IS SPEECH THE RIGHT THING FOR YOUR APPLICATION?

Niels Ole Bernsen and Laila Dybkjær

Natural Interactive Systems Group, Odense University,
Forskerparken 10, 5230 Odense M, Denmark
Phone: +45 65 57 35 44 Fax: +45 63 15 72 24
Email: nob@mip.ou.dk, laila@mip.ou.dk

ABSTRACT

Use of speech input to, and speech output from, computer systems is spreading at a growing pace. This means that, increasingly, developers of systems and interfaces are faced with the question of whether or not to use speech input and/or speech output for the applications they are about to build. This paper presents results from a pilot test of a theory-based approach to speech functionality. The test uses a corpus of claims about speech functionality derived from recent literature on speech and multimodality.

1. INTRODUCTION

Use of speech input to, and speech output from, computer systems is spreading at a growing pace. This means that, increasingly, 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. Until recently, the literature has offered no systematic guidance on this issue of speech functionality although there is consensus that early design guidance is highly desirable [1]. This would reduce the risk of having to do quite basic re-design later on due to, e.g., user dissatisfaction or poor system performance. Systematic guidance could benefit from theory but theory alone is not sufficient. Once developed, theory must be transformed into practically useful tools which can be applied by non-theoreticians. This paper presents results from a pilot test of a theory-based approach to speech functionality [3]. The test involves a corpus of claims about speech functionality derived from recent literature on speech and multimodality. If the full test proves successful, the existing proto-tool [2] can be developed into a workable tool that may assist developers of systems and interfaces in deciding when (not) to use speech in their applications.

2. PREVIOUS WORK

It is trivial to argue that speech is not always suited for human-computer information exchange. An equally trivial generalisation is that, sometimes, other modalities are preferable to speech if we want to optimise the human-computer interface from the point of view of information exchange. But 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? It was shown in [3] that this problem is too complex to be realistically resolved through empirical experimentation. The experimental variables are just too many, including task type, communicative act (e.g. alarm), user group, work environment, system type, performance parameters (e.g. more effective), learning parameters (e.g. learning overhead), and cognitive properties (e.g. attention load). The only constant property of claims about speech functionality is that the claims involve, often oblique, reference to objective *modality properties*, such as that speech is omnidirectional or is eyes-free.

Using as data points 120 claims about speech functionality systematically gathered from papers dedicated to the issue [1], it was shown that a mere 18 modality properties, cf. Figure 1, were sufficient to justify, support or correct 106 of the 109 claims that were not flawed in one way or another.

No	MODALITY	MODALITY PROPERTY
MP1	Linguistic input/output	Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information. They are therefore unsuited for specifying detailed information on spatial manipulation and location.
MP2	Linguistic input/output	Linguistic input/output modalities, being unsuited for specifying detailed information on spatial manipulation, lack an adequate vocabulary for describing the manipulations.
MP3	Arbitrary input/output	Arbitrary input/output modalities impose a learning overhead which increases with the number of arbitrary items to be learned.
MP4	Acoustic input/output	Acoustic input/output modalities are omnidirectional.
MP5	Acoustic input/output	Acoustic input/output modalities do not require limb (including haptic) or visual activity.
MP6	Acoustic output	Acoustic output modalities can be used to achieve saliency in low-acoustic environments.
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 (but only small-piece-by-small-piece).
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 **input**/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 and discourse must be distinguished from designer-designed keywords and discourse 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.
MP 19 **NEW**	Analogue graphics input/output	Analogue graphics input/output modalities lack interpretational scope, which makes them eminently suited for conveying high-specificity information. They are therefore unsuited for conveying abstract information.

Figure 1: The 19 modality properties used in the present study. Differences from the original 18 properties have been marked. Underscore and boldface highlight additions.

All claims could be categorised as belonging to one or other of the 13 claims types presented in Figure 2. Eleven of the 13 types were represented in the data.

The 18 modality properties were taken from modality theory, cf. [3, 5], and include all the properties that modality theory could contribute to the data analysis. Modality theory has been developed for unimodal output modalities. Work on input modalities is in progress.

By *justification* of a data point is meant that, given a set of modality properties and a claim on speech functionality, a designer is *practically justified* in making that claim based on that set of 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, claims might be in partial or full conflict with modality theory. In such cases, *correction* was introduced to the claim in question based on reference to a specific modality property. However, it should be noted that even if a positive claim on speech is justified this does not necessarily mean that the designer is then going to use speech. Any recommendation on speech may in principle be overridden by "external" design considerations, such as 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.

T1: Claims recommending combined speech input/output.

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

T3: Claims recommending speech output.

T4: Claims positively comparing speech output to other modalities.

T5: Claims recommending speech input.

T6: Claims positively comparing speech input to other modalities.

T7: Conditional claims on the use of speech.

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

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

T10: Recommendations against the use of speech output.

T11: Claims negatively comparing speech output to other modalities.

T12: Recommendations against the use of speech input.

T13: Claims negatively comparing speech input to other modalities.

Figure 2: The 13 claims types used for categorising data points. T2 and T8 were not represented in the first data.

An interesting point is that most of the 18 modality properties are *not* about speech. The hierarchical nature of modality theory means that the properties of a particular unimodal modality at some level of abstraction are inherited by that modality's daughter nodes and by their daughter nodes etc., cf. [4]. Justification of why a certain speech modality may, e.g., be recommended for a certain interface 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 a taxonomy of modalities. In other words, the problem of speech functionality *cannot* be solved through appeal to properties that are characteristic of all and only the speech modalities.

The fact that only 18 modality properties were needed to account for nearly all the data was considered an encouraging result. The hypothesis based on this first result is that *knowledge of a small set of modality properties might suffice to settle most issues of speech functionality without trial-and-error or recourse to costly empirical investigation.* The test is whether investigation of an equally large control set of claims about speech functionality will show that the original modality properties are largely sufficient for justifying, supporting or correcting those claims.

3. DATA COLLECTION

In order to test the explanatory power of the 18 modality properties (Figure 1) on a new data set, the following protocol was defined:

(i) Data point collection should be done by the author who was not involved in collecting the previous data.

(ii) All references should be post-1993. The 120 data points mentioned in Section 2 were all from a 1993 collection of papers on interactive speech technology [1]. As multimodal interaction has grown in importance since 1993, we wanted to see whether that would be reflected in the new data when selected from papers published in various proceedings and journals in the years following 1993.

(iii) Decisions on discarding data points due to irrelevance or redundancy, must be agreed by both authors.

(iv) Claims categorisations must be agreed by both authors.

(v) Justification of data points should be made first by the author who did not collect the data. Each justification must be agreed by the other author. In case of disagreement solution should be sought through discussion.

A new set of about 200 data points on speech functionality were collected from 25 papers according to (i) and (ii) above. The pilot analysis of the collected data is reported below.

4. THE PILOT TEST

The pilot test concerns a sub-set of the collected data. To enable comparison with the results reported in Section 2, we only included claims of the 13 types shown in Figure 2. Both authors made a first categorisation of the data according to claims type. The author who did the data collection selected, if possible, two claims for each claims type, such that claims were sought from each of the 25 analysed papers. Only one data point of each of claims types T8 and T9 were found. Three papers did not deliver claims of the requisite types. Two other papers each of which only had one relevant claim, were left out because more detailed claims analysis demonstrated that the claim had not been categorised correctly. Among the remaining 20 papers, four were represented twice and the rest once. 24 data points were thus selected for pilot analysis by the author who did not do the data collection.

5. DATA ANALYSES

It is important to bear in mind that this paper deals with very complex data (cf. Figures 3 and 4) which, moreover, have been extracted from their context. The purpose of *data representation* is to express the claims in a comparable and intelligible format which preserves the basic point(s) made by their authors. The purpose is not (a) to co-represent the full context of each data point; nor (b) to make each data point fully explicit with respect to its implicit assumptions; nor (c) to 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".

The analysis showed that of the selected 24 data points, 21 could be either fully justified (Figure 3), or supported to the extent deserved by a partially false claim (Figure 4).

Of the three claims which found no justification or support as deserved, one was beyond the scope of modality theory as it concerned the relative speed of producing information in different modalities. The second claim was extremely vague, as in "modality M1 may be uncomfortable to some users". Such claims are very often true but extremely hard to justify on principled grounds. The third claim could not be justified but only supported due to the fact that input modality theory is incomplete.

Figure 5 shows the modality properties used. Modality properties were used 35 times in justifying, supporting or correcting the 24 claims. There were 27 cases of justification, 7 cases of support, and one correction.

Four modality properties, MP1, MP9, MP15 and MP17, had to be slightly augmented in order to provide full justification or support. All augmentations come straight from Modality Theory, expanding the property derived from Modality Theory to suit the data encountered. One new modality property, MP19, was added which represents a basic insight of Modality Theory and mirrors MP1. The augmentations and the new modality property are shown in Figure 1.

6. CONCLUSION

The results of the pilot analysis of the 24 data points show a rather strong confirmation of the explanatory power of the existing set of modality properties. Only one new output modality property has been found necessary and only one input modality property has been found missing. This is encouraging for the trial with the full data set.

A lesson learnt is that the present set of modality properties should be expressed in full whenever possible, rather than waiting for data which requires this to happen.

If the larger-scale analysis confirms the pilot study, we will proceed to developing a full hypertext/hypermedia design support tool for the web (cf. the proto-tool demonstrator in [3]). It would then seem likely that developers might benefit from a design support tool which provides easy-to-use information on the relevant modality properties and their practical import.

27b. Speech recognition technology is necessary to automate services where the number of service options is large. For example, a restaurant selector service that asks callers which cuisine they would like would be manageable as a speech automated service ("What kind of cuisine would you like?") but unwieldy as a Touch-Tone service ("For Chinese food, press 11; for Italian food, press 12 ...") [13]

Data point 27b. **Generic task** [large number of service options, e.g. restaurant cuisine options]: speech input/output is **performance parameter** [manageable] whereas menu style touch-tone interaction, i.e. haptic [telephone keys] input/speech output, is not. Justified by MP8: "Dynamic output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection." Type: **T2.**

NOTE: The justification implies that the output task might be done by static graphics (text modalities possibly supplemented with images for illustration), cf. MP7.

Figure 3: A justified claim as original and as represented.

167. A few general observations may be made about the adequacy and applicability of each mode: Spoken mode: ... Output as: help, examples, requests, explanation, suggestion. [21, 346]

Data point 167. **Generic tasks** [help, examples, requests, explanation, suggestion]: speech output is **performance parameters** [adequate and applicable]. Supported by MP1: "Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information." Type: **T3.**

NOTE: the reason claim 167 is only supported (not justified) by the quoted part of MP1 is that claim 167 is partly false because it is overly general. In particular, many exemplifications require concrete illustrations of the kind provided by, e.g., graphic or acoustic images, and many explanations require, e.g., graphic diagrams.

Figure 4: A supported claim.

MP	MODALITY	NO. OF CLAIMS ADDRESSED
1	Linguistic input/output	7
2	Linguistic input/output	1
3	Arbitrary input/output	0
4	Acoustic input/output	1
5	Acoustic input/output	3
6	Acoustic output	0
7	Static graphics output	2
8	Dynamic output	5
9	Dynamic acoustic output	1
10	Speech input/output	0
11	Speech input/output	0
12	Speech output	0
13	Synthetic speech output	0
14	Non-spontaneous speech input	1
15	Discourse output	2
16	Discourse input/output	2
17	Spontaneous spoken labels/key-words and discourse input/output	7
18	Notational input/output	2
19	Analogue graphics input/output	1

Figure 5: Modality properties were used 35 times in justifying, supporting and correcting the 24 data points.

7. REFERENCES

1. Baber, C. and Noyes, J. (Eds.): Interactive speech technology. Taylor and Francis, London, 1993.

2. Bernsen, N.O.: Towards a tool for predicting speech functionality. Free Speech Journal, 1996. http://www.cse.ogi.edu/CSLU/fsj/issues/issue1/

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

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

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