TOWARDS A TOOL FOR PREDICTING SPEECH FUNCTIONALITY

Niels Ole Bernsen, The Maersk Institute for Production Technology
Odense University, Denmark

Abstract. In these days of multimodal systems and interfaces, many research teams are investigating the purposes for which novel combinations of modalities can be used. It is easy to forget that we still lack solid foundations for evaluating the functionality of individual families of input/output modalities, such as the speech modalities. The reason why these foundations are missing is the complexity of the problem. Based on the study of particular applications, empirical investigations of speech functionality address points in a vast multi-dimensional design space. At best, solid findings yield low-level generalisations which can be used by designers developing almost identical applications. Furthermore, the conceptual and theoretical apparatus needed to describe these findings in a principled way is largely missing. This paper argues that a shift in perspective can help address issues of modality choice both scientifically and in design practice. Instead of empirically focusing on fragments of the virtually infinite combinatorics of tasks, environments, performance parameters, user groups, cognitive properties etc., the problem of modality functionality is addressed as a problem of choosing between modalities which have very different properties with respect to the representation and exchange of information between user and system. Based on a study of 120 claims on speech functionality from the literature, it is shown that a small set of modality properties are surprisingly powerful in justifying, supporting and correcting the claims set. The paper analyses why modality properties can be used for these purposes and argues that their power could be made available to systems and interface designers who have to make modality choices during early design of speech-related systems and interfaces. Using hypertext, it is illustrated how this power may be harnessed for the purpose of predictively supporting speech modality choice during early systems and interface design.

1. Introduction

Use of speech input to, and speech output from, computer systems is spreading at a growing pace. This means that an increasing number of designers and developers of systems and interfaces are faced with the question of whether to use speech input and/or speech output for the applications they are about to build. The literature offers no systematic guidance on this issue although there is consensus in the field that systematic guidance is highly desirable [2]. Systematic guidance requires theory but theory alone is not sufficient. Once developed, theory must be transformed into practically useful methods or tools which can be applied by non-theoreticians. This paper addresses the issue of theory with the ultimate aim of providing a tool that may assist developers of systems and interfaces in deciding when to use speech in their applications.

A theory of speech functionality must prove its worth in supporting or criticising claims about speech functionality. For this purpose, representative data consisting of claims about speech functionality is required. The paper investigates the possibility of theoretically justifying a large number of claims which have been made for or against the use of speech output and/or speech input to represent information in the information exchange which takes place between user and system in human-computer interaction. The data includes 120 different claims on speech functionality. It is argued that previous efforts to address the functional question of whether to use, or not to use, speech for a certain system and interface design task have been based on insufficient and fragile experimental work, user testing, common sense hypothesising or designer experience from trial-and-error. The scope of the concepts, or domain va-

riables, that have been used to express hypotheses or findings, such as 'task type' or 'user population type', is largely ad hoc and differs from author to author. Moreover, it is argued that an empirical research agenda which would suffice to answer the question of speech functionality is virtually endless given the large number of domain variables involved, and that no scientific anchoring of most of the conceptual structures used so far is in sight. These observations suggest that a change in theoretical perspective might be desirable.

It is proposed to address the question of speech functionality from the point of view of information representation. In the domain of information representation, a theory called modality theory is already in place. Modality theory turns out to have strong justificatory power with respect to the examined claims. This suggests the following hypothesis: the making, during early design, of reliable hypotheses about the suitability or unsuitability of one or more speech modalities for the system and interface design task at hand can often be done based on understanding of the information representation properties of a limited set of input/output modalities. In addition to understanding the relevant modality properties, designers should of course understand their design task, which includes understanding of how the relevant domain variables are instantiated in the design space defined by the design task. If the present data is representative, at least 3 in 4 design decisions concerning speech functionality do not require empirical experimentation, user testing, common sense hypothesising or designer trial-and-error, nor do these decisions require elaborate and as yet non-existent taxonomies and theories of the many domain variables involved in standard requirement specifications. If the above hypothesis is correct, modality theory helps address a problem for the solution of which no other viable approach is in sight. The power of modality theory suggests that the theory might be applied in developing a practically useful tool for advising developers of systems and interfaces on when to use speech input/output. The paper concludes by illustrating a tool which applies modality theory in support of early systems and interface design.

In what follows, Section 2 presents arguments in favour of a change in theoretical perspective. Section 3 briefly presents modality theory. Section 4 introduces the data, describes the method of data transformation and analysis, and presents the modality properties whose justificatory and predictive power will be investigated. Section 5 presents an analysis of the data. Section 6 concludes by illustrating the practical application of results. Appendix 1 presents the data. Appendix 2 presents the scope of the human-computer interaction variables used in presenting the data, as derived from the data.

2. Justifying Speech Functionality

In the terminology of the present paper, 'speech' or 'spoken language' designates several different unimodal input or output modalities. A modality is simply a form (or mode) of representing information as output from, or input to, a computer system [21]. A unimodal modality is a modality which may form a component of a multimodal interface [4]. Some unimodal modalities, including the speech modalities, are perfectly capable of being used alone in exchanging information with computer systems. It is trivial to argue that speech is not suited for every kind of human-computer information exchange, and it is an equally trivial generalisation that, in some cases, other modalities are preferable to speech if we want to optimise the human-computer interface from the point of view of information exchange. On the other hand, sometimes speech actually is suited to the system and interface design task at hand and sometimes speech is preferable to other modalities as well. The hard question is: in which specific cases are these generalisations true? This question not only is a hard one to answer in a principled way; a principled answer might also bring important benefits to systems and interface design practice by removing some of

the uncertainties which presently characterise the choice of speech modalities for the design of particular artefacts.

[combined speech input/output, speech output, or speech input modalities M1 and/or M2 and/or M3 and/or M4 etc.] are
[useful or not useful] for
[**generic task** GT and/or **speech act type** SA and/or **user group** UG and/or **interaction mode** IM and/**or work environment** WE and/or **generic system** GS and/or **performance parameter** PP and/or **learning parameter** LP and/or **cognitive property** CP] and/or
[preferable or non-preferable] to [alternative modalities AM1, AM2, AM3, AM4 etc.] and/or
[useful on conditions] C1, C2, C3 and/or C4 etc.

Figure 1. The minimum complexity of the problem of accounting for the functionality of speech in systems and interface design. Domain variables are in boldface.

Figure 1 shows why the question of when to use, and when not to use, speech in interface design is a hard one. This is because of its underlying complexity which derives from the large number of domain variables involved. Based on the data to be examined later, the minimum complexity of the problem has been expressed semi-formally in Figure 1.

When referring to the 'domain variables' in the speech functionality problem in the remainder of this paper, focus will be on the variables 'generic task' ... 'cognitive property' in Figure 1 rather than on the complete inventory of variables including modalities and conditions. The domain variables are defined through their extensions (or scope) in Appendix 2. Modalities will be at centre stage throughout and no attempt has been made to systematise the conditions referred to in the data. Note that, unless factors could be trivially constrained or eliminated, Figure 1 expresses the minimum complexity of the problem. There may be more relevant domain variables involved than those found in the data and expressed in the data presentation in Appendix 1. 'Privacy protection', for instance, might be such a domain variable. Although present in the data, 'privacy protection' has not been singled out as a domain variable in the data presentation in Appendix 1. Similarly, Figure 1 does not distinguish among the different technologies so familiar to the speech community, such as isolated words input, connected speech input, continuous speech input, speaker-dependent speech input, speaker-independent speech input, synthetic speech output and concatenated speech output. These distinctions are being made so infrequently in the data to be examined that they have been omitted from the expression in Figure 1.

Yet the complexity expressed in Figure 1 is huge. If, in order to solve the problem of speech functionality, we were to empirically investigate each and every possible combination of the domain variables in Figure 1 (cf. Appendix 2), then (a) we would never finish the task, and (b) we would certainly never finish the task in time to be able to provide much needed support of modality choice in early systems and interface design. At best, we would end up with a very large, albeit still incomplete, number of low-level generalisations based on having made all possible mistakes at least once. The generalisations in question would look more or less like the following from the data:

Data point 11 (Table 2, Appendix 1). **Speech act** [instruction] + **generic task** [follow procedure, e.g. for using a video recorder or an ATM]: speech output can be useful.
Assumption: The procedures to be followed require limb and visual activity.

Data point 121 (Table 10, Appendix 1). **Generic multimedia systems** [office] + **generic task** [speech input text editing + keyboard text entry]: speech input has no advantage in **performance parameters** [speed, accuracy, and ease of use].

Figure 2. Examples of low-level generalisations on speech functionality for systems and interface design. Domain variables are in boldface.

Data point 11 in Figure 2 should be read as follows: speech output can be useful for providing instructions to users during their performance of procedural tasks requiring limb and visual activity. The relevant text in [1] says: "An increasingly popular use of speech output is in providing instructions for users of sophisticated technology, such as video casette recorders or ATMs. It has proved a popular application in toys, and can provide information on how to use a product". Data point 121 in Figure 2 should be read thus: speech input has no advantage in speed, accuracy, and ease of use when used for text editing in keyboard-based office text-entry systems. The relevant text in [30] says: "One solution to overcome the limitations inherent in the recognizer technology has been the suggestion of a multi-modal system for office applications. For example, all editing and formatting commands should be given verbally, while text is entered via the keyboard. ... but the benefits to the users are not immediately obvious. When considering the indices of speed, accuracy, ease of use, there appears to be little advantage in introducing speech to the system". Note that claim (11) in Figure 2 is only valid provided that at least one unstated assumption is being added. Lacking theoretical foundations, the claims about speech functionality made in the literature cannot be expected to be fully explicit nor to incorporate all the assumptions necessary. In order not to unnecessarily complicate data presentation and analysis, such assumptions have been added sparingly below. This means that more assumptions might have to be added to ensure the validity of a certain claim.

However, the problem of speech functionality is even harder than presented in Figure 1 and illustrated in Figure 2. It is a well known fact that decades of effort in human-computer interaction (HCI) and linguistic research has failed to produce principled, stable, general and generally accepted taxonomies of the domains of the following variables: 'generic task', 'speech act type', 'user population', 'interaction mode', 'work environment', 'generic system', 'performance parameter', 'learning parameter' and 'cognitive property'. Arguably, work on speech acts has made the most progress since Searle's seminal yet much too abstract taxonomy [31]. Today's spoken language dialogue designers and theoreticians still work towards integrating and extending the many existing, more detailed speech acts taxonomies in the absence of global consensus on speech acts annotation in spoken dialogue [12]. This implies that, were we to continue generating low-level generalisations, such as those exemplified in Figure 2, we would in most cases not know how to state them in any systematic fashion. Each researcher would have to invent systematic ways of expressing them, just as the author has done with respect to the data points in Figure 2 and Appendix 1 (surveyed in Appendix 2), and as some of the data points authors did prior to the present study.

It seems clear that in the situation just outlined, there is a need for a theory-based approach which could help answer the design question: "Should one or several of the spoken language modalities be considered for systems and interface design task SID(tx) or not?". There is, moreover, an important reason why we should expect that the theory-based approach in question will be found in modality theory and not primarily from among the many approaches which seek to create taxonomies for individual domain variables such as 'generic task', 'user population' or 'work environment' (Figure 1). The reason is apparent from the examples in Figure 2. It is that actual claims about the functionality of speech do not uniformly include any single such domain variable. Appendix 1 demonstrates that specific claims about

the functionality of speech may involve none, one or several domain variables. Theoretical mastery of any single such variable, in other words, would not solve the problem. What the claims authors have done, obviously, is to abstract away, sometimes inccorrectly, as we shall see, from the other domain variables instantiated in the experiments or design solutions they have examined, focusing on relationships involving only some domain variables. In fact, the only variable shared by all claims about speech functionality is the variable 'modality'. On reflection, this is hardly surprising, of course, but even trivial points are sometimes easily ignored. This strongly suggests that, when seeking theory-based solutions to the speech functionality problem, attention might profitably be directed to the theory of modalities of information representation.

The basic idea of the proposed approach is the following. Suppose that the requirements specification for a specific systems and interface design task SID(t1) includes information to the effect that, e.g., the user needs hands and eyes free operation, and suppose that we already know that some modalities of information representation M1-Mn allow hands and eyes free operation. Together, these two pieces of information imply that the modalities M1-Mn can be suggested as potentially appropriate modalities for the system to be designed. The fact that, for instance, modality M1 allows hands and eyes free operation is called a modality property of M1. Knowledge of modality properties thus allows the systems developer to carry out a mapping from requirements specification information onto candidate modalities. Moreover, to do the mapping, the developer does not need any systematic knowledge of the domain variables in the form of as yet non-existent task taxonomies or taxonomies of performace parameters (cf. Appendix 2). The developer only needs the requirements specification information, however expressed, together with knowledge of the relevant modality properties. As we shall see in Section 6, this is already a lot to ask and even preliminary taxonomies of the domain variables may help but at least the developer needs not wait for breakthroughs in HCI to invoke theory-based support on when (not) to use speech in an application to be developed.

3. Modality Theory

For about four years, we have been developing an approach called modality theory which does not focus on taxonomies and theories of generic tasks, speech act types, user populations, interaction modes, work environments, generic systems, performance parameters, learning parameters or cognitive properties. Rather, modality theory focuses on the types (or modalities) of information to be exchanged between user and system during task performance. Modality theory has been developed for unimodal output modalities and work on input modalities is in progress. The theory provides a complete taxonomy of unimodal output modalities in the physical media of graphics, acoustics and haptics (or touch). The taxonomy is hierarchically organised at up to four levels of abstraction called the super level, the generic level, the atomic level and the sub-atomic level, respectively (Figure 3). The generic level has been systematically generated from the following small set of binary and ternary distinctions between basic properties characterising information representation: linguistic/non-linguistic, analogue/non-analogue, arbitrary/non-arbitrary, static/dynamic and graphics/acoustics/haptics. Space limitations do not allow discussion of these distinctions [see 4, 5]. The super level is of less theoretical significance as it merely represents one among several possible classifications of the modalities at the lower levels. Being in most cases too abstract or general for interface design purposes, the generic level has formed the basis of generation-by-decomposition of the atomic level through the addition of further sets of basic property distinctions. The process of unimodal modality generation from basic properties are guided by principles of completeness, orthogonality, relevance and intuitiveness of the generated results, i.e. the unimodal modalities [7]. The principles of relevance and intuitiveness sometimes warrant reduction of the number

of generated unimodal modalities. For instance, Figure 3 distinguishes at the generic level between static and dynamic graphic language using non-analogue signs (modalities 5 and 8) but does not explicitly distinguish between static and dynamic acoustic and haptic language using non-analogue signs (modalities 6 and 7). Since most acoustic language is dynamic and dynamic haptic output, such as dynamic Braille or dynamic haptic images, is still scarce for technological reasons, these distinctions are not needed at this point. The principle of completeness is not violated through these reductions because the relevant phenomena are still being represented in the modality documents (see below).

Generally speaking, the atomic level represents individual unimodal modalities at the level of abstraction at which interface designers are used to thinking about output modalities. In some cases, however, further generation-by-decomposition of the sub-atomic level from the atomic level has been necessary to achieve sufficiently fine-grained distinctions among modalities [7]. The properties of each unimodal modality are defined, analysed and illustrated in a hypertext/hypermedia 'modality document' in the Modality Theory Workbench and Demonstrator. Supporting theoretical concepts of modality theory are being separately defined, analysed and illustrated in 'lexicon documents' [8]. The modality and lexicon documents express modality theory proper whereas Figure 3 presents the underlying taxonomy. The Modality Theory Workbench and Demonstrator is currently being ported from OMNIS 7 into html. Based on the taxonomy and theory of unimodal output modalities, all possible multimodal representations in the media of graphics, acoustics and haptics can be generated by composition from unimodal modalities or analysed by decomposition into their component unimodal modalities.

Modality theory will be applied in the following way in this paper. The speech output modalities are numbered 2 and 6 at the generic level of the taxonomy in Figure 3, and 6a, 6b and 6c at the atomic level. Modality 2 is static and dynamic acoustic language using analogue or 'iconographic' signs, such as hieroglyphs. Modality 6 is static and dynamic acoustic language using non-analogue signs, i.e. most of ordinary speech. Modality 2 can be ignored for technical reasons: the relatively few recognisable analogue signs, or 'onomatopoetica', in spoken language have been included in modality 6. This allows us to ignore modality 2 for interface design support purposes, which is why modality 2 does not have any daughter nodes in Figure 3 [4]. This is another example of pragmatic reduction in the number of unimodal modalities discussed in the preceeding paragraph. Based on modality 6, modalities 6a, 6b and 6c have been generated-by-decomposition through introduction of the distinction, which is not unique to speech, between discourse, labels/keywords and notation. Spoken discourse is the basic form of spoken language. Discourse is different from text because the former is basically situation-dependent whereas the latter is basically situation-independent. This may account for many of the differences between, i.a. spoken discourse and written text, such as the differences in grammar and the particular rhetorical potential of spoken discourse [7]. Spoken labels/keywords represent the use of single phrases to convey information, such as in spoken 'earcons'. Spoken notation is the use of spoken language to express any kind of linguistic notation [7]. Note, finally, that the atomic-level modalities 8d-8f are graphical modalities which are most commonly used in the form of the graphical speaking faces that help listeners decode synthetic speech output.

The hierarchical nature of the taxonomy means that the properties of a particular unimodal modality at a certain level of abstraction are inherited by all that modality's daughter nodes and by their daughter nodes etc. For instance, if linguistic modalities share a certain property which in this case is being expressed at the super level, then the speech modalities at the generic and atomic levels of the taxonomy will inherit that property because speech constitutes a sub-set of the linguistic modalities. The implication is that a justification of why a certain speech modality may, e.g., be recommended for a specific in-

terface design task does not have to derive from a property which is peculiar to speech but may well derive from the fact that the speech modality has inherited that property from higher up in the taxonomy. In Sections 4 and 5 below, 18 particular modality properties will be used in justifying or supporting most of the examined data (see Figure 4). Only 3 of these properties are characteristics of

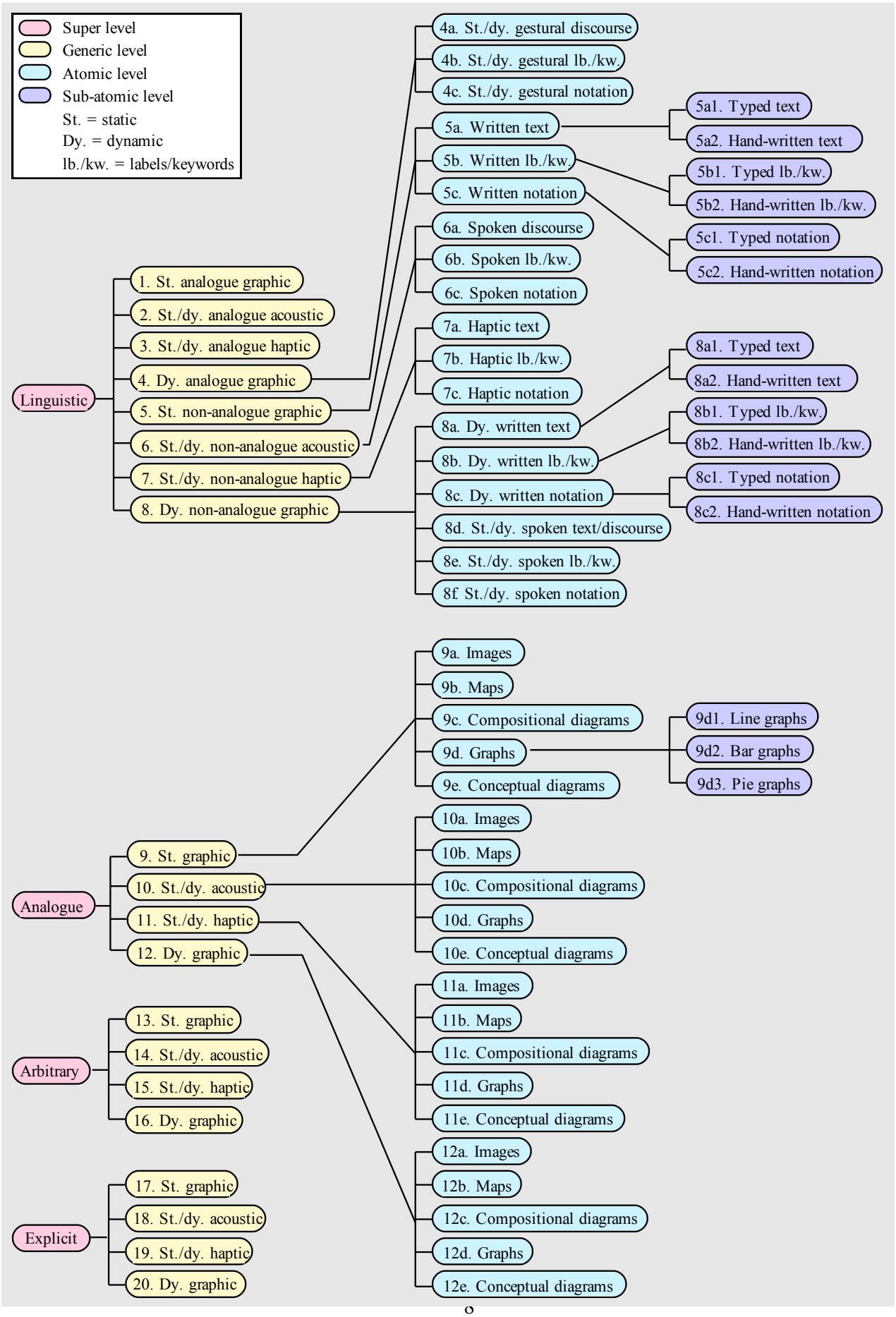
Super level
Generic level
Atomic level
Sub-atomic level
St. = static
Dy. = dynamic
lb./kw. = labels/keywords
Linguistic
1. St. analogue graphic
2. St./dy. analogue acoustic
3. St./dy. analogue haptic
4. Dy. analogue graphic
5. St. non-analogue graphic
6. St./dy. non-analogue acoustic
7. St./dy. non-analogue haptic
8. Dy. non-analogue graphic
4a. St./dy. gestural discourse
4b. St./dy. gestural lb./kw.
4c. St./dy. gestural notation
5a. Written text
5b. Written lb./kw.
5c. Written notation
5a1. Typed text
5a2. Hand-written text
5b1. Typed lb./kw.
5b2. Hand-written lb./kw.
5c1. Typed notation
5c2. Hand-written notation
6a. Spoken discourse
6b. Spoken lb./kw.
6c. Spoken notation
7a. Haptic text
7b. Haptic lb./kw.
7c. Haptic notation
8a. Dy. written text
8b. Dy. written lb./kw.
8c. Dy. written notation
8a1. Typed text
8a2. Hand-written text
8b1. Typed lb./kw.
8b2. Hand-written lb./kw.
8c1. Typed notation
8c2. Hand-written notation
8d. St./dy. spoken text/discourse
8e. St./dy. spoken lb./kw.
8f. St./dy. spoken notation
Analogue
9. St. graphic
10. St./dy. acoustic
11. St./dy. haptic
12. Dy. graphic
9a. Images
9b. Maps
9c. Compositional diagrams
9d. Graphs
9e. Conceptual diagrams
9d1. Line graphs
9d2. Bar graphs
9d3. Pie graphs
10a. Images
10b. Maps
10c. Compositional diagrams
10d. Graphs
10e. Conceptual diagrams
11a. Images
11b. Maps
11c. Compositional diagrams
11d. Graphs
11e. Conceptual diagrams
12a. Images
12b. Maps
12c. Compositional diagrams
12d. Graphs
12e. Conceptual diagrams
Arbitrary
13. St. graphic
14. St./dy. acoustic
15. St./dy. haptic
16. Dy. graphic
Explicit
17. St. graphic
18. St./dy. acoustic
19. St./dy. haptic
20. Dy. graphic

Figure 3. The taxonomy of unimodal output modalities. The four levels are, from left to right: super level, generic level, atomic level and sub-atomic level.

speech as such. The remaining properties have been derived either from higher up in the taxonomy or only belong to some, but not all, of the speech modalities. In other words, the problem of speech functionality cannot be solved through appeal to properties which are characteristic of all and only the speech modalities.

To properly understand the relationship between the taxonomy of unimodal output modalities (Figure 3) and the list of modality properties used in justifying claims about speech functionality (Figure 4), the reader should note that there is no one-to-one correspondence between the unimodal modalities distinguished in Figure 3 and the individual modality properties listed in Figure 4. To be sure, the correspondence does hold in some cases. For instance, modality property 1 (MP1) in Figure 4 refers to the super-level linguistic modality in Figure 3. Strictly speaking, however, one-to-one correspondence only holds for MPs 1-3 and MP12. For instance, MP4 deals with acoustic input/output but Figure 3 does not show an 'acoustic' modality or, rather, does not show the acoustic medium independently of the modalities presented in that medium. The reasons for this lack of one-to-one correspondence are twofold. First, the taxonomy in Figure 3 only addresses output modalities whereas some of the modality properties in Figure 4 characterise input modalities (see below). Secondly, each unimodal modality in Figure 3 is uniquely defined from a set of basic properties. Modality theory, as distinct from the taxonomy in Figure 3 and as expressed in the modality documents (see above), analyses all these properties in order to provide a comprehensive understanding of the modalities in the taxonomy. It is these modality property analyses which are being used in Figure 4. Modality theory thus includes a large number of generalisations of the form "Acoustic output is ..." or "Discourse output is ..." which go into the understanding of individual unimodal modalities but which are only visible at taxonomy level in so far as the terms "acoustics" or "discourse" are being used to characterise individual unimodal modalities.

Some of the modality properties to be used in justifications in Section 5 are properties of both speech output and speech input. Modality theory has not yet been fully developed for the representation of input information, as will be revealed through its incapability of justifying some of the data points. However, the input modality properties used in data justification below are properties which are common to input and output. These properties already form part of input modality theory.

4. Data, Method and Modality Properties

4.1 Data Selection and Global Data Structuring

The 120 claims on speech functionality to be analysed below have been systematically gathered from a collection of papers dedicated to the issue of speech functionality [2]. Redundant claims have been excluded because of the limited statistical value of the data. All non-redundant claims of the following forms have been included: recommendations for or against combined speech input/output, speech output or speech input; recommendations for or against combined speech input/output, speech output or speech input as compared to the use of non-speech modalities or speech/non-speech combinations; and conditional claims on the use of speech, i.e. claims which state that if some speech modality is to be used, then it should be used subject to the condition that, e.g., headphones are used in public spaces to protect privacy. Claims of the following form were excluded from consideration: claims comparing speech/non-speech multimodal combinations with other speech/non-speech multimodal combinations. The claims are presented in 11 tables in Appendix 1 showing 11 different types of claim about speech functionality, as follows:

Table 1. Claims recommending combined speech input/output.
Table 2. Claims recommending speech output.
Table 3. Claims positively comparing speech output to other modalities.
Table 4. Claims recommending speech input.
Table 5. Claims positively comparing speech input to other modalities.
Table 6. Conditional claims on the use of speech.
Table 7. Claims negatively comparing combined speech input/output to other modalities.
Table 8. Recommendations against the use of speech output.
Table 9. Claims negatively comparing speech output to other modalities.
Table 10. Recommendations against the use of speech input.
Table 11. Claims negatively comparing speech input to other modalities.

The idea has been to present the claims following a simple and intuitive ordering principle suggested by the claims themselves, which might make the claims overviews themselves useful for various purposes. Of the 13 possible types of claim, two were not found among the data, i.e. claims positively comparing combined speech input/output to other modalities and recommendations against the use of combined speech input/output. It is hypothesised that this does not significantly reduce the representativeness of the data set.

4.2 Data Standardisation

It is important to bear in mind in what follows that this paper deals with very complex data (cf. Figure 1) which, moreover, have been extracted from their context. The purpose of data representation has been to express the data in a comparable and intelligible format which preserves the basic point made by their authors. The purpose has not been to (a) co-represent the full context of each data point; nor to (b) make each data point fully explicit with respect to its implicit assumptions; nor to (c) create a fully formalisable representation. (c) would probably be beyond current state-of-the-art, and (a) and (b) would have meant producing lengthy and partly speculative renderings of the data, which would conspire to defeat the practical aims of the analysis and discussion in what follows. The data, as rendered, therefore remain partially "messy".

For the purposes of this paper, each of the 120 data points expresses a single claim about the functionality of speech. This is one reason why it has been necessary to transform the claims as expressed in [2] into the standard format used in Tables 1 to 11. In the original papers, authors often cluster several claims together. A second reason for data standardisation is to achieve a maximum of brevity and succinctness. Considerable care has been taken to ensure that each standardised claim was semantically equivalent to the original claim. A colleague read the papers in [2] and critically reviewed the claims standardisations. Thirdly, the domain variables in Figure 1 have been used in expressing the standardised claims. The domain variables 'generic task', 'speech act', 'user group', 'interaction mode', 'work environment', 'generic system', 'performance parameter', 'learning parameter' and 'cognitive property' have been boldfaced in the tables (cf. Figure 2). This provides ease of access to the complexity of the speech functionality problem. Tables 1 to 11 demonstrate how lack of appropriate analytical frameworks often leads to claims which are not sufficiently specific or which have been scoped in unclear or ambiguous ways. Sometimes, assumptions about the intended scope of a certain claim have been added (cf. Figure 2). The amount of assumptions has been kept at a minimum, however. In other cases, a comment has been added, for instance in order to clarify a point made in the claim. Comments are otherwise of many different kinds. A particular class of comments deals with external devices. The reader

may have noticed that external devices have not been mentioned so far. Claims often appeal to input devices, such as 'keyboard', or to input notions such as "written input" which could be either typed keyboard input, hand-written input to the computer, or even hand-written input on paper in case the comparison is between a non-computerised task and computerisation of the task through use of speech input. From the point of view of modality theory, hand-writing and typing are different input modalities which are useful for inputting different types of information. In addition, the phrase "hand-writing on paper" is often intended to mean more than just that and to include the input of analogue information, for instance through the use of arrows from one place in a text to another, encirclings etc. Speech can more easily replace the hand-writing itself than it can replace the analogue spatial input, which is why such detailed distinctions between different forms of information have to be brought out when analysing particular claims. Similarly, the keyboard can be used for inputting different linguistic modalities, such as typed notation, typed labels/keywords and typed text as well as for inputting (rather cumbersomely) analogue spatial manipulation information. In all such cases, the comments aim to sort out the issues involved in a claim. Among the domain variables used in this paper, the variable closest to input devices is 'interaction mode'. This, admittedly ad hoc, domain variable subsumes add-on devices for transmitting acoustic input/output, such as wireless devices or headphones. These devices are unambiguous from a modality theory point of view.

Given the machinery just described and the additional fact that, necessarily, standardised claims have been presented out-of-context in the tables, it is important to emphasise that attribution of a claim to an author only indicates that the claim has been lifted from that author's writings. Attribution does not imply that the author can be assumed to hold, or to have held, the opinion about speech functionality expressed in the claim. In many cases, of course, the author does, or did, hold the opinion expressed in the claim. In other cases, however, the author might want to hedge the claim, contextualise the claim, or further specify the claim in ways that have not been made clear in the data representation below, or the author did merely quote an actual or hypothetical claim in order to make it the subject of critical analysis or empirical disconfirmation.

Based on the methodology of data standardisation described, it seems justified to hypothesise that the claims set is representative of standard approaches to the speech functionality issue. More precisely, representativeness is hypothesised with respect to (i) the sources of claims, such as experimental work, user testing, common sense hypothesising or designer experience; (ii) the variables involved in claims, such as 'speech input', 'generic task', 'cognitive property' or alternative 'modality' solutions; (iii) the types of claim, such as recommendation, negative comparison or conditional recommenda-

No	MODALITY	MODALITY PROPERTY
MP1	Linguistic input/output	Linguistic input/output modalities have interpretational scope. They are therefore unsuited for specifying detailed information on spatial manipulation.
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.
MP7	Static graphics	Static graphic modalities allow the simultaneous representation of large amounts of information for free visual inspection.
MP8	Dynamic output	Dynamic 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.
MP10	Speech input/output	Speech input/output modalities, being temporal (serial and transient) and non-spatial, should be presented sequentially rather than in parallel.
MP11	Speech input/output	Speech input/output modalities in native or known languages have very high saliency.
MP12	Speech output	Speech output modalities may simplify graphic displays for ease of visual inspection.
MP13	Synthetic speech output	Synthetic speech output modalities, being less intelligible than natural speech output, increase cognitive processing load.
MP14	Non-spontaneous speech input	Non-spontaneous speech input modalities (isolated words, connected words) are unnatural and add cognitive processing load.
MP15	Discourse output	Discourse output modalities have strong rhetorical potential.
MP16	Discourse input/output	Discourse input/output modalities are situation-dependent.
MP17	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). (Note that spontaneous keywords must be distinguished from designer-designed keywords which are not necessarily natural to the actual users.)
MP18	Notational input/output	Notational input/output modalities impose a learning overhead which increases with the number of items to be learned.

Figure 4. The 18 modality properties (MPs) used in justifying, supporting or correcting the claims presented in Tables 1-11 in Appendix 1. The underlining of terms in column 2 indicates hypertext links and is explained in Section 6.

tion; (iv) the epistemic modifiers involved in claims, such as "may be preferable to", "use this modality for", "was perceived to be useful" or "is dubious compared to"; and (v) the (sometimes insufficient) scoping and level of precision of claims. Many more claims could have been added from investigations into speech functionality done since 1993. As argued already (Section 1), however, adding claims on speech functionality made in the literature between 1993 and 1996 would primarily produce a larger data set. It would not produce anything qualitatively different from what we already have, such as an exhaustive set of data on speech functionality covering all possible combinations of all relevant domain variables. The complexity of the problem is just too large to be exhausted through low-level generalisations on tasks, users, environments etc.

4.3 Data Justification

When the claims had been standardised, modality theory was searched for properties of modalities which might serve to justify the claims under investigation. The result was 18 such properties, each of which would serve to justify at least one claim, sometimes in conjunction with other modality properties. In some cases, although no modality property was found which could fully justify a certain claim, that property could nevertheless support the claim to a greater or lesser extent. In other cases, neither justification nor support could be found in modality theory for a certain claim which would therefore be marked as one for which no justification had been found. Not surprisingly, claims from any of those

three categories might sometimes be in partial of full conflict with modality theory. In such cases, correction was introduced to the claim in question based on reference to a specific modality property. The cases of justification, support, no justification and correction will be examined in Section 5, thereby providing contextualised versions of these notions. The notion of (full) 'justifica-tion' used in the data analysis may be explained as follows. It amounts to the claim that, given a set of modality properties and a specific claim on speech functionality, a designer is practically justified in making that claim on speech functionality based on that set of modality properties. In other words, armed with the modality properties, the designer would in principle be able to make the claim even without the benefit of the particular source of the claim (Section 4.2). Inevitably, some of the claims in the data turned out to be expressed at the level of modality theory itself. When considered true, these claims have been included among those fully justified by modality theory.

The modality properties (MPs) used in claims justification are presented in Figure 4. Without going into unnecessary theoretical detail, a few words of explanation follow. In MP1, 'interpretational scope' refers to a basic limitation in the expressiveness of linguistic modalities compared to analogue modalities [6]. For present purposes, MP1 may be reduced to its second half which deals with spatial information. In MP3, 'arbitrary' modalities are representations whose meaning has been decided on ad hoc, such as the ad hoc introduction of particular sounds in acoustic alarms. In MPs 7 and 8, 'freedom of visual/perceptual inspection' means that the user has all the time desired to decode particular representations. This is true of, e.g., static graphic representations. However, as stated in MP9, even dynamic acoustic representations can be made interactively static by replaying them. In MPs 15 and 16, 'discourse' means the basic form of free speech exchange which is situation-dependent and rhetorical. From the point of view of the user, discourse output is preferable to 'spoken labels/keywords' output in the sense that, being free-form and unconstrained in length, discourse can remove the ambiguities which we often encounter in labels or keywords, be they spoken, graphic or haptic. Similarly, discourse input is preferable to (designer-determined) spoken labels/keywords input in the sense that users do not have to remember the particular keywords they must use in order to make their speech application execute. Discourse is also preferable to spoken 'notation' (MP18) in the sense that notation, being an add-on to natural language rather than a part of it, imposes an additional learning overhead which may not be appropriate in all applications. Spoken input through a fixed (or designer-determined) set of keywords imposes a similar learning overhead to that of spoken input notation. It should be emphasised that the modality properties MP1-MP18 are simply those that were required to justify as many of the claims as possible. The set of modality properties does not have any form of theoretical closure and modality theory could have provided more, or other, properties had the data been different.

5. Analysis of the Data

The data points are presented in standard form in Tables 1-11 in Appendix 1. Each claim has been evaluated from the point of view of the modality properties presented in Figure 4. For each claim is indicated in the tables whether the claim is justified, supported or left unexplained by the modality properties as well as whether the modality properties imply corrections to the claim. Of the 120 data points, 91 were justified by reference to one or more of the modality properties listed in Figure 4; 15 were supported by modality properties; and no justification was found in 14 cases. Corrections by reference to modality properties were made in 9 cases.

5.1 Justification

The fact that 3 out of 4 claims on speech functionality could be justified through reference to a small number of modality properties, suggests the following hypothesis: the making, during early design, of reliable claims about the suitability or unsuitability of one or more speech modalities for aspects of the system and interface design task at hand can often be done based on understanding of the information representation properties of a limited set of input/output modalities. In addition to understanding the relevant modality properties, designers should of course understand their design task, which includes understanding of how the relevant domain variables are instantiated in the design space defined by the design task. If the present data is representative, 3 in 4 design recommendations concerning speech functionality do not require empirical experimentation, user testing, common sense hypothesising or designer trial-and-error, nor do these decisions require elaborate and as yet non-existent taxonomies and theories of the many domain variables involved in standard requirement specifications (Figure 1). If the above hypothesis is correct, modality theory helps address a problem for the solution of which no other viable approach is in sight.

In interpreting this result, it should be borne in mind that the design recommendations on speech functionality in question, whether based on modality theory or on empirical methods and intuition, are not decisions to actually use speech in the design of a particular artefact. The latter decisions are 'holistic' or highly contextual, i.e. they must take into account all the peculiarities of the design space and the specified requirements, and often have to trade them off against one another. It is hard to believe that these decisions and trade-offs can be made the subject of explicit generalisations which uniquely determine the selection of particular modalities in context. It is therefore not surprising that no such decisions and generalisations were found in the data. Rather, the design recommendations for which modality theory, on the one hand, and empirical methods and intuition on the other, can provide justifications, are recommendations to consider to use speech, or not to use speech, or to consider to use speech rather than some alternative modality, or not to do so, or to consider using speech on certain stated conditions, given certain properties of the design space under consideration as characterised by the requirements specification. Such design recommendations are important in early design and development because they serve to constrain the design space with respect to the available candidate modalities for the design task. It follows that the recommendations may in principle be overridden by other design considerations, such as, to take a simple example, the absence of speech synthesisers in the machines to be used for an application for which synthetic speech would otherwise have been a good choice. In other words, predicting speech applicability, or predicting modality applicability in general, is always a 'ceteris paribus'-matter: if everything else equally favours the competing modalities, then use modality Mx because of its modality property MPy.

In what follows, justification from modality properties is illustrated through selected examples from Tables 1 through 11. The justified claims differ from one another in many respects. Figure 5 shows a "straight" justification from a single modality property, which does not need auxiliary assumptions (contrast Figure 2, data point 11). The claim addresses comparison between speech output and static text with respect to the domain variable 'cognitive property'. The cautious epistemic modifier "may be preferable to" is quite common in the data. Sometimes the modifier reflects the impossibility of unconditional prediction as discussed in the preceeding paragraph.

Data point 100 (Appendix 1, Table 3). Speech output may be preferable to static text for **cognitive property** [setting a mood]. Justified by MP15: "Discourse output modalities have strong rhetorical potential".

Figure 5. A straightforward justification of a comparative claim based on one modality property. Domain variables are in boldface.

Figure 6 shows a justification from a relatively large number of modality properties. The justification requires an auxiliary assumption concerning the domain variable 'work environment'. The claim addresses the use of speech output with respect to the variable 'speech act'. The epistemic modifier "use when" is taken in the sense of a recommendation rather than as the unconditional claim "always use". This is because the context of the claim is: "If you are using speech output for alarms, then [only use it for single alarms, not for multiple simultaneous alarms]." As one reviewer has pointed out, the recommendation to use speech is in this example a fairly strong one. This is because the conditions on the disrecommendations (MP8 and MP10) are more specific than the conditions on the recommendations (MP6 and MP11). The implementer is justified in concluding that speech alarms are a good idea (when people can be expected to hear them), unless there are a lot of them, which is a special case in which they're disfavoured.

Data point 37 (Appendix 1, Table 2). Output **speech act** [alarm]: use when single alarms.
Assumption: The claim deals with low-acoustic process control environments.
Justified on the stated assumption by MP10: "Speech input/output modalities, being temporal (serial and transient) and non-spatial, should be presented sequentially rather than in parallel." MP6: "Acoustic output modalities can be used to achieve saliency in low-acoustic environments." MP8: "Dynamic output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." MP11: "Speech input/output modalities in native or known languages have very high saliency."

Figure 6. A complex justification from several modality properties. Domain variables are in boldface.

Figure 7 shows justification of a claim which is at the same level as the justifying modality property. As in the modality properties presented in Figure 4, there are no auxiliary assumptions about other domain variables and no other domain variables are included in the claim. Note, however, that the justifying modality property has larger scope than the claim it justifies because the modality property concerns both speech input and speech output. The context from which it was taken shows that claim 9 is in fact a recommendation to use speech output. But note that the justifying modality property (MP4) which is being used in justifying the recommendatory claim 9, may just as well be used in justifying a corresponding, negative claim (see claim 6 in Table 8 of Appendix 1). This illustrates the 'holistic' nature of design decisions on whether to actually incorporate a specific modality in the designed artefact. Whether a modality property is good or bad, useful or harmful, depends on other variables of the design task, such as the work environment. A subtle reading of claim 9, therefore, is that it presupposes a generic task/work environment combination in which the omnidirectionality of speech may be beneficial rather than harmful.

Data point 9 (Appendix 1, Table 2). Speech output can be displayed to several simultaneously. Justified by MP4: "Acoustic input/output modalities are omnidirectional."

Figure 7. A claim which closely resembles the modality properties presented in Figure 4.

Figure 8 shows a justification of a negative claim concerning connected speech which includes a considerable number of domain variables: 'generic task', 'user group' and 'cognitive property'. This usually indicates that the claim represents a very low-level generalisation (see also Figure 2). MP14 provides

good reasons why office workers would feel better off not having to do large-text entry using connected words input.

Data point 120 (Appendix 1, Table 10). **Generic task** [large-text entry] + **user group** [office workers]: cannot be expected to **cognitive property** [accept] connected speech input. Justified by MP14: "Non-spontaneous speech input modalities (isolated words, connected words) are unnatural and add cognitive processing load."

Figure 8. A claim which involves a large number of domain variables (in boldface).

5.2 Support

The hypothesis that understanding of a limited set of modality properties suffices to support reasoning about a large number of design issues concerning modality choice, becomes re-inforced through consideration of the 29 cases which were not fully justified during analysis of the data. Let us look at these in turn, starting with the 15 cases of support.

Figure 9 shows a typical example of support of a claim which is complex in terms of the number of domain variables involved. The claim compares speech input with typed language input. The epistemic modifier "is likely to" is a rather strong one. MP17, therefore, cannot fully justify claim 119. Even though the users are non-expert typists, data-entry tasks differ widely along dimensions such as size of the data set and source of the data to be entered. Similarly, non-expert typists have different typing skills. It is not evident that the fact that speech is natural (in some modalities) generalises across all these differences to justify the claim that speech input will always be faster than typed language data entry.

Data point 119 (Appendix 1, Table 5). **Generic task** [data-entry] + **user group** [non-experts]: speech input is likely to be **performance parameter** [faster] than haptic [keyboard] modality. 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)."

Figure 9. A claim which is supported, but not justified, by modality theory. Domain variables are in boldface.

Table 1. 113: comparative empirical claim.
Table 2. 122: claim about actual use of speech modalities, not about properties relevant to modality preference.
Table 3. 98, 99, 102: tentative comparative empirical claims.
105: comparative empirical claim.
Table 4. 59: comparative empirical finding.
71, 117: very detailed recommendations.
Table 5. 64, 119: tentative comparative empirical claims.
90: comparative empirical finding.
Table 6. - Table 7. - Table 8. - Table 9. -
Table 11. 73, 84, 92: comparative empirical findings.

Figure 10. The nature and distribution of the 15 supported claims across the different types of claim presented in Tables 1 through 11. Numbers refer to data points.

Figure 10 shows that the support cases were distributed relatively evenly across the 11 different types of claim. However, 10 out of the 15 "merely supported" claims explicitly compare speech modalities to other modalities (Tables 3, 5 and 11). In addition, claims 113 (Table 1) and 59 (Table 4) are implicitly comparative, i.e. they compare speech modalities to unspecified alternative modalities. One claim (122 in Table 2) has a form which appears to make it inaccessible to full theoretical justification. Finally, two claims, i.e. 71 and 117 in Table 4 address very detailed and complex relationships among several variables. This suggests the following hypotheses:

(a) claims that compare speech and non-speech modalities in one or several respects are comparatively difficult to justify through reference to modality properties. On closer inspection, most of these claims turn out to be "holistic" ones, i.e. they claim that, considering all relevant domain variables, modality M1 is preferable to modality M2 in some respect. Such claims cannot be fully justified through generalisations (cf. Section 5.1);

(b) claims addressing very complex relationships among several variables are difficult to justify through appeal to modality properties. Figure 11 shows such a claim. The author claims (cf. also claim 70, Table 11) that speech input is slower than haptic (push-button) input. Speech input is therefore inferior to haptic input with respect to time-critical input. However, subtract the time-criticality but preserve the criticality and speech input becomes worth considering. It is difficult to envision a theory which could yield predictions based on distinctions of such subtlety. Yet MP5 supports the claim through reference to the need for heads-up orientation in critical situations in aviation.

Data point 71 (Appendix 1, Table 4). **Generic task** [non-time-critical but critical command input in aviation]: consider speech input. Supported by MP5: "Acoustic input/output modalities do not require limb (including haptic) or visual activity."

Figure 11. A claim about complex relationships. Domain variables are in boldface.

In fact, the two hypotheses (a) and (b) above both point to the difficulty of producing theoretical justifications of complex relationships. As long as the theories we use in justifying modality choice, whatever their nature, have limited complexity, there will always exist claims which are too complex to be justified by those theories. In such cases, empirical investigation is the only solution.

5.3 No Justification

The 14 cases of no justification are far from being evenly distributed across the different types of claim, as shown in Figure 12.

Table 1. - Table 2. - Table 3. - **Table 4.** 40: claim unclear. **Table 5.** 16, 79, 80: claims possibly false. 81, 82, 83: arguments questionable. 65, 68: tentative comparative empirical claims. Point to lacks in modality theory for input. 67: claim unclear. **Table 6.** 104: unclear point.

123: claim possibly false.
Table 7. - Table 8. - Table 9. - Table 10. -
Table 11.15: claim probably false.
70: comparative empirical claim. Points to lacks in modality theory for input.

Figure 12. The nature and distribution of claims for which no justification was found across the different types of claim presented in Tables 1 through 11. Numbers refer to data points.

The no justification cases are found mostly (9 cases) in Table 5 which positively compares speech input modalities to non-speech modalities. However, this is merely an anomaly in the data, due to the fact that one author began by listing a series of comparative claims favouring speech input modalities only to shoot those claims down later. Mostly, then, the cases of no justification are cases of unclear, questionable, possibly or probably false claims (11 cases). And of course, no sound theory can justify what is not the case. If we subtract these 11 claims from the claims set, we obtain a total of 109 claims which may deserve justification of support. The three interesting cases of no justification are claims 65 and 68 in Table 5 and 70 in Table 11. These cases should be justifiable by a fully developed theory of input modalities. Claims 65 and 68 address the limited accuracy of haptic input modalities through the keyboard. Claim 70 addresses the comparatively high speed of using push-buttons. Arguably, an important contribution of input modality theory should be to enable appropriate mappings from input information onto input devices, such as keyboards and push-buttons [9,11,25].

5.4 Correction

Except for one claims correction, the 9 cases of claims correction through reference to modality properties all address claims recommending speech, as shown in Figure 13. The exception is the sweepingly negative claim 15 about speech input in Table 11 which conflicts with a considerable number of justified claims. It appears that advocates of speech modalities are more vulnerable to criticism than their opponents. Four of the corrections derive from overlooking the possibility of using spoken notation (claims 113, 20, 19 and 39). Since spoken notation is a modality which has been generated by modality theory and which otherwise does not appear to be in widespread current use, the fact that it has been ignored is hardly surprising. Two claims (41 and 79) are too general considering the actual properties of speech. One claim (82) ignores an important property of static graphics.

Table 1. 113: speech notation overlooked.
Table 2. 20: speech notation overlooked.
Table 3.24: correct conclusion from wrong premises.
Table 4. 19, 39: speech notation overlooked.
41: too general claim given the properties of speech.
Table 5. 79: too general claim given the properties of speech (spatial manipulation).
82: dubious claim given the properties of static graphics.
Table 6. - Table 7. - Table 8. - Table 9. - Table 10. -
Table 11.15: sweepingly negative claim about speech.

Figure 13. The nature and distribution of claims corrected through reference to modality properties. Numbers refer to data points.

Perhaps the most interesting case of correction occurs with respect to claim 24 (Figure 14). In this complex argument, the assumption first states that 'acoustic non-speech' can be positively characterised as arbitrary acoustics, i.e. acoustics which bear arbitrary relationships to their ad hoc assigned meanings.

The correction then points out that the reason why arbitrary acoustics output are inferior to speech output is not that humans cannot discriminate between different types of arbitrary acoustic representations. Appreciation of music is probably dependent upon the ability to discriminate between large numbers of individually different sounds each of which does not carry any particular meaning. Rather, the inferiority of arbitrary acoustics is due to the learning overhead which is needed for humans to learn the meaning of many different arbitrary acoustic representations (MP3). Finally, MP17 argues why claim 24 is true after all. This is not because arbitrary acoustics cannot be made arbitrarily expressive of the meaning of different alarms but because speech output is "expressive for free" to humans. This example illustrates why we need a more firm and articulate theoretical background on which to think and reason about the properties of speech and other modalities.

Data point 24 (Appendix 1, Table 3). Many individual **speech acts** [warnings]: speech output is preferable to acoustic non-speech because of its expressiveness and **cognitive property** [human discrimination capacities]. *Assumption:* The acoustic non-speech referred to is the arbitrary acoustic modality. *Correction:* humans would appear able to discriminate between hundreds of sounds. Rather, S-O is preferable to acoustic non-speech if the latter is an arbitrary modality. Corrected by MP3: "Arbitrary input/output modalities impose a learning overhead which increases with the number of arbitrary items to be learned." 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).

Figure 14. A corrected claim about speech output. Domain variables are in boldface.

5.5 Assumption and Comment

A total of 22 explicit assumptions were made in the course of evaluating the claims. As said already, this number is considerably lower than what would have been needed to create 120 fully explicit claims. Claims were commented on in 33 cases. The comments serve purposes of clarification and criticism (cf. Section 4.2).

5.6 Roles of the Modality Properties

Figure 15 shows the respective roles of the 18 modality properties in justifying, supporting and correcting the claims. 7 modality properties were used more than 10 times for these purposes. Note that only one of these (MP11) expresses a property of speech input/output as such. Most of the frequently used modality properties express properties of super-level or generic-level modalities, such as linguistic input/output in general (MP1, MP2), acoustic input/output in general (MP4, MP5), or dynamic output in general (MP8). MP17 expresses a property of the most common speech input/output modalities.

If we look contents-wise at the roles of the most frequently used modality properties, the following picture emerges. Reasoning from linguistic input/output properties (MP1, MP2) mostly addressed the severe limitations of speech in communicating detailed spatial information. Reasoning from acoustic input/output properties (MP4, MP5) mostly addressed the omnidirectionality and limbs-free/eyes-free properties of speech. Reasoning from dynamic output properties (MP8) mostly addressed the limitations of speech output due to the limitations of human memory and attention. Reasoning from speech input/output properties (MP11) mostly addressed the particular saliency of spoken language. Finally, rea-

soning from properties of spontaneous spoken discourse and labels/keywords (MP17) mostly addressed the naturalness of speech.

MP	MODALITY	NO. OF CLAIMS ADDRESSED			
		Jst.	Sup.	Cor.	Total
1	**Linguistic input/output**	17	3	1	**21**
2	**Linguistic input/output**	16	3	1	**20**
3	**Arbitrary input/output**	1		1	**2**
4	**Acoustic input/output**	16	1	2	**19**
5	**Acoustic input/output**	27	5	1	**33**
6	**Acoustic output**	5			**5**
7	**Static graphics output**	7	1	1	**9**
8	**Dynamic output**	16	3		**19**
9	**Dynamic acoustic output**	2			**2**
10	**Speech input/output**	7			**7**
11	**Speech input/output**	13	1	1	**15**
12	**Speech output**	5	2		**7**
13	**Synthetic speech output**	1			**1**
14	**Non-spontaneous speech input**	3			**3**
15	**Discourse output**	5	2		**7**
16	**Discourse input/output**	4	2	1	**7**
17	**Spontaneous spoken labels/keywords and discourse input/output**	10	3	1	**14**
18	**Notational input/output**			4	**4**

Figure 15. The number of claims addressed, i.e. justified, supported or corrected, by each modality property (MP). The frequently used (>10 times) properties are MPs 1, 2, 4, 5, 8, 11 and 17.

The role of the "technical" speech modalities, such as synthetic speech output (MP13) or non-spontaneous speech input (MP14), is limited because the data did not address such properties very frequently. The quality of speech recognition and speech synthesis continues to rise, which makes it understandable that authors have refrained from discussion of moving targets. On the other hand, there is a need for information on how different recognition rates and qualities of synthetic speech affect the user acceptance of applications. Except for one modality property, the modality properties are used mostly for justification, less for support and still less for correction. This roughly corresponds to the overall proportion of justification, support and correction of the data (i.e., 91: 15: 9). The exception is MP18 which has been used solely for correction. This anomaly is due to the fact that we are not as used to thinking about spoken notation as we are to thinking about written notation in logic, mathematics, programming etc. Spoken notation was consistently overlooked in the data.

6. Concluding Discussion

This paper has examined a large number of claims about the functionality of spoken language modalities for interface design. The assumption has been that the claims set is representative of the way in which systems and interface designers and researchers in speech technologies and human-computer interaction reason about speech-related modality choice. The claims demonstrate that we are dealing with tremendously complex issues when reasoning about modality choice. Issues, moreover, which so far have not been adequately resolved conceptually, theoretically or empirically. Despite the complexity, however, at least one aspect of the problem domain remains constant as demonstrated in Tables 1 through 11 in Appendix 1: when reasoning about modality choice we always reason about properties of modalities for information representation and exchange. In a particular case of modality choice in early design we may not be reasoning explicitly about, e.g., generic tasks, user groups, performance parameters or cognitive issues but, by definition, as it were, we always reason about modality properties. This suggests taking a closer look at how knowledge about modality properties might be used to support reasoning about modality choice. From an already existing descriptive and taxonomic theory of modalities, 18 modality properties were selected based on their relevance to the reasoning about speech which was present in the claims set. This relatively small set of modality properties was shown to be sufficient to:

- justify modality recommendations in 91 out of the 109 cases which deserve justification;

- support modality recommendations in 15 out of the 18 remaining cases which deserve justification;

- correct claims in 9 cases.

The way this reasoning works is that knowledge of modality properties allows us to carry out a mapping from requirements specification information onto candidate modalities for the systems and interface development task at hand. In the 3 cases which deserved justification or support but which could neither be fully justified nor supported, the reason was found to be limitations in our current theory of input modalities. To these 3 cases may be added the 15 cases which could only be supported, but not fully justified, through reference to modality properties because of the sheer complexity of the relationships addressed in the claims. These 18 cases share the characteristic that they are too complex for modality theory as it stands.

The style of data presentation and analysis in the present paper has been “data presentation and analysis through approximation” rather than the rigorous and formalisable approach of the exact sciences. This has been necessary because of the complexity of the problem addressed as well as the nature of the data. A formalisable rendering of each data point would probably be beyond current state-of-the-art in addition to being considerably longer and more complex than the data presentations in Appendix 1. Nevertheless, the result is that an understanding of modality properties appears to be remarkably powerful in justifying and supporting particular modality recommendations to be made during systems and interface design. Moreover, similar power appears to be lacking in alternative approaches to this problem. The conclusion is that modality theory represents a principled and stable approach whose justificatory power should be made available to systems and interface designers who have to make modality choices during early design of speech-related systems and interfaces. This can be done independently of the lack of stable and comprehensive theoretical frameworks in other domains of HCI. Empirical studies will still be needed of high-complexity speech interaction problems, but indications are that application of modality properties will enable researchers and systems designers to use empirical methods more discriminatingly. This is desirable because of the difficulty of generalising empirical results more widely.

If this conclusion is true, the next question becomes that of how to actually use available knowledge of modality properties in practical systems and interface design. Justification, in and by itself, is not prediction. It has taken a significant amount of work to analyse the 120 claims, transform them into the standard format used in Tables 1 through 11 below and identify the relevant modality properties from modality theory. Systems and interface designers are not going to spend this amount of effort when addressing problems of modality choice in design practice. Nor are they likely to find the modality properties presented in Figure 4 sufficiently helpful. What these properties lack, as they stand, to be of help in design practice, seems first and foremost to be concretisation and illustration of their import with respect to practical design decision making. Underneath any generalisation lie a wealth of concrete instances, or practical cases, which by way of illustration can help suggest how to understand and apply the generalisation. In the present paper, such instances are found in Appendices 1 and 2. To mention an extremely simple example, MP4 states that acoustic input/output modalities are omnidirectional. MP4 justifies claim 9 (Table 2, Appendix 1) that speech output can be displayed to several people simultaneously. A link between MP4 and claim 9 might draw attention to the fact which might otherwise be overlooked, that the latter follows from the former.

What is needed, therefore, is a technology which can establish on-line links between the modality properties in Figure 4 and the concretisations and illustrations provided in the appendices. Hypertext is such a technology. To illustrate the idea for the purposes of this paper, hypertext links have been created between each modality property and the concrete claims concerning modality choice which that modality property serves to justify, support, fail to explain or correct. It is obvious that hypertext can do much more than that, such as maintaining links between each domain variable in the claims contained in Tables 1 through 11 and the domain variable concretisation lists in Appendix 2, or maintaining search links between a designer-specified set of domain variable instantiations derived from the design task at hand, and a list of claims and modality properties which address those instantiations in the claims set. However, pending further testing of the main conclusion of this paper, we propose at this point to merely illustrate how hypertext might help provide illustrations and concretisations of the modality properties.

A first priority in future work is to test the conclusions of this paper on a separate set of speech-related modality choice claims gathered from the 1993-1996 literature on the subject. This work is in progress. Secondly, we would like to attempt to develop full hypertext support of modality property illustration and concretisation and test the developed system in design practice. Thirdly, of course, and only if the two first steps show real promise, an obvious continuation of the work is to integrate the speech-related modalities examined in this paper into a more general interface design support system able to handle a wide selection of modality choice problems.

REFERENCES

[1] Baber, C.: Speech output. In Baber, C. and Noyes, J. (Eds.): *Interactive Speech Technology*. London: Taylor & Francis 1993, 21-24.

[2] Baber, C. and Noyes, J. (Eds.): *Interactive Speech Technology*. London: Taylor & Francis 1993.

[3] Baber, C., Stanton, N. A. and Stockley, A.: Can speech be used for alarm displays in 'process control' type tasks? *Behaviour & Information Technology* 11 (4), 1992, 216-26.

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

[5] Bernsen, N. O.: A revised generation of the taxonomy of output modalities. *Esprit Basic Research project AMODEUS-2 Working Paper* RP5-TM-WP11, 1994.

[6] Bernsen, N. O.: Why are Analogue Graphics and Natural Language both Needed in HCI? In F. Paterno (Ed.): *Interactive Systems: Design, Specification, and Verification.* Focus on Computer Graphics. Springer Verlag 1995, 235-51.

[7] Bernsen, N. O.: A reference model for output information in intelligent multimedia presentation systems. In G. P. Faconti and T. Rist (Eds.): *Proceedings of the ECAI '96 Workshop: Towards a standard Reference Model for Intelligent Multimedia Presentation Systems.* 12th European Conference on Artificial Intelligence, Budapest, August 1996. To appear in *Computer Standards and Interfaces* 1997.

[8] Bernsen, N. O. and Lu, S.: A software demonstrator of modality theory. In Bastide, R. and Palanque, P. (Eds.): *Design, Specification and Verification of Interactive Systems '95.* Springer Verlag 1995, 242-61.

[9] Bleser, T. W. & Sibert, J.: Toto: a tool for selecting interaction techniques. In *Proceedings of user interface software and technology (Snowbird, Utah, Oct. 1990)* . New York: ACM, 1990, 135-42.

[10] Brandetti, M. D'Orta, P., Ferretti, M. and Scarci, S.: Experiments on the usage of a voice activated text editor. *Proceedings of Speech '88,7th FASE Symposium,* Edinburgh, 1988, 1305-10.

[11] Card, S. K., MacKinlay, J. D. and Robertson, G. G.: A morphological analysis of the design space of input devices. *ACM Transactions on Information Systems* , 9 (2), 1991, 99-122.

[12] Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., and Anderson, A.: The coding of dialogue structure in a corpus. In J. A. Andernach, S. P. van de Burgt, and G. F. van der Hoeve (Eds.): *Proceedings of the Twente Workshop on Language Technology: Corpus-based approaches to dialogue modelling.* Universteit Twente, Enschede, The Netherlands, 1995, 25-34.

[13] Coler C. R.: In-flight testing of automatic speech recognition systems. *Speech Tech* '84 1 (1), 1984, 95-8.

[14] Cowley, C. K., Miles, C. and Jones, D. M.: The incorporation of synthetic speech into the human-computer interface. *Contemporary Ergonomics.* London: Taylor and Francis 1990.

[15] Damper, R. I.: Speech as an interface medium: how can it best be used? In Baber, C. and Noyes, J. (Eds.): *Interactive Speech Technology*. London: Taylor & Francis 1993, 59-72.

[16] Gorden, D. F.: Voice recognition and systems activation for aircrew and weapon system interaction. 1989 (source unknown).

[17] Gould, J. W., Conti, J. and Hovanyecz, T.: Composing letters with a simulated listening typewriter. *Communications of the ACM,* 26, 1983, 295-308.

[18] Hapeshi, K.: Design guidelines for using speech in interactive multimedia systems. In Baber, C. and Noyes, J. (Eds.): *Interactive Speech Technology*. London: Taylor & Francis 1993, 177-88.

[19] Helander, M. G., Moody, T. S. and Joost, M. G.: Systems design for automated speech recognition. In M. G. Helander (Ed.): *Handbook of Human-Computer Interaction.* Amsterdam: Elsevier 1988.

[20] Helander, M. G.: Foreword. In Baber, C. and Noyes, J. (Eds.): *Interactive Speech Technology*. London: Taylor & Francis 1993, ix-xii.

[21] Hovy, E. and Arens, Y.: When is a picture worth a thousand words? Allocation of modalities in multimedia communication. Paper presented at the *AAAI Symposium on Human-Computer Interfaces*, Stanford 1990.

[22] Laycock, J. and Peckham, J. B.: Improving piloting performance whilst using direct voice input. *RAE Technical Report* 80019, 1980.

[23] Lee, K-F.: *Automatic Speech Recognition: the Development of the SPHINX System.* Dordrecht: Kluwer 1989.

[24] Lewis, E.: Interactive speech in computer-aided learning. In Baber, C. and Noyes, J. (Eds.): *Interactive Speech Technology*. London: Taylor & Francis 1993, 37-44.

[25] Mackinlay, J., Card, S. K. and Robertson, G. G.: A semantic analysis of the design space of input devices. *Human-Computer Interaction*, 5, 1990, 145-90.

[26] Martin, G. L.: Utility of speech input in user-computer interfaces. *International Journal of Man Machine Studies* 30 (4), 1989, 355-75.

[27] Murray, I. R., Newell, A. F., Arnott, J. L. and Cairns, A. Y.: Listening typewriters in use: some practical studies. In Baber, C. and Noyes, J. (Eds.): *Interactive Speech Technology*. London: Taylor & Francis 1993, 99-108.

[28] Nicholson, R. T.: *ACM Transactions on Office Automation Systems* 3 (3), 1985, 307-14.

[29] Noyes, J. M. and Frankish, C. F.: A review of speech recognition applications in the office. *Behaviour and Information Technology* 8 (6), 1989, 475-86.

[30] Noyes, J.: Speech technology in the future. In Baber, C. and Noyes, J. (Eds.): *Interactive Speech Technology*. London: Taylor & Francis 1993, 189-208.

[31] Searle, J. R.: A taxonomy of illocutionary acts. In J. R. Searle: *Expression and Meaning.* Cambridge: Cambridge University Press 1979, 1-29.

[32] Stanton, N.: Speech-based alarm displays. In Baber, C. and Noyes, J. (Eds.): *Interactive Speech Technology*. London: Taylor & Francis 1993, 45-56.

[33] Starr, A. F. C.: Is control by voice the right answer for the avionics environment? In Baber, C. and Noyes, J. (Eds.): *Interactive Speech Technology*. London: Taylor & Francis 1993, 85-98.

[34] Tucker, P. and Jones, D. M.: Voice as a medium for document annotation. In Baber, C. and Noyes, J. (Eds.): *Interactive Speech Technology*. London: Taylor & Francis 1993, 109-18.

[35] Usher, D. M.: Automatic speech recognition and mobile radio. In Baber, C. and Noyes, J. (Eds.): *Interactive Speech Technology*. London: Taylor & Francis 1993, 73-84.

[36] Van Nes, F. L.: Multimedia workstations for the office. IPO (Eindhoven) *Annual Progress Report* 23, 1988.

[37] Warner, N. and Harris, S.: Voice-controlled avionics programs, progress and prognosis. *Speech Tech* '84, 1 (1), 1984, 110-23.

[38] White, R. G. and Beckett, P.: Increased aircraft survivability using direct voice input. *Technical Memo* FS(F) 515, RAE, Farnborough, UK, 1983.

[39] White, R. W., Parks, D. L. and Smith, W. D.: Potential flight applications for voice recognition and synthesis systems. *Sixth AIAA/IEEE Digital Avionics System Conference,* 84-2661-CP, 1984.

Acknowledgements. Many thanks are due to Claire Dormann who did part of the original data entry, examined the claims standardisations and detected several problems in the modality property presentation. I am grateful to Laila Dybkjær who developed Figure 3 and provided many valuable comments and criticisms on the draft version of the paper. The four anonymous reviewers suggested numerous improvements on the original submission. I wish to express my gratitude for their tremendus effort.

APPENDIX 1. THE DATA

The data is presented in Tables 1-11 below. The tables should be interpreted as follows:

No (column heading) = data point number marked in [2]. Although the numbered data points run to 124, there are only 120 data points. No data points correspond to the numbers 5, 28, 49 and 107. The corresponding data points were deleted because of redundancies with already represented data points.

Rf (column heading) = reference to the author(s) who made or mentioned the claim made in that row.

gc (Rf column) = general claim (no particular author quoted) made in that row.

Modalities (column heading, comparative tables only) = Specifies the non-speech modality or -modalities to which some speech modality is being compared in the claim in that row.

Mo (column heading) = abbreviation for Modalities.

CLAIMS ... (column heading) = states the type of claim presented in a particular table. The claims type should be read as prefix to the actual claims made in this column.

Justification (column heading) = this column evaluates claims from the point of view of modality theory.

MP/MPs (justification column) = modality property/properties as numbered in Figure 4.

Jst. (justification column) = the claim in that row is being considered fully justified by reference to the MPs referred to in the Jst. cell.

Sup. (justification column) = the claim in that row is being considered to some extent supported by reference to the MPs referred to in the Sup. cell.

No jst. (justification column) = the claim in that row could not be justified by reference to any of the MPs in Figure 4.

Cor. (justification column) = the claim in that row is being considered corrected by reference to the MPs referred to in the Cor. cell. The correction is stated along with the claim in that row.

The body of (standardised) claims presented in the claims rows are self-explanatory. Assumptions and comments have been explained in Section 4.2 above. *S-I/O* means combined speech input/output, *S-O* means speech output and *S-I* means speech input. Note that the modality theory properties (MPs) only provide full (Jst.) or partial (Sup.) justifications of why speech modalities should be *considered* by interface designers. Proving that, in a specific design case, a particular speech modality is the *only* input and/or output modality that meets the stated requirements, demands a more complex argument for which, in most cases, the examined claims do not provide sufficient information.

No	Rf	CLAIMS RECOMMENDING COMBINED SPEECH INPUT/OUTPUT	Justification
1	20	**Generic task** [dual, hands and/or eyes occupied].	MP **5 Jst.**
3	29	**User group** [people who have difficulty using computers]. *Assumption:* For various reasons (computer illiteracy, visual deficiency, manipulation deficiency), these users have difficulty using standard GUI interfaces.	MPs **5,17 Jst.**
42	15	**Generic tasks** [hands/eyes busy]: Niche market S-I/O applications.	MP **5 Jst.**
43	15	**Generic tasks** [mobility is needed]: Niche market S-I/O applications.	MP **4 Jst.**
44	15	**Interaction mode** [computer access over the telephone]: Niche market S-I/O applications.	MP **5 Jst.**
45	15	**User group** [physically disabled users]: Niche market S-I/O applications. *Assumption:* These users are visually or mobility deficient.	MP **5 Jst.**
113	30	**Generic systems** [complex]: S-I/O reduces **learning parameter** [interaction training time]. *Correction:* Speech notation S-I/O modalities do not. Mastery of any kind of notation requires training. *Comment:* The claim is too holistic to be fully justified.	MP **17 Sup. + Cor.** by MP **18**
114	30	**Interaction mode** [radio link] + S-I/O offer increased **performance parameter** [user mobility]. *Comment:* The radio link part seems trivial.	MP **4,5 Jst.**
116	30	Special **user group** [the disabled]. *Assumption:* These users are visually or mobility disabled.	MP **5 Jst.**

Table 1. The 9 cases in which combined speech input/output is recommended for consideration. Full justification was provided in 8 cases, support in 1 case, and correction in one of these.

No	Rf	CLAIMS RECOMMENDING SPEECH OUTPUT	Justification
8	1	**Speech act** [warning] which requires **cognitive property** [attention-grabbing].	MP **11 Jst.**
9	1	S-O can be displayed to several simultaneously.	MP **4 Jst.**
11	1	**Speech act** [command] + **generic task** [follow procedure, e.g. for using a video recorder or an ATM]: S-O was perceived to be useful. *Assumption:* The procedures to follow require limb and visual activity.	MPs **4,5 Jst.**
12	24	Special **user groups** [pre-school children, the blind who have not learnt Braille].	MPs **5,17 Jst.**
18	gc	S-O is **cognitive property** [attention-catching].	MP **11 Jst.**
20	gc	S-O requires **learning parameter** [no learning overhead]. *Correction:* Speech notation S-I/O modalities do.	MP **17 Jst. + Cor.** by MP **18**
21	gc	S-O reduces visual clutter in graphic displays.	MP **12 Jst.**
23	gc	S-O has omnidirectionality.	MP **4 Jst.**
25	32	Output **speech act** [alarm]: consider S-O to reduce **cognitive property** [visual attention load].	MPs **5,12 Jst.**
35	32	Output **speech act** [alarm]: use when **performance parameter** [immediate response] is required. *Assumption:* The claim deals with low-acoustic environments.	MPs **6,8,11 Jst.**
36	32	Output **speech act** [alarm]: use when **performance parameter** [mobile operator]. *Assumption:* The claim deals with low-acoustic environments.	MP **4, Jst.**
37	32	Output **speech act** [alarm]: use when single alarms. *Assumption:* The claim deals with low-acoustic environments.	MPs **6,8,10,11 Jst.**
38	32	Output **speech act** [alarm]: use when **generic task** [serial fault management]. *Assumption:* The claim deals with low-acoustic environments.	MPs **6,8,10,11 Jst.**
56	33	**Generic task** [aviation control and data input]: S-O offers simplification of control panel.	MP **12 Jst.**
122	30	**Generic tasks** [guidance, warnings, instructions, read-aloud text, feedback]: *synthetic* S-O has been used. *Comment:* Support from claims 8,11,12,18,21,25 above. The claim is too holistic to be fully justified.	MPs **4,5,11,12, 17 Sup.**

Table 2. The 15 cases in which speech output is recommended for consideration. Full justification was provided in 14 cases, support in 1 case, and correction in one of these. Note that 122 mentions synthetic speech output. This is rare in the data.

No	Rf	Modalities	CLAIMS POSITIVELY COMPARING SPEECH OUTPUT TO OTHER MODALITIES	Justification
26	3	**S-O vs. static text**	**Generic task** [single fault alarm and management in process control]: prefer S-O over (interactively) static text: it **cognitive property** [catches attention] better. *Assumption:* The claim deals with low-acoustic environments.	MPs **6,11 Jst.**
98	34	static text	S-O may be preferable to static text for expressing low complexity information.	MP **8 Sup.**
99	34	static text	S-O may be preferable to static text for expressing complex information in which details need not **cognitive property** [be attended to].	MP **8 Sup.**
100	34	static text	S-O may be preferable to static text for **cognitive property** [setting a mood].	MP **15 Jst.**
101	34	static text	S-O: may be preferable to static text for **cognitive property** [persuasive] information.	MP **15 Jst.**
102	34	static text	**Generic system** [electronic multimedia document]: S-O may be **performance parameter** [more effective] than static text for introductory information and some types of comment.	MPs **15, 16 Sup.**
105	18	static text	**Generic task** [general instruction]: use S-O for short lists and text for longer lists.	MPs **7,8 Sup.**
24	32	**S-O vs. acoustic non-speech**	Many individual **speech acts** [warnings]: S-O is preferable to acoustic non-speech because of its expressiveness and **cognitive property** [human discrimination capacities]. *Assumption:* The acoustic non-speech referred to is the arbitrary acoustic modality. *Correction:* humans would appear able to discriminate between hundreds of sounds. Rather, S-O is preferable to acoustic non-speech if the latter is an arbitrary modality.	MP **17 Jst. + Cor.** by MP **3**

Table 3. The 8 cases of claims which positively compared speech output to other modalities. Full justification was provided in 4 cases and support in 4 cases. Correction was proposed in one of these cases.

No	Rf	CLAIMS RECOMMENDING SPEECH INPUT	Justification
19	gc	S-I requires **learning parameter** [no learning overhead]. *Correction:* Spoken notation S-I/O modalities do.	MP **17 Jst. + Cor.** by MP **18**
22	gc	S-I has omnidirectionality.	MP **4 Jst.**
39	23	S-I is **cognitive property** [natural] for humans. *Correction:* Only S-I using spontaneous spoken keywords and discourse S-I/O are natural for humans.	MP **17 Jst. + Cor.** by MP **18**
40	23	S-I is **performance parameter** [fast]. *Comment:* Unclear claim: faster than what, for whom, in which environments, for which tasks etc.?	**No jst.**
41	23	S-I is a **performance parameter** [location free] input modality. *Correction:* There are limits to how loudly one can shout. This is implicit in MP 4. Otherwise, mobile S-I input devices must be used.	MP **4 Jst. + Cor.** by MP **4**
46	15	**Generic task** [1-out-of N selection + N (vocabulary size) of the order of hundreds]: consider S-I.	MP **17 Jst.**
50	35	**Generic task** [process plant control]: S-I + **interaction mode** [wireless device] allow **performance parameter** [mobile control]. *Comment:* Depends on the task. The wireless device part seems trivial.	MP **5 Jst.**
51	33	**Generic task** [aviation control and data input]: consider S-I. *Assumption:* The tasks do not involve spatial manipulation.	MP **5 Jst.**
52	33	**Generic task** [aviation control and data input]: S-I offers **performance parameter** [hands and eyes free operation].	MP **5 Jst.**
53	33	**Generic task** [aviation control and data input]: S-I offers **cognitive property** [reduction in visual workload].	MPs **5,12 Jst.**
54	33	**Generic task** [aviation control and data input]: S-I offers **performance parameter** [ease of operation] in the workplace.	MPs **5,12 Jst.**
55	33	**Generic task** [aviation control and data input]: S-I offers **performance parameter** [increased control capabilities].	MP **5 Jst.**
57	16	**Generic task** [aviation control and data input]: empirical results: S-I is advantageous when pilot is **performance parameter** [busy]. *Assumption:* The pilot is busy doing heads-up work.	MP **5 Jst.**
58	16	**Generic task** [aviation control and data input]: empirical results: S-I is **performance parameter** [safer]. *Assumption:* Safety increases if the pilot can concentrate on the heads-up work.	MP **5 Jst.**
59	16	**Generic task** [aviation control and data input]: empirical results: S-I is **performance parameter** [easier].	MPs **5,12 Sup.**
60	39	**Generic task** [aviation control (tuning radios by identification, setting up landing system), data input (to navigation system) and queries (system status, operating a checklist) to system]: consider S-I. *Assumption:* The tasks do not involve spatial manipulation.	MP **5 Jst.**
63	37	**Generic task** [dual task aviation control]: empirical results: S-I resolved **performance parameter** [workload conflicts] in selected tasks. *Assumption:* The selected tasks required heads-up work.	MP **5 Jst.**
71	33	**Generic task** [non-time-critical but critical command input in aviation]: consider S-I.	MP **5 Sup.**
95	28	**Generic task** [text annotation]: S-I may be appropriate for short simple global comments about the document, e.g. handling instructions.	MPs**16,17 Jst.**
112	18	**Generic system** [multimedia]: S-I may be used for frequently used command and selection input.	MP **5 Jst.**
117	30	**Generic system** [inquiry-based] + **generic task** [quality control] having small vocabulary and/or rigid syntax: *speaker independent* S-I may be useful. *Comment:* To the speech community this claim would appear obvious.	MP **5 Sup.**
124	26	S-I may extend user efficiency by providing an extra input modality.	MPs **4,5 Jst.**

Table 4. The 22 cases in which speech input is recommended for consideration. Full justification was provided in 18 cases, support in 3, and no justification in 1 case. Correction was proposed in three of these cases. Note that 117 mentions speaker-independent speech input. This is rare in the data.

No	Rf	Modalities	CLAIMS POSITIVELY COMPARING SPEECH INPUT TO OTHER MODALITIES	Justification
16	26	**S-I vs. typed input**	**Generic task** [text entry] + **performance parameter** [efficiency]: prefer S-I to typed input. *Comment:* The suitability of S-I for text entry remains moot.	**No jst.**
79	34	**S-I vs. written input**	**Generic task** [text annotation]: S-I is preferable to written input because it is **performance parameter** [faster]. *Correction:* This claim may well be false if the annotation task includes spatial manipulation.	**No jst. + Cor. by** MPs **1,2**
80	34	written input	**Generic task** [text annotation]: S-I is preferable to written input because it **performance parameter** [avoids shorthand]. *Comment:* This claim may well be false.	**No jst.**
81	34	written input	**Generic task** [text annotation]: S-I is preferable to written input because it reduces **cognitive property** [reading load]. *Comment:* Is not obvious that reduced reading load is in itself an advantage. The reduction of reading load is trivial.	**No jst.**
82	34	written input	**Generic task** [text annotation]: S-I is preferable to written input because it leaves the page uncluttered. *Correction:* Is not obvious that reduced clutter is an advantage.	**No jst. + Cor.** by MP **7**
83	34	written input	**Generic task** [text annotation]: S-I is preferable to written input because it encourages **performance parameter** [elaboration]. *Comment:* Is not obvious that this is the case.	**No jst.**
89	34	written input	**Generic task** [text annotation]: empirical results: S-I is more personal and nuanced than written input.	MPs **15, 16 Jst.**
90	28	written input	**Generic task** [text annotation]: empirical results: S-I is suited only for short, personal, informal communications to a limited number of addressees. *Comment:* No justification was found for the "only" part of the claim.	MPs **15, 16 Sup.**
91	28	written input	**Generic task** [text annotation]: empirical results: S-I may supplement written input.	MPs **5, 15, 16 Jst.**
64	22	**S-I vs. keyboard**	**Generic task** [aviation data input]: S-I may be **performance parameter** [faster] than haptic [keyboard] modality. *Assumption:* In heads-up situations.	MP **5 Sup.**
65	22	keyboard	**Generic task** [aviation data input]: S-I may be **performance parameter** [more accurate] than haptic [keyboard] modality.	**No jst.**
66	22	keyboard	**Generic task** [heads up aviation control] + **cognitive property** [high workload]: empirical results: S-I gave **performance parameter** [better performance] than haptic [keyboard] modality.	MP **5 Jst.**
67	13	keyboard	**Speech act** [command] input: S-I may be **performance parameter** [faster] than haptic [keyboard] modality. *Comment:* Unclear: faster in which environments, for which tasks?	**No jst.**
68	13	keyboard	**Speech act** [command] input: S-I may be **performance parameter** [more accurate] than haptic [keyboard] modality.	**No jst.**
119	30	keyboard	**Generic task** [data-entry] + **user group** [non-experts]: S-I is likely to be **performance parameter** [faster] than haptic [keyboard] modality.	MP **17 Sup.**
62	38	**S-I vs. keyboard or touch**	**Generic task** [heads up aviation control]: empirical results: S-I is preferable to haptic [keyboard and touch] modalities because of reducing **cognitive property** [spatial and temporal distraction] of pilot from the task of flying the aircraft.	MP **5 Jst.**

Table 5. The 16 cases of claims which positively compared speech input to other modalities. Full justification was provided only in 4 cases, support in 3 cases, and no justification in 9 cases. Correction was proposed in 2 of these. Note that written input may well be hand-written input on paper.

No	Rf	Mo	CONDITIONAL CLAIMS ON THE USE OF SPEECH	Justification
4	15	**S-I/O**	There are unwarranted claims on the **cognitive property** [naturalness] of speech for HCI. *Comment:* MP17 provides a correct version of this claim.	MP **17 Jst.**
13	24	I/O	S-I/O: If meaning is not grasped, **performance parameter** [repetition] is needed. *Comment:* This is, in fact, true of all input.	MP **8 Jst.**
14	24	I/O	S-I/O is **cognitive property** [attention-catching]. This may require **interaction mode** [microphones and/or headphones] in some **work environments**.	MPs **4,11 Jst.**
7	1	**S-O**	S-O implies **cognitive property** [cognitive processing limitations] with respect to the amount of information that can be attended to and remembered.	MP **8 Jst.**
10	1	O	**Speech act** [advice]: many irrelevant S-Os may cause people to disable S-O.	MP **11 Jst.**
103	18	O	**Work environment** [public spaces]: Consider using **interaction mode** [headphones] for S-O to protect privacy.	MPs **4,11 Jst.**
104	18	O	**Generic task** [general instruction]: use S-O sparingly. *Comment:* It is not clear which point is being made here.	**No jst.**
110	18	O	**Generic task** [learning]: if using S-O, add facilities for reviewing for **learning parameter** [enhancement of long-term retention].	MPs **8,9 Jst.**
115	30	O	S-O is good in some **generic systems,** bad **cognitive property** (irritating, annoying) in others.	MPs **4,11,13 Jst.**
17	24	**S-I**	**Performance parameter** [effectiveness] depends on the **task**. *Comment:* This follows from claims 63,117 in Table 4, 64,119 in Table 5, 47, 61,72, 75,76,87,93,94,96,121 in Table 10 and 85,86,88,92,97 in Table 11.	MPs **1,2,5,7,8,17 Jst.**
48	15	I	S-I is not adequate for all **information, tasks, users, environments, etc.** *Comment:* This follows from results on S-I in Tables 9 and 10.	MPs **1,2,6,8, 10,11 Jst.**
74	27	I	**Generic task** [text entry and editing]: S-I, being **performance parameter** [slow and inefficient], is only acceptable if haptic modalities are unavailable. *Comment:* Note that haptic modalities include devices that enable spatial manipulation.	MPs **1,5 Jst.**
77	27	I	**Work environment** [office]: may inhibit use of S-I applications [word processors].	MPs **4,11 Jst.**
78	34	I	**Generic task** [text annotation]: different annotation types may require different input modalities (typed language, spoken language, hand-written language). *Comment:* Follows from claims 89,90,91 in Table 5.	MPs **15,16 Jst.**
111	18	I	**Work environment** [office]: S-I, being public, can have negative effects.	MPs **4,11 Jst.**
123	30	I	If S-I is used then use S-O. *Comment:* This may well be false.	**No jst.**

Table 6. The 16 cases of conditional claims on the use of combined speech input/output, speech output and speech input, respectively. Full justification was provided in 14 cases and no justification in 2 cases.

No	Rf	Modalities	COMPARATIVE CLAIMS AGAINST THE USE OF SPEECH INPUT/OUTPUT	Justification
2	19	**S-I/O vs. visual-O + haptic-I**	**Generic task** [spatial manipulation]: S-I/O is inferior to visual output and haptic input. *Assumption:* The haptic input in question enables spatial manipulation.	MPs **1, 2,7 Jst.**

Table 7. The only case of a claim which negatively compared speech input/output to other modalities. Full justification was provided.

No	Rf	RECOMMENDATIONS AGAINST THE USE OF SPEECH OUTPUT	Justification
6	1	**Generic system** [ATMs] which requires privacy protection: Do not prefer S-O.	MP **4 Jst.**
31	32	Output **speech act** [alarm]: avoid S-O when there is a **cognitive property** [memory] component. *Comment:* Note MP9, however.	MP **8 Jst.**
32	32	Output **speech act** [alarm]: avoid S-O when there may be a **performance parameter** [delay] before the fault is attended to. *Comment:* Note MP9, however.	MP **8 Jst.**
33	32	Output **speech act** [alarm]: avoid S-O when there may be several simultaneous alarms.	MP **10 Jst.**
34	32	Output **speech act** [alarm]: avoid when spatial reference to the information source is important.	MP **4 Jst.**
106	18	S-O is serial and transient and therefore **cognitive property** [burdens memory].	MP **8 Jst.**

Table 8. The 6 cases of claims which recommends against the use of speech output. Full justification was provided in all 6 cases.

No	Rf	Modalities	COMPARATIVE CLAIMS AGAINST THE USE OF SPEECH OUTPUT	Justification
27	3	**S-O vs, static gr. text**	**Generic task** [multi-alarm management in process control]: prefer (interactively) static text output over S-O.	MPs **7, 10, Jst.**
30	32	static gr. text	S-O may lock people out of the interaction for its duration. Static visual displays can **performance parameter** [be sampled] when convenient.	MPs **7,8, Jst.**
108	18	static gr. text	**Generic task** [learning]: S-O, being serial and transient, affords less **learning parameter** [elaboration options] for learning than graphic text.	MPs **8, 10, Jst.**
109	18	static gr. text	**Generic task** [general instruction]: S-O: prefer text for longer messages if review is not possible.	MPs **7,8, 9 Jst.**
29	32	**S-O vs. static text/S-O and static text**	**Generic task** [alarm management in process control]: empirical results suggest **performance parameter** [better performance] in text and speech, and text only conditions than in the speech-only condition.	MPs **7,8, 10 Jst.**

Table 9. The 5 cases of claims which negatively compared speech output to other modalities. Full justification was provided in all 5 cases.

No	Rf	RECOMMENDATIONS AGAINST THE USE OF SPEECH INPUT	Justification
47	15	**Generic tasks** [analogue quantify, position]: avoid S-I. *Comment:* These operations include non-linguistic spatial manipulation.	MPs **1,2 Jst.**
61	33	**Generic task** [continuous + time critical aviation control and data input (e.g. volume control in communications equipment, selection of positions of flaps, speed breaks and trims)]: avoid S-I. *Comment:* These operations include precise, non-linguistic spatial positioning.	MPs **1,2 Jst.**
69	33	**Generic task** [aviation]: rigid, *isolated words* S-I dialogue is **cognitive property** [unnatural] and adds **cognitive property** [workload].	MP **14 Jst.**
72	27	**Generic task** [text editing]: empirical results: S-I is **performance parameter** [slow, cursor movements are difficult]. *Comment:* These operations involve non-linguistic spatial manipulation.	MPs **1,2 Jst.**
75	27	**Generic task** [text editing]: S-I is **performance parameter** [very difficult for cursor control and describing locations in the text]. *Comment:* These operations involve spatial manipulation.	MPs **1,2 Jst.**
76	27	**Generic task** [text editing]: Users have difficulty **cognitive property** [inventing S-I commands] for editing operators. *Comment:* These operations involve non-linguistic spatial manipulation.	MPs **1,2 Jst.**
87	34	**Generic task** [text annotation]: empirical results: S-I users experience **cognitive property** [articulation impairment]. *Comment:* There may be more reasons involved than MPs 1 and 2.	MPs **1,2 Jst.**
93	36	**Generic task** [text annotation]: empirical results: S-I is **performance parameter** [unsuited] for higher-level text structure editing (i.e. sentence and paragraph level). *Comment:* These operations involve non-linguistic spatial manipulation.	MPs **1,2 Jst.**
94	34	**Generic task** [text annotation]: empirical results: S-I is **performance parameter** [inefficient] for 'cut and paste' editing. *Comment:* These operations involve non-linguistic spatial manipulation.	MPs **1,2 Jst.**
96	14	**Generic task** [text annotation]: S-I is **performance parameter** [inefficient] for describing spatial information.	MPs **1,2 Jst.**
118	30	*Isolated word* S-I: **performance parameter** [difficult] and **cognitive property** [irritating] for humans to pause between words, especially in stressful situations.	MP **14 Jst.**
120	30	**Generic task** [large-text entry] + **user group** [office workers]: cannot be expected to **cognitive property** [accept] *connected* S-I.	MP **14 Jst.**
121	30	**Generic multimedia systems** [office] + **generic task** [S-I text editing + keyboard text entry]: S-I has no advantage in **performance parameters** [speed, accuracy, and ease of use]. *Comment:* Justification from claims 72,75,76,87,93,94,96 above.	MPs **1,2 Jst.**

Table 10. The 13 cases of claims which recommend against the use of speech input. Full justification was provided in all 13 cases. Note that 69 and 118 mention isolated words recognition and that 120 mentions connected speech recognition. This is rare in the data.

No	Rf	Modali-ties	COMPARATIVE CLAIMS AGAINST THE USE OF SPEECH INPUT	Justific-ation
15	10, 17	**Stan-dard input devices**	S-I is non-advantageous compared to standard input devices. *Correction:* A sweeping, holistic claim which is in potential conflict with all of the S-I and S-I/O recommendations listed in this paper.	**No jst. + Cor. by** MPs **4,5, 16,17**
84	34	**Written input**	**Generic task** [text annotation]: empirical results: S-I leads to **perfor-mance parameter** [fewer annotations] than written input. *Comment:* MP1 and MP2 make this a negative claim about S-I. *Assumption:* The "written input" contains non-linguistic, analogue spatial representations.	MPs **1,2 Sup.**
85	34	Written input	**Generic task** [text annotation]: empirical result: S-I is inferior to written input wrt. **performance parameter** [ease of review].	MPs **7,8 Jst.**
86	34	Written input	**Generic task** [text annotation]: empirical result: S-I is inferior to written input wrt. **performance parameter** [attachment to right place in text]. *Assumption:* The "written input" contains non-linguistic, analogue spatial representations.	MPs **1,2 Jst.**
88	34	Written input	**Generic task** [text editing]: empirical results: S-I is inferior to written input for higher-level text structure editing. *Assumption:* The "written input" contains non-linguistic, analogue spatial representations.	MPs **1,2 Jst.**
92	34	Written input	**Generic task** [proof reading]: empirical results: S-I is generally non-advantageous compared to written input. *Assumption:* The "written input" contains non-linguistic, analogue spatial representations.	MPs **1,2 Sup.**
97	34	Written input	**Generic task** [text annotation]: S-I is dubious compared to written input. *Assumption:* The "written input" contains non-linguistic, analogue spatial representations.	MPs **1,2 Jst.**
73	27	**Key-board**	**Generic task** [text entry and editing]: empirical results: S-I is **perfor-mance parameter** [slower and less efficient] than haptic modality [key-board]. *Comment:* The suitability of S-I for text entry remains moot. The key-board is a complex device which may also enable non-linguistic spatial manipulation.	MPs **1,2 Sup.**
70	33	**Push-buttons**	**Speech act** [time-critical command] input: S-I is **performance parameter** [slower] than haptic [push-button] modality.	**No jst.**

Table 11. The 9 cases of claims which negatively compared speech input to other modalities. Full justification was provided in 4 cases, support in 3 cases, and no justification in 2 cases. Correction was proposed in one of these cases.

APPENDIX 2. SCOPE OF THE DOMAIN VARIABLES OCCURRING IN THE CLAIMS IN TABLES 1-11

The actual scope of the domain variables, such as 'generic task' or 'user group', which have been used to express the claims in Tables 1 through 11 in Appendix 1, is presented below.

Generic task
dual, hands and/or eyes occupied (2 occurrences)
mobility is needed
follow procedure
serial fault management
guidance, warnings, instructions, read-aloud text, feedback
single fault alarm and management in process control
general instruction (2 occurrences)
alarm management in process control
1-out-of N selection + N vocabulary size of the order of hundreds
process plant control
aviation control and data input (9 occurrences)
aviation control, data input and queries
dual task aviation control
non-time-critical but critical command input
text annotation (17 occurrences)
quality control
text entry
aviation data input (2 occurrences)
heads up aviation control (2 occurrences)
text entry and editing (2 occurrences)
data entry
general instruction
learning
spatial manipulation
multi-alarm management in process control
learning
analogue quantify, position
continuous + time critical aviation control and data input
aviation
text editing (4 occurrences)
large-text entry
S-I text editing + keyboard text entry
proof reading

Speech act
warning (2 occurrences)
alarm (8 occurrences)
command (2 occurrences)
time-critical command
instruction
advice

User group
people who have difficulty using computers
physically disabled users
the disabled
pre-school children, the blind who have not learnt Braille
non-experts
office workers

Interaction mode
computer access over the telephone
radio link
wireless device
microphones and/or headphones
headphones

Work environment
public spaces
office (2 occurrences)

Generic system
complex
electronic multimedia document
multimedia
inquiry-based

ATMs
office

Performance parameter
user mobility
immediate response
mobile operator
more effective
better performance
fast
location free
mobile control
hands and eyes free operation
ease of operation
increased control capabilities
busy
safer
easier
workload conflicts
efficiency
avoids shorthand
elaboration
faster (3 occurrences)
more accurate
better performance
more accurate
slower and less efficient
slower
repetition
effectiveness
slow and inefficient,
delay
be sampled
slow, cursor movements are difficult
very difficult for cursor control and describing locations in the text
inefficient (2 occurrences)
unsuited
difficult
speed, accuracy, and ease of use
fewer annotations
ease of review
attachment to right place in text

Learning parameter
interaction training time
no learning overhead (2 occurrences)
enhancement of long-term retention
elaboration options

Cognitive property
attention-catching (4 occurrences)
be attended to
spatial and temporal distraction
setting a mood
persuasive
human discrimination capacities
reduction in visual workload
reading load
high workload
naturalness (2 occurrences)
unnatural
cognitive processing limitations
memory
burdens memory
workload
inventing S-I commands
articulation impairment
irritating
irritating, annoying
acceptance