generic task** [avoiding to convey syntactic meaning of grouping among algebraic symbols]: speech output is inferior to static graphic (text) output. Supported by MP25: ”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.” Claims type: **T11**

**NOTE:** This claim may seem surprising. One might have thought that, universally, the avoidance of ambiguity is an asset of a particular representation. However, in this case the authors want to encourage “active reading” of algebra and argue that active reading requires the reader to impose an interpretation on otherwise ambiguous symbols. In this case, the expressiveness of speech is counter-productive! Modality Theory has no property to the effect that a static graphic algebraic representation preserves syntactic ambiguity. To include such a claim would probably lead to an explosion in the number of Modality Properties.

**True.**

**41.** A verbose stream of relentless speech is more likely to overwhelm a listener when it lacks any prosodic cues. [13, 59]

Data point 41. Verbose, lengthy speech output lacking prosodic cues is likely to be **cognitive property** [overwhelming]. Justified by MP13: "Synthetic speech output modalities, being less intelligible than natural speech output, increase cognitive processing load." Claims type: **T7**

**True.**

**42.** For an auditory system an extra facility has to be added that is not needed in a visual system. This is the notion of current. In a visual display the current selection is indicated by some means in a persistent fashion (i.e., by highlighting in reverse video). The auditory display is transient, so the current selection on focus of attention also disappears. So the action current is needed to give this focus. [13, 64]

Data point 42. **Generic task** [replacing static graphics output] for **user group** [the visually disabled]: speech output needs to simulate the notion of what is in current focus. Justified by MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." Claims type: **T11**

**NOTE:** Due to its transience, speech lacks a variety of means of expression which are available in graphics, such as highlighting, which is even available in dynamic graphics. One may, for instance, "highlight" a word by stressing it, but it does not stay stressed because of the transience of speech. By contrast, a highlighted word does stay highlighted even in dynamic graphics for as long as it remains visible to the user.

**True.**

**43.** Combining a move or action with an object or target forms a command that falls naturally into a spoken form. For example "beginning of expression", "next term" and "previous character" emerge easily from the set of actions and targets as intuitive commands. [13, 65]

Data point 43. **Generic task** [navigating spoken output algebra] + **user group** [the visually disabled]: some spoken input commands (e.g. "next term") do **performance parameter** [emerge easily and intuitively]. Corrected by MP17: "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.)" Claims type: **T5**

**NOTE:** Do not expect *any* designer-designed keywords to emerge naturally or intuitively with users.

**False.**

**44.** A glance gives information about overall structure and complexity and can give the reader expectations about the expression. In addition the reader may review the expression for any unknown or difficult symbols. Such a glance is usually not available to a blind reader. With a spoken presentation it is not possible to take an abstract or high-level view, and reading is usually reduced to a bottom-up process of integrating a series of symbols that have been heard in a temporal, "left-to-right" manner. [13, 72]

Data point 44. **Generic task** [perceiving at-a-glance structure and complexity of algebra expressions]: is **performance parameter** [not possible] with spoken output. Justified by MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." And MP9: "Dynamic acoustic output modalities can be made interactively static (but only small-piece-by-small-piece)." Claims type: **T10**

**True.**

**45.** Speeded-up speech, which retains structural cues such as division into terms and the grouping of objects into complex items, fulfills some of the criteria for a glance. The only information lacking is the type of the object being represented. [13, 73]

Data point 45. **Generic task** [perceiving at-a-glance structure and complexity of algebra expressions] + **user group** [the visually disabled]: can (only) be partly simulated by speeded-up and temporally structured output speech. Supported by MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." Claims type: **T11**

**NOTE:** MP8 simply points out that speech, being dynamic, has the dynamic character that characterises a glance. MP8 cannot justify the temporal structure - part of the claim. However, Modality Theory actually *is* capable of justifying this claim because it provides a complete list of the information channels offered by different modalities. Comparing the information channels offered by two different modalities enables identification of how far those of one modality can be mapped into those of the other modality. For instance, the spatial grouping offered by static 2D graphics and used in algebra texts can be mapped into temporal grouping among spoken expressions. However, this part of Modality Theory is not easily expressible in simple and relatively coarse-grained Modality Properties. Information channels are referred to, and only in a global fashion, in MP25.

**True.**

**46.** Omission errors formed the largest category of errors. There are memory limits to how many objects or groups of objects listeners can maintain after hearing them. [13, 79-80]

Data point 46. **Generic task** [listing many objects or groups of objects]: speech output affords **cognitive property** [limited retainability]. Justified by MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." Claims type: **T10**

**NOTE**: It may be argued that MP8 does not justify the part of the claim which states that "Omission errors formed the largest category of errors." However, MP8 does justify that many omission errors are to be expected, and that might be sufficient information for the developer when considering whether to use spoken output.

**True**

**47.** A glance or overview of the structure of [complex algebra] information can be provided by combining the prosodic features of the spoken output, the hiding of complexity, and the use of earcons to convey information. [13, 89]

Data point 47. **Generic task** [perceiving at-a-glance the structure and complexity of algebra expressions]: use combination of spoken output and sound images. Supported by MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." Claims type: **T14**

**NOTE:** Rather weak support is provided by MP8 which points out that speech, being dynamic, has the dynamic character that characterises a glance. To provide a full justification, analysis would have to go into the information channels characterising speech and other acoustic modalities, such as earcons. However, this part of Modality Theory is not easily expressible in simple and relatively coarse-grained Modality Properties. Information channels are referred to, and only in a global fashion, in MP25.

**True.**

# Article 14

**48.** Interfaces involving spoken ... input could be particularly effective for interacting with dynamic map systems, largely because these technologies support the mobility [walking, driving et.] that is required by users during navigational tasks. [14, 95]

Data point 48. **Generic task** [mobile interaction with dynamic maps, e.g. whilst walking or driving]: a speech input interface component could be **performance parameter** [particularly effective]. Justified by MP5: "Acoustic input/output modalities do not require limb (including haptic) or visual activity." Claims type: **T14**

**NOTE:** The careful wording of the claim "Interfaces involving spoken ... input". It is not being claimed that speech could suffice for the task, only that speech might be a useful interface ingredient. Otherwise, the claim would be susceptible to criticism from, e.g., MP1. Note also that the so-called "dynamic maps" are static graphic maps which are interactively dynamic.

**True.**

**49.** Speech allows the hands and eyes to be busy, which is particularly valuable when users are in motion. [14, 95]

Data point 49. For **performance parameter** [user mobility, such as walking or driving]: consider speech input which is hands-free and eyes-free. Justified by MP5: "Acoustic input/output modalities do not require limb (including haptic) or visual activity." Claims type: **T5**

**NOTE:** This is an example of a claim which comes with a "built-in" justification which is equivalent to a Modality Property. In addition, the claim points out a particular implication of MP5, i.e. the usefulness to user mobility. One might wish that more claims had this character!

**True.**

**50.** Speech allows the hands and eyes to be busy, which is particularly valuable when users are ... in natural field settings. [14, 95]

Data point 50. **Work environment** [natural field settings]: consider speech input which is hands-free and eyes-free. Justified by MP5: "Acoustic input/output modalities do not require limb (including haptic) or visual activity." Claims type: **T5**

**True.**

**51.** It [speech input] also offers speed. [14, 95]

Data point 51. Speech input offers **performance parameter** [speed]. No justification. Claims type: **T5**

**NOTE:** This claim is simply too vague to evaluate. The speed offered by speech input is highly relative to the task.

**To vague to justify or support.**

**52.** It [speech input] also offers high-bandwidth information. [14, 95]

Data point 52. Speech input offers high-bandwidth information. Justified by MP22: "Linguistic text and discourse input/output modalities have very high expressiveness." Claims type: **T5**

**NOTE:** The more precise statement MP22. Some linguistic modalities, such as keywords or notation, are in most cases much less expressive than unconstrained text or discourse.

**True.**

**53.** It [speech input] also offers relative ease of use. [14, 95]

> Data point 53. Speech input offers **performance parameter** [relative ease of use]. Supported by MP17, Corrected by MP17: "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.)". Claims type: **T5**
>
> **NOTE:** The more precise statement MP17. Some speech input modalities, such as designer-designed keywords, do *not* offer relative ease of use.
>
> **Partly true.**

**54.** It is known that users tend to prefer speech for describing objects and events, sets and subsets of objects, out-of-view objects, and past and future temporal states, as well as for issuing commands for actions or iterative actions. [14, 95]

> Data point 54. **Speech acts** [describing objects and events, sets and subsets of objects, out-of-view objects, past and future temporal states, issuing commands for actions or iterative actions]: speech input has **cognitive property** [preferred]. Justified by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP17: "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)." Claims type: **T5**
>
> **True.**

**55.** As human language technologies, spoken ... input have the advantage of permitting users to engage in more natural information-seeking dialogues with map systems. [14, 95]

> Data point 55. **Generic task** [information-seeking dialogues with map systems]: speech input is **cognitive property** [natural]. Justified by MP17: "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)." Claims type: **T5**
>
> **True.**

**56.** Compared with other input modes, [spoken and pen-based input] can more easily support spontaneous and flexible description of map objects, events, and spatial layouts, as well as their interrelation. [14, 95]

Data point 56. **Generic task** [description of map objects, events, and spatial layouts, as well as their interrelation]: speech input combined with haptic (pen-based) input can more easily support **performance parameter** [spontaneous and flexible description] than other input modalities. Supported by MP1: "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." And MP17: "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)." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T14**

**NOTE:** Nobody can tell if the claim is true because that would require comparison with all other possible modality combinations. Still, the claim is a plausible one that deserves to be made.

**Too unlimited to justify.**

**57.** Together, spoken and pen-based input provide complementary capabilities, which can function as a set of power tools for managing complex information. [14, 95]

Data point 57. **Generic task** [managing complex information, e.g. digital roadmaps]: speech input combined with haptic (pen-based) input are powerful and complementary. Justified by MP22: "Linguistic text and discourse input/output modalities have very high expressiveness." And MP1: "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." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T14**

**NOTE:** The power and complementarity of the mentioned input modality combination is fully laid out by the quoted Modality Properties.

**True.**

**58.** Compared with linguistic modes of interaction such as speech, pen, or keyboard, direct manipulation queries are relatively inflexible in their coverage and are unable to support spontaneous description. In contrast, a language-based map system would permit users to automatically locate out-of-view entities through simple description of landmarks and streets, as in "Where are the for-sale homes farthest from Yuba Faul?". Direct manipulation can be cumbersome or even infeasible for supporting a function like automatic map location and, when possible, the resulting "manual queries" tend to be less efficient than linguistic ones.

This is especially true when information is densely calibrated, as in time or cost, because in these cases users must make fine-grained slider manipulations and their ability to pick a precise value may not be possible given the system's available granularity. [14, 99]

Data point 58. **Generic task** [locate out-of-view roadmap entities]: spontaneous linguistic input is **performance parameter** [more flexible] than haptic (direct slider manipulation) queries. Justified by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP20: "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." Claims type: **T6**

**True.**

**59.** In comparison with manual input modes, spoken input is a relatively imprecise and inefficient mode for indicating locations. As a means of specifying lines, these disadvantages of speech input are compounded further. For tracing irregular routes or outlining spatial areas, spoken input can be infeasible altogether. [14, 100]

Data point 59. Compared to haptic (deixis/pointing) input, speech input is **performance parameters** [relatively imprecise and inefficient, or worse] for indicating locations, specifying lines, tracing irregular routes or outlining spatial areas. Justified by MP1: "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." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T13**

*Assumption:* the "manual input modes" mean direct manipulation.

**True.**

**60.** As expected, unimodal writing [input] was significantly slower than ... speech input. [14, 116]

Data point 60. **Generic task** [interaction with digital roadmaps]: speech input was **performance parameter** [significantly faster] than haptic (hand-written) input. No justification. Claims type: **T6**

**NOTE:** Input and/or output speed, although admittedly important to efficiency, is a highly device-, task- and user skill-dependent notion which is difficult or impossible to generalise. This is why Modality Theory has little to say about it except sometimes by implication. Empirical study may be needed.

**True.**

**61.** This data clarifies that multimodal input [pen + speech input] is [not] faster than speech in the ... non-spatial [task domains]. Analyses of both the verbal and quantitative simulation domains revealed no significant difference in task completion time between spoken and multimodal input. [14, 116]

Data point 61. **Generic task** [non-spatial, verbal and quantitative, e.g. conference registration]: combined speech input and haptic (pen-based) input was **performance parameter** [not faster] than speech-only input. Justified by MP1: "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." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T6**

**True.**

**62.** People reported that they liked using speech and pen for different functions ... Participants most frequently reported preferring the pen for indicating locations, scrolling, adding drawn objects to the map, ... requesting distance calculations between objects [16, 118] [and] designation of points, lines, and areas. [14, 124]

Data point 62. **Generic task** [indicating locations, scrolling, adding drawn objects, requesting distance calculations between objects, designating points, lines, and areas in interaction with digital roadmaps]: haptic (pen-based) input was **cognitive property** [preferred] to speech input. Justified by MP1: "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." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T13**

**True.**

**63.** They generally preferred speech for functions like requesting automatic location of out-of-view landmarks, labeling map-based content, and issuing descriptive commands (e.g., to specify a real estate selection constraint, such as "No houses in a flood zone"). [16, 118] ... they preferred speech for describing objects and sets of objects [14, 124]

Data point 63. **Generic task** [requesting automatic location of out-of-view landmarks, labeling map-based content, and issuing descriptive commands in interaction with digital roadmaps]: speech input was **cognitive parameter** [preferred] to haptic (pen-based) input. Justified by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T6**

**True.**

**64.** In large part, the error prone, slow, disfluent, and generally unacceptable nature of speech-only input to maps can be traced directly to people's difficulty articulating spatially oriented descriptions. [14, 120]

Data point 64. **Generic task** [interaction with digital roadmaps]: speech input is **performance parameters** [error prone, slow, disfluent, unacceptable] because of **cognitive property** [difficulty articulating spatially oriented descriptions]. Justified by MP2: "Linguistic input/output modalities, being unsuited for specifying detailed information on spatial manipulation, lack an adequate vocabulary for describing the manipulations.". Claims type: **T12**

**True.**

**65.** In brief, the performance advantages of multimodal [pen + speech input] over speech-only map interaction include: shorter and less complex constructions. [14, 120]

Data point 65. **Generic task** [interaction with digital roadmaps]: speech input produces **performance parameter** [longer and more complex linguistic constructions] than speech input combined with haptic (pen-based) input. Justified by MP1: "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." And MP2: "Linguistic input/output modalities, being unsuited for specifying detailed information on spatial manipulation, lack an adequate vocabulary for describing the manipulations." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T14**

**True.**

**66.** In brief, the performance advantages of multimodal [pen + speech input] over speech-only map interaction include: 10% faster task completion. [14, 120]

Data point 66. **Generic task** [interaction with digital roadmaps]: combined speech input and haptic (pen-based) input produces **performance parameter** [10% faster task completion] than speech-only input. Justified by MP1: "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." And MP2: "Linguistic input/output modalities, being unsuited for specifying detailed information on spatial manipulation, lack an adequate vocabulary for describing the manipulations." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T14**

**NOTE:** The 10% figure cannot be justified, of course, but the exact figure hardly matters to the developer. What matters is that there is a significant difference in task performance speed.

**True.**

**67.** In brief, the performance advantages of multimodal [pen + speech input] over speech-only map interaction include: 36% fewer task-critical content errors. [14, 120]

Data point 67. **Generic task** [interaction with digital roadmaps]: combined speech input and haptic (pen-based) input produces **performance parameter** [36% fewer task-critical content errors] than speech-only input. Justified by MP1: "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." And MP2: "Linguistic input/output modalities, being unsuited for specifying detailed information on spatial manipulation, lack an adequate vocabulary for describing the manipulations." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T14**

**NOTE:** The 36% figure cannot be justified, of course, but the exact figure hardly matters to the developer. What matters is that there is a significant difference in task performance error.

**True.**

**68.** In brief, the performance advantages of multimodal [pen + speech input] over speech-only map interaction include: 50% fewer spontaneous disfluencies. [14, 120]

Data point 68. **Generic task** [interaction with digital roadmaps]: combined speech input and haptic (pen-based) input produces **performance parameter** [50% fewer spontaneous disfluencies] than speech-only input. Justified by MP1: "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." And MP2: "Linguistic input/output modalities, being unsuited for specifying detailed information on spatial manipulation, lack an adequate vocabulary for describing the manipulations." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T14**

**NOTE:** The 50% figure cannot be justified, of course, but the exact figure hardly matters to the developer. What matters is that there is a significant difference wrt. spontaneous disfluencies.

**True.**

**69.** In brief, the performance advantages of multimodal [pen + speech input] over speech-only map interaction include: 95-100% preference for multimodal interaction. [14, 120]

Data point 69. **Generic task** [interaction with digital roadmaps]: speech input is **cognitive property** [almost never preferred] to speech input combined with haptic (pen-based) input. Justified by MP1: "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." And MP2: "Linguistic input/output modalities, being unsuited for specifying detailed information on spatial manipulation, lack an adequate vocabulary for describing the manipulations." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T14**

**True.**

**70.** This preference [for pen + speech input] clearly was most pronounced when manipulating complex visual-spatial displays. [14, 124]

Data point 70. **Generic task** [manipulating complex visual-spatial displays]: combined speech input and haptic (pen-based) input was **cognitive property** [clearly preferred] to speech-only input. Justified by MP1: "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." And MP2: "Linguistic input/output modalities, being unsuited for specifying detailed information on spatial manipulation, lack an adequate vocabulary for describing the manipulations." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T14**

**True.**

# Article 15

**71.** Form interfaces tend to be more like appliances, whereas query languages tend to be more like composable tools. Speech input might be used to support either, depending on the level of granularity of the actions that can be interpreted. One could imagine a speech act that is the equivalent of "make me some bread" or a lengthy series of utterances that details every step of a recipe. [15, 137]

> Data point 71. **Generic tasks** [form filling, querying in a query language]: speech input might be used. Justified by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." Claims type: **T5**
>
> **NOTE:** This claim is a modest one and easy to justify.
>
> **True.**

**72.** However, in databases with many object types and relations among them, it is unclear whether users would be able to formulate queries using natural language expressions that can be matched accurately to underlying objects. [15, 145]

> Data point 72. **Generic task** [querying in a query language] + **generic system** [complex relational database]: unclear if **user group** [ordinary users] can use speech input to **performance parameter** [accurately match to underlying objects]. Justified by MP18: "Notational input/output modalities impose a learning overhead which increases with the number of items to be learned." Claims type: **T12**
>
> *Assumption 1:* natural language does not match unambiguously into the formal (notational) query language assumed by the database organisation.
>
> *Assumption 2:* the claim is about ordinary users.
>
> *Assumption 3:* ordinary users do not master the formal (notational) query language assumed by the database organisation.
>
> **NOTE:** Three assumptions have been added. In general, claims often rest on unstated assumptions. These are not always made explicit in notes to claims but this has been done, illustratively, so to speak, in this case.
>
> **True.**

**73.** [For querying of a complex relational database], combinations of speech-based referring expressions and menus that remind users of available attributes might be most effective. [15, 145]

> Data point 73. **Generic task** [querying in a query language] + **generic system** [complex relational database]: speech input producing **speech act** [referring expressions] combined with static graphic (menus of attributes) output might be **performance parameter** [most effective]. Justified by MP18: "Notational input/output modalities impose a learning overhead which increases with the number of items to be learned." And MP7: "Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction." Claims type: **T14**

**NOTE:** The static graphics output keywords will help the ordinary user keep within the vocabulary and domain which the system can handle. That this actually is the most effective approach is not completely evident, however, but depends of unstated interface details and various comparisons with other solutions.

**True.**

**74.** Set description in Visage is performed in multiple ways. The first is drill-down: selecting a relation from a menu along which to navigate from one object to a set of other objects. Often, drill-down must occur in multiple steps across multiple relations (e.g., from a military unit to its subordinate units to the warehouse where the latter get their supplies to the crews that manage the warehouse, etc.). Under these circumstances, other forms of input are likely to complement drill-down and be more efficient, for example, a simple spoken request, such as "what crews support this division?". Systems that support queries like these would be very powerful but, as Cohen and Oviatt (1995) pointed out, require significant interpretation and robustness with respect to the numerous ways people are likely to refer to the same relations. [15, 154]

Data point 74. **Generic task** [drill-down to select a relation from a static graphic keyword menu along which to navigate from one object to a set of other objects]: speech input is likely to complement haptic (mouse selection drill-down) input and be **performance parameter** [more efficient]. Justified by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP20: "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." Claims type: **T14**

**NOTE:** As the last sentence points out, such multimodel systems are non-trivial to build.

**True.**

**75.** In addition to drill-down, set description in Visage is also supported using dynamic query sliders and related painting techniques. Both these techniques enable one to define sets by specifying ranges for quantitative attributes. ... Although Visage supports creation of multiple visualizations as well as dynamic query sliders, there are substantial operations if the goal is merely to specify a single expression. Spoken descriptions of the attribute ranges would be much more efficient (e.g., "select units that have more than 30 jeeps and more than 100 people"). [15, 155]

Data point 75. **Generic task** [specifying attribute ranges for atomic queries]: speech input is **performance parameter** [much more efficient] than haptic (mouse dynamic query sliders and painting techniques) input. Justified by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP20: "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." Claims type: **T6**

**NOTE:** The claim proposes a combination of static graphic output information as used in an existing system, and spoken input "shortcuts".

**True.**

**76.** On the other hand, both techniques [dynamic query sliders and related painting techniques] serve other important functions not easily performed by speech inputs. Painting multidimensional charts [painting can be used to select elements within one frame] or other frame types enables one to define sets by enumeration [One may e.g. "ask" to select units that have more than 30 jeeps and more than 100 people] —especially when a pattern of elements is used to define the set. Dynamic query sliders provide continuous control of quantitative variables with immediate feedback, thus enabling selection of subsets based on patterns that emerge because of the animation. The important point here is that combining speech and these techniques provides great potential for supporting different extremes of the set creation dimension. Indeed, sets defined using either speech or direct manipulation can be further refined by the other. [15, 155]

Data point 76. **Generic task** [specifying attribute ranges for atomic queries]: combination of speech input and haptic (mouse dynamic query sliders and painting techniques) input has **performance parameter** [great potential]. Justified by MP1: "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." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T14**

**True.**

**77.** It may still be possible to eliminate some of the burden of navigating through menus and cascading drill-down operations if spoken requests remain close to database relations. An example is a speech request that enumerates the relations to be traversed: "Drill-down to subordinates, to warehouses, to crews ... ." This sequence can be much more quickly conveyed than with menu navigation and has the advantage that the end point of the path can be displayed without having to view all the intermediate steps when these are not relevant. Providing drill-down menus together with speech traversal provides the opportunity for users to learn sufficient database structure to make the transition to functional speech requests and still provides help when needed. [15, 158]

Data point 77. **Generic task** [querying] + **generic system** [complex hierarchical inventory database]: combination of hierarchical static graphics (text menus) output and speech input traversal of these may help **user group** [ordinary users] learn to do the queries **performance parameter** [faster] by eventually using

speech input only. Justified by MP7: ”Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction.” And MP17: “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.)” And MP20: ”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.” Claims type: **T14**

**NOTE:** The idea here is that the static output graphics keywords can help users learn enough to eventually do the task by speech-only.

**True.**

**78.** Speech is useful for retrieving objects by name or referring expression, especially when objects are not currently visible to the user. [17, 169] In particular, we proposed the use of speech to augment interaction with visualizations to: express queries that refer to object sets intensionally or by name, especially when they are not visible or would otherwise require numerous navigation operations. [15, 181]

Data point 78. **Generic task** [retrieval of graphics objects that are not currently visible or whose retrieval require numerous navigation operations]: speech input is **performance parameter** [useful] for retrieval by **speech act** [name or referring expression]. Justified by MP1: “Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information.” And MP20: ”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.” Claims type: **T14**

**NOTE:** That MP1 deals with linguistic expressions generally and not just speech. However, it follows that (also) speech is useful for the purpose and, if you can speak it, why type it?

**True.**

**79.** SageBook visualizations can be assigned names that must be typed in a dialogue box and perused using a folder-style metaphor. Although an interface like the Macintosh Finder would be helpful for requesting visualizations by name, speech would be more efficient. [15, 169]

Data point 79. **Generic task** [graphics object retrieval]: speech input is **performance parameter** [more efficient] than haptic (typed) input for retrieving not-currently-visible objects by **speech act** [naming]. No justification. Claims type: **T6**

**NOTE:** When the user knows the name of a particular graphics object, it would no doubt be faster for most users to speak the name than to type it. However, input and/or output speed, although admittedly important to efficiency, is a highly device-, task- and user skill-dependent notion which is difficult or impossible to generalise. This is why Modality Theory has little to say about it except sometimes by implication. Empirical study may be needed.

**True.**

**80.** Speech also enables referring to graphical properties without having to refer to particular objects. For example, to request graphics that use color or size in SageBrush, one must select a grapheme first (e.g., a line), then select its color or thickness property. However, this limits the retrieval set to visualizations where line thickness is used. In contrast, spoken referring expressions can be made without restricting them to object types (e.g., "find visualization that use color or size"). Likewise, it is easier to refer to specific graphical or data values using speech (e.g., "find visualizations that use red and blue," "find visualizations with circles and diamonds," "find the visualization showing data for 1990 to 1995"). Finally, speech would support composing queries or requests that combine properties of both data and graphics (e.g., "find the red and blue charts showing interest rates by year," "find charts with vertical bars and dates along the x-axis"). [15, 169-170]

Data point 80. **Generic task** [data and graphics visualisations retrieval by graphical properties and/or data values]: speech input **performance parameter** [makes it easier] than haptic (mouse selection) input to retrieve objects. Justified by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information. "And MP20: "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." Claims type: **T6**

**True.**

**81.** Many of the [design] operations [i.e. those involving discrete expressions] currently performed using direct manipulation might also be performed using speech. [15, 170]

Data point 81. **Generic task** [data and graphics object retrieval using discrete expressions]: speech input can do the operations currently done by haptic (mouse selection) input. Justified by MP1: "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." Claims type: **T5**

**NOTE:** The claim contrasts discrete input expressions and continuous input information. The continuous input information is exactly the spatial information for which speech is unsuited according to MP1.

**True.**

**82.** Speech inputs, coupled with direct manipulation pointing, would provide a broader set of operations than simple graphical property changes. For example, one could change the bar chart on the right to a table by circling it and saying "Change this to a table." One could remove and add data attributes in a similar way (e.g., "remove the circles," "remove Vehicle Type from the bars," "change bar color to show Cargo Weight"). More complex changes are also worth exploring, such as "reorganize the shipment chart so Vehicle Type is along the y-axis and Team is shown in color." ... multimodal interfaces have great potential for supporting both ends of the continuum, providing fine-grain, composable design operations and higher level abstract expressions. [15, 170-171]

Data point 82. **Generic task** [complex changes to data graphics, e.g. bar chart change to table]: combined speech input and haptic (mouse pointing gesture) input have **performance parameter** [great potential for broadening the scope of possible operations]. Justified by MP1: "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." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T14**

**True.**

**83.** Virtually all of the design operations involve discrete expressions and can therefore be performed using either direct manipulation or spoken inputs. The main exception to this is the spatial arrangement of graphemes in clusters. One can specify the relative positions of circles, bars, and other graphemes. Chart spaces can be aligned spatially as well. The specification of spatial arrangements like these requires continuous controls in the same sense that the scale and pan of frames in Visage, discussed previously. [15, 171]

Data point 83. **Generic task** [spatial arrangement of graphemes in clusters, e.g. circles and bars]: speech input is **performance parameter** [inferior] to haptic (mouse direct manipulation) input. Justified by MP1: "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." And MP2: "Linguistic input/output modalities, being unsuited for specifying detailed information on spatial manipulation, lack an adequate vocabulary for describing the manipulations." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T13**

**True.**

**84.** One impediment to the use of speech is the lack of familiar terminology for referring to all the design elements of visualizations. Although users can easily refer to colors, bars, and

lines, our experience is that they lack terms for referring to horizontal and vertical axes, the difference between simple and interval bars, ..., gauges, nodes and links in networks, and other objects. [15, 172]

Data point 84. **Generic task** [complex design operations on data graphics, e.g. bar charts]: **user group** [ordinary users] lack speech input terminology for referring to many elements of visualizations (e.g. horizontal and vertical axes, nodes and links in networks). No justification. Claims type: **T12**

**NOTE:** Arguably, this claim is outside the scope of Modality Theory altogether, as it concerns the quite general fact that ordinary users cannot be expected to master specialist domains and their terminologies. Users having, e.g., a speech interface for this application are likely to acquire the terminology needed provided that they get appropriate graphics output in menus etc. Empirical study may be needed.

**True.**

**85.** Nonetheless, the lack of a common spoken vocabulary [for referring to horizontal and vertical axes, the difference between simple and interval bars, ..., gauges, nodes and links in networks, and other objects] is likely to be mitigated by multimodal interfaces. Users can point to elements of visualizations and request components to be removed and copied in new ones. [15, 172]

Data point 85. **Generic task** [complex design operations on data graphics, e.g. bar charts]: combined speech input and haptic (mouse pointing gesture) input into graphic output space is likely to **performance parameter** [mitigate] the problem that **user group** [ordinary users] **cognitive property** [lack speech input terminology for referring to many elements of visualizations (e.g. horizontal and vertical axes, nodes and links in networks)]. Justified by MP1: "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." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T14**

**True.**

**86.** Direct manipulation techniques like the ones described here are critical for supporting dynamic continuous changes to object appearance. However, there is also great potential for composite actions to be created from the primitive SDM (Selective Dynamic Manipulation) operations, but initiated by spoken commands. Spoken commands would express the intent of multiple composed primitive actions (e.g., "make the red objects more visible," "shrink everything but the red objects," "compare the red and black objects"). [15, 179]

Data point 86. **Generic task** [dynamic continuous operations on graphics objects, e.g. shrink, compare]: speech input has **performance parameter** [great potential for expressing the intent of multiple composed primitive direct manipulation actions]. Justified by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." Claims type: **T5**

**True.**

**87.** We proposed several functions that would be better served by multimodal interfaces providing speech inputs that complement direct manipulation interfaces. In particular, we proposed the use of speech to augment interaction with visualizations to: Controlling viewpoints and appropriate levels of zoom for maps and large 3D spaces (e.g., "show 59th Street and Broadway", "show Chile"). [15, 181]

Data point 87. **Generic task** [controlling viewpoints and appropriate levels of zoom in viewing maps and large 3D spaces, e.g. "Show Chile"]: speech input added to haptic (mouse manipulation) input, can **performance parameter** [augment interaction]. Justified by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T14**

**True.**

# Article 16

**88**. If the task sub-goal requires abstract procedural information then prefer linguistic media with speech [to linguistic media with text and to visual media] for simple, short operations. [16, 238]

Data point 88. **Generic task** [rendering abstract procedural information]: prefer speech output to text or analogue output modalities for simple, short operations. Supported by MP1**:** "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP5: "Acoustic input/output modalities do not require limb (including haptic) or visual activity." And MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." Corrected by MP4: "Acoustic input/output modalities are omnidirectional." Claims type: **T4**

**NOTE:** This claim is an example of an if-then rule for modality choice. We have abandoned this approach ourselves because it tends to lead to an uncontrollable and non-maintainable multitude of badly scoped claims (rules). The claim illustrates the problems of correct scoping: if the environment is noisy, the claim is false. If the environment is quiet and speech does not disturb others, and if the task is heads-up or hands-busy, the claim is reasonable. Still, it is quite possible

that, for some tasks requiring inspection of screen output, graphic text (hand-written or typed) output might not be, on occasion, as suitable as speech output.

**Partly true.**

**89**. If the task sub-goal requires abstract procedural information then prefer linguistic media with ... text for complex, longer procedure. [16, 238]

Data point 89. **Generic task** [rendering abstract procedural information]: prefer linguistic (text) output modalities to speech output for complex, longer procedure. Justified by MP8: ”Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection.” Claims type: **T11**

**NOTE:** Assume that the procedure is heads-up/eyes-busy. In that case, text has the drawback of being eyes-busy as well. Still, the argument would be that, if the user cannot remember the instructions for the procedure to be performed, it is a decidedly secondary advantage that the user has eyes-free to perform that procedure. So, in a certain context, one Modality Property may decide the solution to a problem even though other Modality Properties, abstractly speaking, would count against that solution. In the assumed case, MP5 would count against using text.

**True.**

**90.** If the task sub-goal requires attributes with descriptive information for situation, physical objects, then prefer visual media with still images [to linguistic media]. [16, 238]

Data point 90. **Generic task** [rendering descriptive information for situation, physical objects]: prefer static graphic images output to linguistic output modalities. Supported by MP23, Corrected by MP23: ”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.” Claims type: **T11**

**NOTE:** The reason why this claim is only partly true, and seriously so, is apparent from MP23. The category of images comprise much more than static graphic images. Depending on the information to be presented on objects or situations, and to which user group, one may prefer static graphic images (“still images”), dynamic graphics images, static haptic images or dynamic haptic images.

**Partly true.**

**91.** If the task sub-goal requires attributes with descriptive information for abstract object properties or values, then prefer linguistic media [to visual media]. [16, 238]

Data point 91. **Generic task** [rendering descriptive information for abstract object properties or values]: prefer linguistic output modalities to analogue graphics output. Justified by MP1**:** "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP23: "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." Claims type: **T4**

*Assumption:* by 'visual media' the author means graphic images.

**NOTE:** The confused distinction between "linguistic media" and "visual media". Text, whether hand-written or typed, can be linguistic *and* visual (text can also be non-visual, e.g. when it is haptic).

**True. Conceptually confused.**

**92.** If the task sub-goal requires rules or heuristics for decision making then use linguistic media as text. Speech may be used, but beware: speech is not persistent. [16, 238]

Data point 92. **Generic task** [rendering rules or heuristics for decision making]: sometimes prefer text output modalities to speech output because speech is not persistent. Justified by MP1**:** "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." Claims type: **T11**

**True.**

**93.** For event information use audio media for sound warning but present the context of event messages using text for descriptive/status information and image for physical/spatial detail. [16, 239]

Data point 93. **Generic task** [rendering event information]: combine acoustic output modalities for sound warning with text output modalities for descriptive/status information and image output modalities for physical/spatial detail. Supported by MP4: "Acoustic input/output modalities are omnidirectional." And MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP23: "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." Claims type: **T14**

**NOTE:** This claim is worth making even if it is clearly an over-generalisation. The quoted MPs only serve to justify the "positive" side of this very complex claim. Many more MPs could be quoted to point out why alternative modalities might be less suited for the purposes mentioned in the claim. Similarly, several MPs could be quoted that would serve to correct the claim. 'Event information' is

a very general term and it is therefore very likely that the recommended modality combination will not be optimal in some cases: sound warnings may not be needed or may be useless in a noisy environment (cf. MP6), images may be less informative than graphs, text + images may be replaced by graphs, etc. The claim, therefore, can only be supported, not justified.

**Often true.**

# Article 17

**94.** With a limited spoken vocabulary and a well-structured grammar, speech input is successful for a number of "hands-busy" and/or "eyes-busy" tasks ... This includes a range of applications such as quality control and inspections, stock control, parcel sorting ..., baggage handling ..., meter reading ..., and direct speech input to computers for medical and dental procedures ... Speech input is credited for improving time on task for each application. [17, 591]

Data point 94. **Generic tasks** ["hands-busy" and/or "eyes-busy", e.g. quality control, inspection, stock control, parcel sorting, baggage handling, meter reading, medical procedures, dental procedures]: speech input is credited for **performance parameter** [improving task performance time]. Justified by MP5: "Acoustic input/output modalities do not require limb (including haptic) or visual activity." Claims type: **T5**

**NOTE:** The support provided by MP5 is so strong that, arguably, it counts as a justification.

**True.**

**95**. With a limited spoken vocabulary and a well-structured grammar, speech input is successful for a number of "hands-busy" and/or "eyes-busy" tasks ... This includes a range of applications such as quality control and inspections, stock control, parcel sorting ..., baggage handling ..., meter reading ..., and direct speech input to computers for medical and dental procedures ... Speech input is credited for improving time on task for each application. [17, 591]

Data point 95. **Generic task** [hands-busy and/or eyes-busy, e.g. quality control, inspection, stock control, parcel sorting, baggage handling, meter reading, medical procedures, dental procedures]: speech input can improve **performance parameter** [time on task]. Justified by MP5: "Acoustic input/output modalities do not require limb (including haptic) or visual activity." Claims type: **T5**

**True.**

**96**. Immediate data entry, a second speech input success factor, can reduce the number of input errors caused by memory lapse or transcription error. [17, 591]

Data point 96. **Generic task** [immediate data entry]: speech input can reduce **performance parameter** [number of input errors] caused by **cognitive property** [memory lapse] or **performance parameter** [transcription error]. Justified by MP5: "Acoustic input/output modalities do not require limb (including haptic) or visual activity." Claims type: **T5**

*Assumption:* speech input does not cause as many speech recognition errors as transcription causes transcription errors. This assumption is not necessarily true today.

**NOTE:** The modest "can" at the centre of the claim. It is not being claimed that speech is superior to other modalities for all kinds of immediate data entry, only that speech sometimes can do what is being claimed. Obviously, the advantage over other input modalities of using speech for immediate data entry grows with the added time required for using those other modalities. So the claim would be most true when the task itself keeps hands and/or eyes busy.

**True.**

# Article 18

**97.** Speech is easy to generate. [18, 5]

Data point 97. Speech input is **performance parameter** [easy to generate]. Justified by MP17: "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)." Claims type: **T5**

**True.**

**98.** ... most people can speak much faster than they can write or type. [18, 5]

Data point 98. Speech input is **performance parameter** [much faster to produce] than haptic (text) input for **user group** [most people]. No justification. Claims type: **T6**

*Assumption:* the claim concerns standard input tasks for the creation of typed text or discourse, such as dictation.

**NOTE:** Input and/or output speed, although admittedly important to efficiency, is a highly device-, task- and user skill-dependent notion which is difficult or impossible to generalise. This is why Modality Theory has little to say about it except sometimes by implication. Empirical study may be needed.

**True.**

**99.** Also a case can be made that the conventions for formulating speech are much less demanding than those for writing. [18, 5]

Data point 99. Speech input is **performance parameter** [less demanding to produce in terms of the conventions that have to be followed] than text modalities. Justified by MP16: "Discourse input/output modalities are situation-dependent." And MP24: "Text input/output modalities are basically situation-independent." Claims type: **T6**

**NOTE:** It is the situation-independent character of text which has imposed on it a degree of formality which is not required of speech. Speech is for people who synchronously share a situation, text is for people who, asynchronously, find themselves in different situations.

**True.**

**100.** ... a written message can be easily previewed to ascertain its structure and likely content. After reading it can easily be re-viewed and specific parts re-visited. This is an advantage for the writer as well as the reader as it makes the message easier to edit. Even with a very well-designed user interface for playing back recorded messages ... it is always awkward to review selected parts of a spoken message and impossible to preview its structure. [18, 5-6]

Data point 100. Speech output is **performance parameters** [awkward/less easy to re-view/re-visit/edit for structure and content] and [impossible to preview for structure] compared with text modalities. Justified by MP7: "Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction." And MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." And MP9: "Dynamic acoustic output modalities can be made interactively static (but only small-piece-by-small-piece)." Claims type: **T11**

**True.**

**101.** The main result was that it took longer to perform [the] information sharing tasks when only written communication was allowed [rather than voice-only]. [18, 6]

Data point 101. **Generic task** [information sharing through human-human communication]: speech input/output is **performance parameter** [faster] than text input/output modalities. No justification. Claims type: **T2**

**NOTE:** Input and/or output speed, although admittedly important to efficiency, is a highly device-, task- and user skill-dependent notion which is difficult or impossible to generalise. This is why Modality Theory has little to say about it except sometimes by implication. Empirical study may be needed.

**True.**

**102.** The main result was that it took longer to perform [the] information sharing tasks when only written communication was allowed [rather than writing-and-voice]. [18, 6]

Data point 102. **Generic task** [information sharing through human-human communication]: combined speech input/output and haptic (hand-written text) input/output were **performance parameter** [faster] than text-only modalities. No justification. Claims type: **T14**

**NOTE:** Input and/or output speed, although admittedly important to efficiency, is a highly device-, task- and user skill-dependent notion which is difficult or impossible to generalise. This is why Modality Theory has little to say about it except sometimes by implication. Empirical study may be needed.

**True.**

**103.** They had MBA students annotate written abstracts to help the author revise them. The abstracts contained copy-editing, structural and semantic errors. Copy-editing problems were best dealt with by text annotations. That is, when only allowed to use text they corrected more of these errors than when only allowed to use speech. The authors describe copy-editing errors as 'local' in contrast to the more 'global' structural and semantic errors. [18, 6]

Data point 103. **Generic task** [text copy-editing error annotation]: text (hand-writing) input gave **performance parameter** [more thorough] annotation than speech input. Justified by MP1: "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." And MP2: "Linguistic input/output modalities, being unsuited for specifying detailed information on spatial manipulation, lack an adequate vocabulary for describing the manipulations." And MP21: "Haptic deictic input gesture is

eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T13**

**NOTE:** The hypothesis supported by MP1 and MP21 is that the deictic character of the pen makes it much more easy to indicate copy errors than when using speech. It is, e.g., much easier to just correct a misspelled word by hand than to have to start by saying "in line x word no. y from the left, [word], change letters such-and-such into letters rather-such-and-such".

**True.**

**104.** They had MBA students annotate written abstracts to help the author revise them. The abstracts contained copy-editing, structural and semantic errors. ... The authors describe copy-editing errors as 'local' in contrast to the more 'global' structural and semantic errors. These [latter] were more likely to be corrected when the annotations had to be by speech. [18, 6]

Data point 104. **Generic task** [annotation of structural and semantic errors in texts]: speech input gave **performance parameter** [more thorough] annotation than text (hand-writing) input. Justified by MP16: "Discourse input/output modalities are situation-dependent." And MP24: "Text input/output modalities are basically situation-independent." And MP15: "Discourse input/output modalities have strong rhetorical potential." And MP25: "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." Claims type: **T6**

**NOTE:** The hypothesis supported by MP16 and MP24 is that the situation-independent character of text makes text input more cumbersome than speech input for the described task. In addition, it is more easy to make nuanced comments in speech.

**True.**

**105.** They conclude that speech is more expressive than text. [18, 6]

Data point 105. Speech input/output is more expressive than text input/output. Justified by MP15: "Discourse input/output modalities have strong rhetorical potential." And MP16: "Discourse input/output modalities are situation-dependent." And MP24: "Text input/output modalities are basically situation-independent." And MP25: "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." Claims type: **T6**

**NOTE:** The quoted Modality Properties can be seen to reflect the fact that, word-by-word, speech has more information channels than text.

**True.**

**106.** Annotations were more likely to be polite or complimentary in tone when spoken compared to when they were written. This affected the recipients who judged reviewers who composed in speech more favourably than those who composed in text. [18, 7]

Data point 106. **Generic task** [annotation of errors in texts]: speech input was more likely to be **cognitive property** [polite or complimentary in tone] than haptic text (hand-writing) input. Supported by MP15: ”Discourse input/output modalities have strong rhetorical potential.” And MP25: ”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.” Claims type: **T6**

**NOTE:** This support seems pretty strong but it cannot be excluded, for instance, that one would find cultural differences wrt. the amount of politeness included in the textual annotations. Whether culturally dependent or not, differences in politeness would seem too detailed for inclusion in a manageable set of Modality Properties.

**True.**

**107.** A system offering speech and written annotation simultaneously might be advantageous .... maximising the communicative value of each message [18, 7-8]

Data point 107. **Generic task** [annotation of errors in texts]: combined speech input and haptic text [hand-writing] input might be **performance parameters** [advantageous, maximise the communicative value of each message] compared to speech input-only and haptic text (hand-writing) input-only. Justified by MP15: ”Discourse input/output modalities have strong rhetorical potential.” And MP16: “Discourse input/output modalities are situation-dependent.” And MP21: ”Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information.” And MP24: ”Text input/output modalities are basically situation-independent.” And MP25: ”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.” Claims type: **T14**

**NOTE:** The claim is a rather modest one (“might be”), and can therefore be justified by providing the reasons why, in many cases, speech is likely to get more information across than hand-writing, as well as why hand-writing sometimes is likely to get more information across.

**True.**

**108.** Perhaps the most important benefit of using speech and writing together is that it makes it easy to separate the ‘talk’ from what is talked about. [18, 7]

Data point 108. Perhaps the most important benefit of combined speech input/output, haptic (hand-written text) input and static graphic (text) output may be the **performance parameter** [easy separation of the 'talk' from what is talked about]. Corrected by MP7 "Static graphic modalities allow the simultaneous representation of large amounts of information for free visual inspection." And MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." And MP12: "Speech output modalities may complement graphic displays for ease of visual inspection." Claims type: **T14**

**NOTE:** The basic difference in ease of perceptual inspection between speech and text would seem to constitute the most important benefit of using speech and writing together. People do not seem to have difficulties separating the talk from what is talked about in speech-only conversation.

**False.**

**109.** Writing and drawing can be used as a shared artefact, a permanent record of what is being discussed. This permits deixis and the efficient use of language in the discussion whether that discussion is spoken or written. [18, 8]

Data point 109. Combined static graphic (text and drawing) output can **performance parameter** [be shared] and permits haptic (deixis) input and **performance parameter** [efficient use] of linguistic input. Justified by MP7: "Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction." Claims type: **T14**

**NOTE:** The claim essentially points out the permanent and shared nature of static graphics output. The "efficient use of language" -bit should not mislead. Language is (or can be) quite efficient without support from shared artefacts.

**True.**

# Article 19

**110.** User satisfaction [with speech interfaces] depends on socio-professional category. [19, 343]

Data point 110. Speech input/output: **cognitive property** [satisfaction] depends on **user group** [socio-professional category]. No justification. Claims type: **T7**

**NOTE:** The claim is extremely general and non-specific. Users' socio-professional category may play *some* role in the user satisfaction caused by all or most kinds of interfaces. From an informative and useful claim we want to know, e.g., which categories, or what the dependency consists in.

To vague to justify or support.

**111.** Speech is desirable in certain situations. ... The learning of the interface is usually faster. [19, 343]

Data point 111. Speech input/output interfaces are **learning parameter** [faster to learn]. Supported by MP17: ”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).” Claims type: **T1**

**NOTE:** The claim does not answer the question: “faster than what?” and is impossible to evaluate more exactly.

**Too vague to justify.**

**112.** Speech is desirable in certain situations. Error-repair is usually more efficient. [19, 343]

Data point 112. Speech input interfaces allow **performance parameter** [more efficient error repair]. No justification. Claims type: **T1**

**NOTE:** The claim does not answer the question: “more efficient than what?” and is impossible to evaluate more exactly.

**Too vague to justify or support.**

**113.** The context [for using speech input/output] may be restrictive (noise, confidentiality). [19, 343]

Data point 113. Use of speech input/output is **performance parameter** [restricted] by **work environment** [noise and confidentiality]. Justified by MP4: “Acoustic input/output modalities are omnidirectional.” Claims type: **T7**

**True.**

**114.** Speech is desirable in certain situations. ... But the machine’s linguistic level (that is, the level of language understood by the machine) requires an adaptation from the user. [19, 343]

Data point 114. **Generic system** [current limited speech understanding systems]: users must **learning parameter** [adapt their language]. Justified by MP14: “Non-spontaneous speech input modalities (isolated words, connected words) are unnatural and add cognitive processing load.” And MP17: ”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.)” Claims type: **T7**

**True.**

**115.** The spoken mode is superior to the written mode in that the keyboard limits input speed. [19, 344]

Data point 115. Speech input is **performance parameter** [superior] compared to haptic (typed language) input which **performance parameter** [involves a keyboard that limits input speed]. No justification. Claims type: **T6**

**NOTE:** One interpretation is the true claim that most people can speak much faster than they can write or type. However, the present claim does not contain qualifications such as 'most people', 'for most tasks" etc. The claim is clearly an over-generalisation. As it stands, it is probably best characterised as a claim that is too vague to justify or support: it is not helpful to the understanding of speech functionality.

**Too vague to justify or support.**

**116.** The spoken mode is superior to the written mode in that ... [the latter] mobilises user sensori-motor resources. [19, 344]

Data point 116. Speech input is **performance parameter** [superior] compared to haptic (typed language) input which **cognitive property** [mobilises user sensori-motor resources]. No justification. Claims type: **T6**

**NOTE:** This claim might be interpreted as expressing the Modality Property that speech does not require haptic or visual activity. However, the present claim does not contain qualifications such as 'most people', 'for most tasks" etc. It is certainly true that the keyboard mobilises user sensori-motor resources but it does not follow that speech input is superior. Some tasks benefit from typing, such as the write-up of formalisms. And long-term use of speech instead of typing may have damaging effects on users. Etc. The claim is clearly an over-generalisation. But the real problem may be that Modality Properties - supposing that the claim is intended to express a Modality Property - are generally "neutral". The fact that, e.g., speech is non-visual is not "good in itself" but its value depends on the circumstances. As it stands, it is probably best characterised as a claim that is too vague to justify or support: it is not helpful to the understanding of speech functionality.

**Too vague to justify or support.**

**117.** A few general observations may be made about the adequacy and applicability of each mode: Spoken mode: - Input as: commands. [19, 346]

Data point 117. For **speech act** [command]: speech input is **performance parameter** [adequate]. Supported and Corrected by MP17: "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.)" Claims type: **T5**

**NOTE:** The correction is that speech input is adequate for spontaneous commands and non-spontaneous commands which the users have had a chance of learning by rote. Sometimes they do not have that chance, either because they do not use the system often enough or because there are just too many commands to remember.

**Partly true.**

**118.** A few general observations may be made about the adequacy and applicability of each mode: Spoken mode: ... - Output as: help, examples, requests, explanation, suggestion. [19, 346]

Data point 118. **Generic tasks** [providing help, examples, requests, explanation, suggestion]: speech output is **performance parameters** [adequate]. Supported by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." Corrected by MP19: "Analogue graphics input/output modalities lack interpretational scope, which makes them eminently suited for conveying high-specificity information." Corrected by MP23: "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." Claims type: **T3**

**NOTE:** The reason the claim is only supported (not justified) by the quoted part of MP1 is that the claim is partly false because it is a blatant over-generalalisation. In particular, many exemplifications require concrete illustrations of the kind provided by, e.g., graphic or acoustic images, and many explanations require, e.g., graphic diagrams.

**Partly true.**

**119.** Users tend to specialise the use of the different modes [speech and gesture], where speech is used, for example, for repetition or for commands which do not require looking at the screen for precise positioning. [19, 370]

Data point 119. **Generic task** [graphic object creation, e.g. black circles]: speech input is **cognitive property** [preferred] to haptic (gesture) input for **speech act** [commands not requiring precise spatial positioning]. Justified by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP20: "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." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T5**

**NOTE:** As long as the objects to be created can be easily specified in speech, users are likely to prefer speech to haptic manipulation which would require additional operations with the haptic device.

**True.**

# Article 20

**120.** Synthetic speech is widely used to enable blind people to receive output from computer systems. However, speech is slow to use compared with vision and places far higher demands on short-term memory. These problems are particularly apparent when exploring large data structures such as lists and tables. [20, 51]

Data point 120. **Generic task** [exploring large data structures, e.g. lists and tables]: speech output is **performance parameter** [slower to use] than graphics (text) output and places **cognitive property** [higher demands on short-term memory]. Justified by MP8: ”Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection.” And MP9: ”Dynamic acoustic output modalities can be made interactively static (but only small-piece-by-small-piece).” And MP7 “Static graphic modalities allow the simultaneous representation of large amounts of information for free visual inspection.” Claims type: **T11**

**True.**

# Article 21

**121.** When generating text from speech, eliminating the use of a keyboard or the intermediary of a typist has the promise ... of attracting non-typists. [21, 431]

Data point 121. **Generic system** [speech-to-text]: promises **cognitive property** [attracting] non-typists compared to using haptic (typed language) input. Justified by MP5: “Acoustic input/output modalities do not require limb (including haptic) or visual activity.” And MP17: ”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).” Claims type: **T6**

**NOTE**: Everything here hinges on the word "promises". If taken at face value, MP17 justifies the promise. However, the promise may be false. Claims about generic systems are known to have many usability pros and cons, in this case, for instance, the long-term effects of speaking constantly.

**True.**

**122.** Studies with speech products to date suggest that skilled typists are slowed down by using a speech dictation system. [21, 431]

Data point 122. Compared to haptic (typed language) input, speech input is **performance parameter** [slower] for **user group** [skilled typists]. No justification. Claims type: **T13**

**NOTE**: Input and/or output speed, although admittedly important to efficiency, is a highly device-, task- and user skill-dependent notion which is difficult or impossible to generalise. This is why Modality Theory has little to say about it except sometimes by implication. Empirical study may be needed.

**True.**

**123.** Studies with speech products to date suggest that ... speech dictation system[s] ... are best suited for non-typists or people with typing disabilities. [21, 431]

Data point 123. Compared to haptic (typed language) input, speech input is **performance parameter** [best suited] for **user groups** [non-typists, people with typing disabilities]. No justification. Claims type: **T7**

**NOTE**: The reason for this claim is that skilled typists can type faster than they can dictate. Input and/or output speed, although admittedly important to efficiency, is a highly device-, task- and user skill-dependent notion which is difficult or impossible to generalise. This is why Modality Theory has little to say about it except sometimes by implication. Empirical study may be needed.

**True.**

**124.** Even some of the doctors who are comfortable with computers found that they are uncomfortable relying on speech recognition. One of the residents we worked with said that he knew how to use computers, and he knew how to dictate reports, but he "felt strange speaking to the computer" and that it made him "feel like he had to talk to a robot". [21, 436]

Data point 124. Speech input is **cognitive property** [uncomfortable] for **user group** [some computer literates]. No justification. Claims type: **T12**

**NOTE**: The simple comment is that, probably, every new input technology has met with conservatism at an early stage. This has nothing to do with speech in particular. Empirical study may be needed.

**True.**

**125.** We found that for the functions that initiate a new report, initiate dictation or finalise the report, use of voice was preferred over use of mouse and keyboard. These are the functions that can be completed from start to finish without any requirement to use the keyboard. However, for those functions that do require some use of the keyboard there was a strong tendency to execute other commands in that function by continuing with the keyboard or using the mouse. This was also true of the editing and correcting of the report that was done predominantly with mouse and keyboard. It was very clear from our observations that users tend to find a pattern of modality usage and stick with it. [21, 437]

Data point 125. **Generic tasks** [medical report creation]: speech input was **cognitive property** [preferred] to haptic (mouse, keyboard) input when **performance parameter** [keyboard independence, e.g. initiate new report, initiate dictation, finalise report], but was **cognitive property** [deferred] to haptic (mouse, keyboard) input when **performance parameter** [keyboard dependence, e.g. edit and correct report]. Justified by MP17: ”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).” And MP5: “Acoustic input/output modalities do not require limb (including haptic) or visual activity.” And MP1: “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.” And MP21: ”Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information.” Claims type: **T14**

**True.**

# Article 22

**126.** A major design goal for QuickSet is to provide the same user input capabilities for handheld, desktop, and wall-sized terminal hardware. We believe that only voice and gesture-[pen-] based interaction comfortably span this range. [22, 32]

Data point 126. **Generic systems** [handheld, desktop, and wall-sized terminal hardware]: only combined speech input and haptic (pen) input are believed to **performance parameters** [comfortably span this range (handheld, desktop, and wall-sized terminal hardware)]. Supported by MP5: “Acoustic input/output modalities do not require limb (including haptic) or visual activity.” Claims type: **T14**

**NOTE**: MP5 implies that speech input can be used with the systems mentioned and that it can be used along with haptics. One can speak whilst moving one’s limbs and all the devices mentioned could include a microphone. It is obvious that pen-based input does not require a keyboard which would sit badly with handheld

devices. However, Modality Theory does not explicitly address devices and can have difficulty doing more than support device claims.

**Too unlimited to justify.**

**127.** QuickSet provides *both* of these modalities [voice and pen] because it has been demonstrated that there exists substantive language, task, performance, and user preference advantages for multimodal interaction over speech-only and gesture-only interaction with map-based tasks ... Specifically, for these tasks, multimodal input results in ... 23% fewer words, as compared to a speech-only interaction. [22, 32]

Data point 127. **Generic task** [map-based, e.g. location descriptions for roadmaps]: combined speech input and haptic (pen) input produce **performance parameter** [23% fewer words] than speech-only interaction. Justified by MP1: "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." And MP2: "Linguistic input/output modalities, being unsuited for specifying detailed information on spatial manipulation, lack an adequate vocabulary for describing the manipulations." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T14**

**NOTE**: The exact percentage mentioned is not being justified, of course.

**True.**

**128.** Multimodal pen-voice interaction is known to be advantageous for small devices. [22, 32]

Data point 128. **Generic system** [small device]: combined speech input and haptic (pen) input is **performance parameter** [advantageous]. Supported by MP5: "Acoustic input/output modalities do not require limb (including haptic) or visual activity." Claims type: **T14**

**NOTE**: MP5 implies that speech input can be used with small devices and that it can be used along with haptics. One can speak whilst moving one's limbs and all the devices mentioned could include a microphone. It is obvious that pen-based input does not require a keyboard which would sit badly with small devices.

However, Modality Theory does not explicitly address devices and can have difficulty doing more than support device claims.

**True.**

**129.** Multimodal pen-voice interaction is known to be advantageous for ... mobile users who may encounter different circumstances. [22, 32]

Data point 129. **User group** [mobile users]: combined speech input and haptic (pen) input is **performance parameter** [advantageous]. Supported by MP5: "Acoustic input/output modalities do not require limb (including haptic) or visual activity." Claims type: **T14**

**NOTE**: MP5 implies that speech input can be used with mobile devices and that it can be used along with haptics. One can speak whilst moving one's limbs and all the devices mentioned could include a microphone. It is obvious that pen-based input does not require a keyboard which would sit badly with mobility. However, Modality Theory does not explicitly address devices and can have difficulty doing more than support device claims.

**True.**

**130.** Multimodal pen-voice interaction is known to be advantageous for ... error avoidance and correction. [22, 32]

Data point 130. Combined speech input and haptic (pen) input is **performance parameter** [advantageous for error avoidance and correction]. No justification. Claims type: **T14**

**NOTE:** This is a non-specific claim which, moreover, does not indicate with which other input modalities the pen-voice combination is being compared in an implicit way.

Too vague to justify or support.

**131.** Multimodal pen-voice interaction is known to be advantageous for ... robustness. [22, 32]

Data point 131. Combined speech input and haptic (pen) input is **performance parameter** [advantageous for robustness]. No justification. Claims type: **T14**

**NOTE:** This is a highly non-specific claim which, moreover, does not indicate with which other input modalities the pen-voice combination is being compared in an implicit way. 'Robustness' is one of those laudatory terms which may mean so much that they have ended up meaning nothing in and by themselves.

**Too vague to justify or support.**

**132.** Spoken interaction with virtual worlds offers distinct advantages over direct manipulation, in that users are able to describe entities and locations that are not in view, can be teleported to those out-of-view locations and entities, and can ask questions about entities in the scene. [22, 33-35]

Data point 132. **Generic task** [interacting with virtual worlds in the graphic output domain]: speech input is preferable to haptic (direct manipulation) input because it affords **speech act** [describing entities not in view], **speech act**

[commanding teleporting] and **speech act** [asking questions about entities in view]. Justified by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T6**

**True.**

**133.** Given that numerous difficult-to-process linguistic phenomena (such as utterance disfluencies) are known to be elevated in lengthy utterances and also to be elevated when people speak locative constituents ..., multimodal interaction that permits pen input to specify locations offers the possibility of more robust recognition. [22, 39]

Data point 133. **Generic task** [specifying locations in interaction with maps]: speech-only input is **performance parameters** [lengthier, more disfluent, more difficult to process] than speech input combined with haptic (pen-based) input. Justified by MP1: "Linguistic input/output modalities ... are ... unsuited for specifying detailed information on spatial manipulation and location." And MP21: "Haptic deictic input gesture is eminently suited for spatial manipulation and indication of spatial location. It is not suited for conveying abstract information." Claims type: **T13**

**True.**

# Article 23

**134.** Rich media (face-to-face and telephone) are proposed to be suitable for resolving equivocal situations. ... Rich oral media facilitate equivocality reduction by enabling individuals to process multiple, conflicting interpretations of a situation. Thus, it is believed that oral media are preferred for communication situations high in equivocality, while written media [written documents] are preferred for communication situations low in equivocality. ... Media high in richness, such as face-to-face interaction and telephone calls, enable negotiation, clarification, explanation and exchange of subjective views. [23, 444-446]

Data point 134. **Generic task** [resolving equivocal situations] + **interaction modes** [face-to-face and telephone]: speech input/output are **performance parameter** [preferable] to graphics (text) input/output. Justified by MP16: "Discourse input/output modalities are situation-dependent." And MP24: "Text input/output modalities are basically situation-independent." Claims type: **T1**

**NOTE**: It is exactly the situatedness of speech and other forms of discourse that makes them superior to text for the task described.

**True.**

**135.** Lean media (written documents) are proposed to be more suitable for reducing uncertainty. ... Rich oral media facilitate equivocality reduction by enabling individuals to process multiple, conflicting interpretations of a situation. Thus, it is believed that oral media

are preferred for communication situations high in equivocality, while written media are preferred for communication situations low in equivocality. ... Media high in richness, such as face-to-face interaction and telephone calls, enable negotiation, clarification, explanation and exchange of subjective views. On the other hand, media low in richness, such as written media, although not appropriate for resolving equivocal issues, are most appropriate for processing large amounts of standard, accurate, objective and quantitative data. [23, 444-446]

Data point 135. **Generic task** [reducing uncertainty, processing large amounts of standard, accurate, objective and quantitative data]: graphics (text) input/output are **performance parameter** [preferable] to speech input/output. Justified by MP7: "Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction." And MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." And MP9: "Dynamic acoustic output modalities can be made interactively static (but only small-piece-by-small-piece)." Claims type: **T9**

*Assumption:* 'reducing uncertainty' primarily refers to the uncertainty that easily arises when people orally discuss large amounts of data with which they are not thoroughly familiar.

**NOTE**: If the above assumption is not being made, it becomes far less clear that speech is inferior to text for uncertainty reduction. The claim, therefore, would appear a potentially confusing one.

**True.**

**136.** First, voice mail provides dynamic verbal cues that reflect a person's tone of voice, inflections and emotions while email can only convey static visual cues in text. Thus, voice mail is richer in terms of its capacity to convey multiple cues. The ability to interpret a communicating partner's tone of voice is a significant advantage of voice mail. Second, voice mail uses natural language which, together with audio cues, provides language variety and language content. In email, while natural language is employed, audio cues are absent, which limits its language variety. Third, the audio nature of voice mail makes it more amenable to the transmission of feelings and emotions ... Thus, personal focus is likely to be higher in voice mail than email. ... In sum, due to voice mail's ability to provide verbal cues and inflections and greater personal focus, in addition to its oral nature, it can be considered a richer medium than email. [23, 447-450]

Data point 136. **Generic system** [voice mail] is more expressive than **generic system** [email] because of providing output speech and non-speech audio cues for emotions that are absent in static graphics (text) output. Justified by MP15: "Discourse input/output modalities have strong rhetorical potential." And MP25: "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." Claims type: **T4**

**NOTE**: Word for word, speech is more expressive than standard writing. But note that this isolated observation does not prove that voice mail is superior to email!

**True.**

**137.** It is cognitively taxing for a user of voice mail to process lengthy messages or complex sequences of messages. [23, 450]

Data point 137. **Generic system** [voice mail] + **generic task** [processing lengthy messages or complex sequences of messages]: speech output is **cognitive property** [taxing]. Justified by MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." And MP9: "Dynamic acoustic output modalities can be made interactively static (but only small-piece-by-small-piece)." Claims type: **T10**

**True.**

**138.** Email messages promote accuracy. Senders of email messages can take their time in composing and editing while voice mail users cannot. [23, 450]

Data point 138. **Generic system** [email] promotes **performance parameter** [accuracy] compared to speech input using **generic system** [voice mail]. Justified by MP7: "Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction." And MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." Claims type: **T13**

**NOTE**: It may be interesting to note that MP7 and MP8 are true both of message *production* and message *reception.* What was said was said and can be hard to take back whereas a written message only becomes binding when it has been dispatched.

**True.**

**139.** Email receivers can increase information accuracy by reading messages slowly or printing a hardcopy. To accomplish the same purpose, voice mail messages have to be transcribed. [23, 450]

Data point 139. For **cognitive property** [increasing information reception accuracy]: static graphics (typed) output using **generic system** [email] is **performance parameter** [easier to use] than speech output using **generic system** [voice mail]. Justified by MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." And MP9: "Dynamic acoustic output modalities can be made interactively static (but only small-piece-by-small-piece)." And MP7 "Static graphic/haptic modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection." Claims type: **T11**

**True.**

**140.** The written nature of email preserves the formality of technical reports, data or formal requests and responses. While quantitative data can be communicated via both media, senders

and receivers of email messages can process quantitative data more efficiently and effectively than voice mail users. [23, 450]

Data point 140. **Generic task** [human-human communication of quantitative data]: graphics (typed) input/output using **generic system** [email] is **cognitive property** [processed more efficiently and effectively] than speech input/output using **generic system** [voice mail]. Justified by MP7: "Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction." And MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." And MP9: "Dynamic acoustic output modalities can be made interactively static (but only small-piece-by-small-piece)." And MP10: "Speech input/output modalities, being temporal (serial and transient) and non-spatial, should be presented sequentially rather than in parallel." Claims type: **T9**

**True.**

**141**. Individuals preferred to communicate via email as opposed to voice mail for situations that require explanation or negotiation. Thus, for the specific case of email and voice mail, the results fail to support the hypothesised relationship derived from MRT [Media Richness Theory] between a medium's richness and its ability to handle equivocality. [23, 456]

Data point 141. **Generic tasks** [explanation or negotiation]: **generic system** [email] was **cognitive property** [preferred] to **generic system** [voice mail]. Justified by MP16: "Discourse input/output modalities are situation-dependent." And MP24: "Text input/output modalities are basically situation-independent." Claims type: **T9**

**NOTE**: This claim and its explanation highlight the hybrid nature of voice mail which makes it comparatively undesirable to use for certain purposes, such as the tasks mentioned. In voice mail, we use a situated modality asynchronously, i.e. in a non-situated way. Speech is basically "designed" for shared-situation communication and explanation and negotiation are paradigm cases of communication which are best resolved using shared-situation (synchronous) communication. However, if the shared situation goes away, as in voice mail, speech becomes very awkward to use and situation-independent modalities, such as email, muscle in and become preferred.

**True.**

**142.** Interviewees generally did not perceive voice mail as an appropriate medium for communicating information to resolve equivocality. Voice mail was preferred for short, spontaneous, one-way drops of information, in contrast to the lengthy, ongoing, prolonged and ambiguous communication typical of equivocal situations. ... "If an issue requires back and forth communication I am much more comfortable on email. Messages are more understandable ... since people have thought them through. Sometimes people don't think

through [a message] on voice mail. They tend to ramble and are not focused. I have a lot of work to do to narrow down the issue." [23, 456]

Data point 142. **Generic tasks** [short, spontaneous, one-way drops of information, as opposed to the lengthy, ongoing, prolonged and ambiguous communication typical of equivocal situations]: **generic system** [voice mail] was **cognitive property** [preferred] to **generic system** [email]. Justified by MP16: "Discourse input/output modalities are situation-dependent." And MP17: "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)." And MP25: "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." Claims type: **T6**

**NOTE**: As long as speakers can communicate in a situation-dependent way, speech will be preferred to typing. Because voice mail is asynchronous, it only, or at most, has a chance vs. email for one-way spontaneous communication. Note that the claim doesn't say what the recipients would have preferred!

**True.**

**143.** In contrast to claims about the usefulness of multiple cues in promoting clarity, facilitating meaning and reducing ambiguity, voice mail's capacity for conveying multiple cues were unimportant in Aerco employees' media choices: "The tone of voice (is) not important in PhoneMail. I don't believe that adds any personal touch." "Verbal cues introduce distortions ... since voice is harder to understand and interpret." [23, 456-457]

Data point 143. **Generic system** [email] output was found to be **cognitive property** [easier to understand and interpret] than **generic system** [voice mail] output by **user group** [Aerco employees]. Justified by MP16: "Discourse input/output modalities are situation-dependent." And MP24: "Text input/output modalities are basically situation-independent." Claims type: **T11**

**NOTE**: When the situation is not shared by speaker and listener, the listener is likely to prefer a situation-independent output modality to a situation-dependent one because the latter introduces uncertainty in the listener who is ignorant of the situation of production of the message.

**True.**

**144.** Interviewees found vmail messages hard to manipulate, store, print and file or send to multiple people. [23, 457]

Data point 144. **Generic system** [voice mail] messages were found **performance parameter** [hard] to manipulate, store, print and file or send to multiple people. Supported by MP8: ”Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection.” And MP9: ”Dynamic acoustic output modalities can be made interactively static (but only small-piece-by-small-piece).” Claims type: **T10**

**NOTE**: The claim is only partly justified by MP8 and MP9 which only address the manipulation difficulty. Modality Theory does not explicitly address devices and can have difficulty doing more than support device claims.

**True.**

**145.** The textuality of electronic mail was found convenient in cases where senders have low voices or heavy accents. The tendency of some people to talk too fast was not a problem for their communication partners when the speakers used email. [23, 459]

Data point 145. **Generic system** [email] output is **cognitive property** [more convenient] than **generic system** [voice mail] output in case of **user group** [low voices, heavy accents, speaking too fast]. Supported by MP8: ”Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection.” And MP17: ”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).” Claims type: **T11**

**NOTE**: MP8 suggests that low voices and speaking too fast is a problem. MP17 suggests that accents can be a problem for the recipient. A full justification would require too fine-grained (and therefore too many) modality properties.

**True.**

**146.** Email users particularly appreciated the ability to download numbers into a spreadsheet package and display them graphically. Graphical displays promote memorability for the recipient. By contrast, recipients of voice mail messages found it difficult to remember numbers unless they transcribed them. [23, 459]

Data point 146. Graphics (typed) output promotes **cognitive property** [memorability] by contrast to speech output. Justified by MP7: ”Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction.” And MP8: ”Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection.” And MP9: ”Dynamic acoustic output modalities can be made interactively static (but only small-piece-by-small-piece).” Claims type: **T11**

**True.**

**147.** The questionnaire data reveals that users preferred email [to voice mail] ... for making comments, annotating documents, making corrections and returning documents to co-authors. [23, 460]

> Data point 147. **Generic tasks** [making comments on documents, annotating documents, making corrections and returning documents to co-authors]: **generic system** [email] was **cognitive property** [preferred] to **generic system** [voice mail]. Justified by MP7: "Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction." And MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." Claims type: **T13**
>
> *Assumption:* the tasks described involve attention to detail.
>
> **NOTE**: Email input can reflect attention to detail much more easily than spontaneous speech.
>
> **True.**

**148.** For receiving messages, respondents clearly preferred email .... Email was viewed as having numerous benefits to message receivers. Email allows recipients to scan quickly across and within messages, enabling them to concentrate on important points and ignore irrelevant ones. Furthermore, when message senders have used email's carbon copy and blind carbon copy features, recipients can easily stay up to date on what is happening and can gain valuable clues about organisational politics and issue urgency by examining who has been involved in a particular communication. While vmail has some scanning and carbon copy features, they are quite limited in comparison to email. In PhoneMail, for example, recipients could scan message headers and listen to them in the order they chose, and they could adjust the speed at which messages were played back. But respondents preferred email ..., mainly due to its visual nature which was believed to provide more mailbox control. [23, 461]

> Data point 148. **Generic task** [receiving messages] + **user group** [Aerco employees]: **generic system** [email] was **cognitive property** [preferred] to **generic system** [voice mail] because email allows **performance parameter** [faster scanning], **cognitive properties** [concentration on important points, ignoring irrelevant points], includes **functionality parameters** [less limited carbon copy and blind carbon copy features], and provides **performance parameter** [more mailbox control] due to its visual nature. Supported by MP7: "Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction." And MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." And MP9: "Dynamic acoustic output modalities can be made interactively static (but only small-piece-by-small-piece)." Claims type: **T11**
>
> **NOTE**: Modality Properties cannot be used in justifying criticisms of system functionality that happens to be missing but might have been available.
>
> **True but partly irrelevant.**

**149.** By contrast, for sending messages, our respondents preferred vmail. ... It seems that senders prefer vmail because it helps them get work off their desks faster. It is generally more accessible and easy to use. The sender who is comfortable with dictation can speak a message informally without worrying about grammar or spelling. [23, 461-462]

Data point 149. **Generic task** [sending messages] + **user group** [Aerco employees]: **generic system** [voice mail] seems to have been **cognitive property** [preferred] to **generic system** [email] because of being **performance parameters** [faster, more accessible and easy to use] and **cognitive property** [more informal]. Justified by MP16: "Discourse input/output modalities are situation-dependent." And MP17: "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)." And MP24: "Text input/output modalities are basically situation-independent." Claims type: **T6**

**NOTE**: The claim does not at all justify the overall superiority of vmail. It basically says that, for the sender, it is faster to "get rid of" a message by speaking than by typing in a particular environment. The basic situation-independent character of text adds formality and limits spontaneity in the production process.

**True.**

**150.** Viewed another way, the findings of this study show that one medium, email, was "dominant" or strongly preferred to the other [vmail] for almost all asynchronous communication tasks. [23, 462]

Data point 150. **Generic tasks** [almost all asynchronous communication tasks]: **generic system** [email] was **cognitive property** [strongly preferred] to **generic system** [voice mail]. Justified by MP16: "Discourse input/output modalities are situation-dependent." And MP24: "Text input/output modalities are basically situation-independent." Claims type: **T9**

**NOTE**: Terse as these justifications are, they pin-point the basic facts that speech is not at all "designed" for asynchronous communication but, rather, for shared-situation dialogue, and that text, in particular static graphics text, is "designed" (or has evolved) for asynchronous communication.

**True.**

**151.** For instance, face-to-face and telephone allow for immediate feedback, the transmission of multiple cues, language variety, etc., whereas written addressed communications and written unaddressed communication do not. [23, 462]

Data point 151. Speech + **interaction mode** [face-to-face, telephone]: allow for immediate feedback, the transmission of multiple cues, language variety, etc. by contrast to graphics (text) input/output. Justified by MP16: "Discourse input/output modalities are situation-dependent." And MP24: "Text input/output modalities are basically situation-independent." And MP25: "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." Claims type: **T1**

**NOTE**: It is exactly the situatedness of speech and other forms of discourse that makes them superior to text for the task described.

**True.**

**152.** Similarly, videoconferencing is frequently claimed to be superior to audio-only conferencing more for its ability to transmit information displays (e.g. meeting overheads) than for its ability to display people's faces and gestures. [23, 463]

Data point 152. **Generic system** [video-conferencing] is frequently claimed to be **performance parameter** [superior] to **generic system** [audio-only conferencing] for its **functionality parameter** [ability to transmit information displays, e.g. meeting overheads]. Justified by MP7: "Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction." And MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." And MP17: "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)." Claims type: **T14**

**NOTE**: Combined speech and static graphics are, indeed, one of the top modality combinations of all times.

**True.**

**153.** Despite vmail's vaunted ability to convey such qualities as compassion, forgiveness or honesty, text digits are easier to process, filter and transfer than voice digits. [23, 463]

Data point 153. **Generic task** [processing, filtering and transferring digits]: speech output is **performance parameter** [less easy] to use than graphics (text) output. Supported by MP7: "Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction." And MP8: "Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." And MP9: "Dynamic acoustic output modalities can be made interactively static (but only small-piece-by-small-piece)." And MP10: "Speech input/output modalities, being temporal (serial and transient) and non-spatial, should be presented sequentially rather than in parallel." Claims type: **T11**

**NOTE**: This comes close to a justification. However, Modality Theory has nothing to say at the moment about the relative ease of data transfer because it does not deal with devices.

**True.**

# 5. Data Analysis

This section shows the data analysis which was used to compute the overall results presented in Section 1.3 above. The numbers followed by colons in the left-hand column refer to the papers from which the data were collected.

| No. | Evaluation | Truth value | Type | MPs | Success |
|---|---|---|---|---|---|
| 1: 1 | No j/s | Too vague j/s | T2 | - | out |
| 2 | No j/s | Too vague j/s | T2 | - | out |
| 3 | Just. | True | T11 | 1,19 | j-hit |
| 4 | Supp. | Too unlimited j | T14 | 1,19 | s-hit |
| 5 | Supp. | Moot | T14 | 7 | s-hit |
| 2: 6 | Just. | True | T6 | 1,5,17 | j-hit |
| 7 | Just. | True | T7 | 17,18 | j-hit |
| 8 | Just. | True | T1 | 1,15,16 | j-hit |
| 9 | Just. | True | T12 | 17,18 | j-hit |
| 3: 10 | Just. | True | T6 | 1,20 | j-hit |
| 11 | Just. | True | T6 | 17,20 | j-hit |
| 12 | Just. | True | T6 | 1,20 | j-hit |
| 13 | Just. | True | T14 | 12 | j-hit |
| 14 | Just. | True | T14 | 12 | j-hit |
| 15 | Just. | True | T14 | 1,7,19 | j-hit |
| 16 | Supp. | True | T13 | 1 | j failure |
| 17 | Supp. | True | T13 | 17 | j failure |
| 4: 18 | Just. | True | T14 | 1,21 | j-hit |
| 5: 19 | Just. | True | T5 | 5 | j-hit |
| 20 | No j/s | False | T5 | - | out |
| 21 | Just. | True | T6 | 3,17 | j-hit |
| 6: 22 | Supp.<br>Corr. | Partly true | T1 | Supp. 17<br>Cor. 17 | s-hit |
| 23 | Supp. | Too unlimited j | T14 | 1,19 | s-hit |
| 7: 24 | Supp. | Too strong j | T7 | 14,15,16,17 | s-hit |
| 8: 25 | Supp. | True | Eq.<br>mm | 17 | j failure |
| 9: 26 | Just. | True | T8 | 6 | j-hit |
| 10: 27 | Just. | True | T10 | 7,8 | j-hit |
| 11: 28 | Just. | True | T6 | 8 | j-hit |
| 29 | Supp. | True | T6 | 17 | j failure |
| 30 | Just. | True | T6 | 8 | j-hit |
| 12: 31 | Just. | True | T5 | 5,17 | j-hit |
| 32 | Just. | True | T10 | 5,8 | j-hit |

| | | | | | |
|---|---|---|---|---|---|
| 33 | Just. | True | T3 | 1,5,17 | j-hit |
| 34 | Supp. | Too unlimited j | T14 | 1,11 | s-hit |
| 35 | Supp. | True | T14 | 1,11 | j failure |
| 36 | Just. | True | T10 | 8 | j-hit |
| 37 | Just. | True | T2 | 1,5,17 | j-hit |
| 13: 38 | Just. | True | T11 | 7,8 | j-hit |
| 39 | Just. | True | T10 | 8,9 | j-hit |
| 40 | Supp. | True | T11 | 25 | j failure |
| 41 | Just. | True | T7 | 13 | j-hit |
| 42 | Just. | True | T11 | 8 | j-hit |
| 43 | Corr. | False | T5 | 17 | c-hit |
| 44 | Just. | True | T10 | 8,9 | j-hit |
| 45 | Supp. | True | T11 | 8 | j failure |
| 46 | Just. | True | T10 | 8 | j-hit |
| 47 | Supp. | True | T14 | 8 | j failure |
| 14: 48 | Just. | True | Rsc | 5 | j-hit |
| 49 | Just. | True | T5 | 5 | j-hit |
| 50 | Just. | True | T5 | 5 | j-hit |
| 51 | No j/s | Too vague j/s | T5 | - | out |
| 52 | Just. | True | T5 | 22 | j-hit |
| 53 | Supp.<br>Corr. | Partly true | T5 | Supp. 17<br>Cor. 17 | s-hit |
| 54 | Just. | True | T5 | 1,17 | j-hit |
| 55 | Just. | True | T5 | 17 | j-hit |
| 56 | Supp. | Too unlimited j | Rsc | 1,17,21 | s-hit |
| 57 | Just. | True | T14 | 1,21,22 | j-hit |
| 58 | Just. | True | T6 | 1,20 | j-hit |
| 59 | Just. | True | T13 | 1,21 | j-hit |
| 60 | No j/s | True | T6 | - | j/s failure |
| 61 | Just. | True | T6 | 1,21 | j-hit |
| 62 | Just. | True | T13 | 1,21 | j-hit |
| 63 | Just. | True | T6 | 1,21 | j-hit |
| 64 | Just. | True | T12 | 2 | j-hit |
| 65 | Just. | True | T14 | 1,2,21 | j-hit |
| 66 | Just. | True | T14 | 1,2,21 | j-hit |
| 67 | Just. | True | T14 | 1,2,21 | j-hit |
| 68 | Just. | True | T14 | 1,2,21 | j-hit |
| 69 | Just. | True | T14 | 1,2,21 | j-hit |
| 70 | Just. | True | T14 | 1,2,21 | j-hit |
| 15: 71 | Just. | True | T5 | 1 | j-hit |
| 72 | Just. | True | T12 | 18 | j-hit |
| 73 | Just. | True | T14 | 7,18 | j-hit |

| | | | | | |
|---|---|---|---|---|---|
| 74 | Just. | True | T14 | 1,20 | j-hit |
| 75 | Just. | True | T6 | 1,20 | j-hit |
| 76 | Just. | True | T14 | 1,21 | j-hit |
| 77 | Just. | True | T14 | 7,17,20 | j-hit |
| 78 | Just. | True | T14 | 1,20 | j-hit |
| 79 | No j/s | True | T6 | - | j/s failure |
| 80 | Just. | True | T6 | 1,20 | j-hit |
| 81 | Just. | True | T5 | 1 | j-hit |
| 82 | Just. | True | T14 | 1,21 | j-hit |
| 83 | Just. | True | T13 | 1,2,21 | j-hit |
| 84 | No j/s | True | T12 | - | j/s failure |
| 85 | Just. | True | T14 | 1,21 | j-hit |
| 86 | Just. | True | T5 | 1 | j-hit |
| 87 | Just. | True | T14 | 1,21 | j-hit |
| 16: 88 | Supp.<br>Corr. | Partly true | T4 | Supp. 1,5,8<br>Corr. 4 | s-hit |
| 89 | Just. | True | T11 | 8 | j-hit |
| 90 | Supp.<br>Corr. | Partly true | T11 | Supp. 23<br>Corr. 23 | s-hit |
| 91 | Just. | True | T4 | 1,23 | j-hit |
| 92 | Just. | True | T11 | 1,8 | j-hit |
| 93 | Supp. | Often true | T14 | 1,4,23 | s-hit |
| 17: 94 | Just. | True | T5 | 5 | j hit |
| 95 | Just. | True | T5 | 5 | j-hit |
| 96 | Just. | True | T5 | 5 | j-hit |
| 18: 97 | Just. | True | T5 | 17 | j-hit |
| 98 | No j/s | True | T6 | - | j/s failure |
| 99 | Just. | True | T6 | 16,24 | j-hit |
| 100 | Just. | True | T11 | 7,8,9 | j-hit |
| 101 | No j/s | True | T2 | - | j/s failure |
| 102 | No j/s | True | T14 | - | j/s failure |
| 103 | Just. | True | T13 | 1,2,21 | j-hit |
| 104 | Just. | True | T6 | 15,16,24,25 | j-hit |
| 105 | Just. | True | T6 | 15,16,24,25 | j-hit |
| 106 | Supp. | True | T6 | 15,25 | j failure |
| 107 | Just. | True | T14 | 15,16,21,24,25 | j-hit |
| 108 | Corr. | False | T14 | 7,8,12 | c-hit |
| 109 | Just. | True | T14 | 7 | j-hit |
| 19: 110 | No j/s | Too vague j/s | T7 | - | out |
| 111 | Supp. | Too vague j | T1 | 17 | s-hit |
| 112 | No j/s | Too vague j/s | T1 | - | out |
| 113 | Just. | True | T7 | 4 | j-hit |

| | | | | | |
|---|---|---|---|---|---|
| 114 | Just. | True | T7 | 14,17 | j-hit |
| 115 | No j/s | Too vague j/s | T6 | - | out |
| 116 | No j/s | Too vague j/s | T6 | - | out |
| 117 | Supp.<br>Corr. | Partly true | T5 | Supp. 17<br>Corr. 17 | s-hit |
| 118 | Supp.<br>Corr. | Partly true | T3 | Supp. 1<br>Corr. 19,23 | s-hit |
| 119 | Just. | True | T5 | 1,20,21 | j-hit |
| 20: 120 | Just. | True | T11 | 7,8,9 | j-hit |
| 21: 121 | Just. | True | T6 | 5,17 | j-hit |
| 122 | No j/s | True | T13 | - | j/s failure |
| 123 | No j/s | True | T7 | - | j/s failure |
| 124 | No j/s | True | T12 | - | j/s failure |
| 125 | Just. | True | T14 | 1,5,17,21 | j-hit |
| 22: 126 | Supp. | Too unlimited j | T14 | 5 | s-hit |
| 127 | Just. | True | T14 | 1,2,21 | j-hit |
| 128 | Supp. | True | T14 | 5 | j failure |
| 129 | Supp. | True | T14 | 5 | j failure |
| 130 | No j/s | Too vague j/s | T14 | - | out |
| 131 | No j/s | Too vague j/s | T14 | - | out |
| 132 | Just. | True | T6 | 1,21 | j-hit |
| 133 | Just. | True | T13 | 1,21 | j-hit |
| 23: 134 | Just. | True | T1 | 16,24 | j-hit |
| 135 | Just. | True | T9 | 7,8,9 | j-hit |
| 136 | Just. | True | T4 | 15,25 | j-hit |
| 137 | Just. | True | T10 | 8,9 | j-hit |
| 138 | Just. | True | T13 | 7,8 | j-hit |
| 139 | Just. | True | T11 | 7,8,9 | j-hit |
| 140 | Just. | True | T9 | 7,8,9,10 | j-hit |
| 141 | Just. | True | T9 | 16,24 | j-hit |
| 142 | Just. | True | T6 | 16,17,25 | j-hit |
| 143 | Just. | True | T11 | 16,24 | j-hit |
| 144 | Supp. | True | T10 | 8,9 | j failure |
| 145 | Supp. | True | T11 | 8,17 | j failure |
| 146 | Just. | True | T11 | 7,8,9 | j-hit |
| 147 | Just. | True | T13 | 7,8 | j-hit |
| 148 | Supp. | Partly irrelevant | T11 | 7,8,9 | s-hit |
| 149 | Just. | True | T6 | 16,17,24 | j-hit |
| 150 | Just. | True | T9 | 16,24 | j-hit |
| 151 | Just. | True | T1 | 16,24,25 | j-hit |
| 152 | Just. | True | T14 | 7,8,17 | j-hit |
| 153 | Supp. | True | T11 | 7,8,9,10 | j failure |

# 6. References


1. Furui, S.: Projects for spoken dialogue systems in a multimedia environment. *Proceedings of the ESCA workshop on Spoken Dialogue Systems,* Vigsø, Denmark, 1995, 9-16.

2. Ball, J. E. and Ling, D. T.: Spoken language processing in the personal conversational agent. *Proceedings of the ESCA workshop on Spoken Dialogue Systems,* Vigsø, Denmark, 1995, 109-112.

3. Wyard, P., Appleby, S., Kaneen, E., Williams, S. and Preston, K.: A combined speech and visual interface to the BT business catalogue. *Proceedings of the ESCA workshop on Spoken Dialogue Systems,* Vigsø, Denmark, 1995, 165-168.

4. Loken-Kim, K., Park, Y., Mizunashi, S., Fais, L. and Morimoto, T.: Verbal gestural behaviours in multimodal spoken language interpreting telecommunications. *Proceedings of Eurospeech '95*, Madrid, 1995, 281-284.

5. Haeb-Umbach, R. and Gamm, S.: Human factors of a voice-controlled car stereo. *Proceedings of Eurospeech '95*, Madrid, 1995, 1453-1456.

6. Nitta, T.: Speech recognition applications in Japan. *Proceedings of ICSLP '94*, Yokohama, Japan, 1994, 671-674.

7. Fraser, N. M. and Thornton, J. H. S.: Vocalist: A robust, portable spoken language dialogue system for telephone applications. *Proceedings of Eurospeech '95*, Madrid, 1995, 1947-1950.

8. Springer, S., Basson, S., Kalyanswamy, A., Man, E. and Yashchin, D.: The Money Talks interactive speech technology assessment: A report from the field. *Proceedings of Eurospeech '95*, Madrid, 1995, 1939-1942.

9. Mellor, B.A., Baber, C. and Tunley, C.: Evaluating automatic speech recognition as a component of multi-input device human-computer interface. *Proceedings of ICSLP '96*, Philadelphia, USA, 1996, 1668-1671.

10. Life, A., Salter, I., Tenem, J. N., Bernard, F., Rosset, S., Bennacef, S. and Lamel, L.: Data collection for the MASK kiosk: WOz vs prototype system. *Proceedings of ICSLP '96*, Philadelphia, USA, 1996, 1672-1675.

11. Basson, S., Springer, S., Fong, C., Leung, H., Man, E., Olson, M., Piterelli, J., Singh, R. and Wong, S.: User participation and compliance in speech automated telecommunications applications. *Proceedings of ICSLP '96*, Philadelphia, USA, 1996, 1680-1683.

12. Mynatt, E. D.: Transforming graphical interfaces into auditory interfaces for blind users. *Human-Computer Interaction,* Vol. 12, No. 1&2, 1997, 7-45.

13. Stevens, R. D., Edwards, A. D. N. and Harling, P. A.: Access to mathematics for visually disabled students through multimodal interaction. *Human-Computer Interaction,* Vol. 12, No. 1&2, 1997, 47-92.

14. Oviatt, S.: Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction,* Vol.12, No. 1&2, 1997, 93-129.

15. Roth, S. F., Chuah, M. C., Kerpedjiev, S., Kolojejchick, J. and Lucas, P.: Towards an information visualization workspace: Combining multiple means of expression. *Human-Computer Interaction,* Vol.12, No. 1&2, 1997, 131-185.

16. Sutcliffe, A.: Task-related information analysis. *Int. J. Human-Computer Studies* 47, 1997, 223-257.

17. Dillon, T. W.: User performance and acceptance of a speech-input interface in a health assessment task. *Int. J. Human-Computer Studies* 47, 1997, 591-602.

18. Daly-Jones, O., Monk, A., Frohlich, D., Geelhoed, E. and Loughran, S.: Multimodal messages: the pen and voice opportunity. *Interacting with Computers* 9, 1997, 1-25.

19. Caelen, J.: Multimodal human-computer interface. In E. Keller (Ed.): *Fundamentals of Speech Synthesis and Speech Recognition.* New York: John Wiley, 1994, 339-373.

20. Pitt, I. J. and Edwards, A. D. N.: An improved auditory interface for the exploration of lists. Seattle, WA: *Proceedings of ACM Multimedia 1997,* 51-61.

21. Lai, J. and Vergo, J.: MedSpeak: Report creation with continuous speech recognition. Atlanta, GA: *Proceedings of CHI 1997,* 431-438.

22. Cohen, P. R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L. and Clow, J.: QuickSet: Multimodal interaction for distributed applications. Seattle, WA: *Proceedings of ACM Multimedia 1997,* 31-40.

23. El-Shinnawy, M. and Marcus, M. L.: The poverty of media richness theory: explaining people's choice of electronic mail vs. voice mail. *Int. J. Human-Computer Studies* 46, 1997, 443-467.

24. Bernsen, N. O.: Foundations of multimodal representations: A taxonomy of representational modalities. *Interacting with Computers* 6, (4), 1994, 347-71.

25. Bernsen, N. O.: Defining a taxonomy of output modalities from an HCI perspective. *Computer Standards and Interfaces,* Special Double Issue, 18, 6-7, 1997, 537-553.

26. Bernsen, N.O.: A Toolbox of output modalities. 1997. http://www.mip.ou.dk/nis/publications/papers/toolbox_paper/index.html

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

28. Bernsen, N. O. and Dybkjær, L.: Is speech the right thing for your application? *Proceedings of the International Conference for Spoken Language Processing, ICSLP'98,* Sydney. Sydney: Australian Speech Science and Technology Association 1998, 3209-3212.

29. Bernsen, N. O. and Dybkjær, L.: Speech in multimodal systems. Paper to appear in the *Proceedings of the ESCA Tutorial and Research Workshop on Interactive Dialogue in Multi-Modal Systems,* Irsee, Germany, June 1999.