A THEORY OF SPEECH IN MULTIMODAL SYSTEMS

Niels Ole Bernsen and Laila Dybkjær

Natural Interactive Systems Laboratory, University of Southern Denmark,
Science Park 10, 5230 Odense M, Denmark
Phone: +45 65 50 35 44   Fax: +45 63 15 72 24
Email: nob@nis.sdu.dk, laila@nis.sdu.dk

ABSTRACT

Increasingly, speech input and/or speech output is being used in combination with other modalities for the representation and exchange of information with, or mediated by, computer systems. Therefore, a growing number of developers of systems and interfaces are faced with the question of whether or not to use speech input and/or speech output in multimodal combinations for the applications they are about to build. This paper presents first results on speech in multimodal systems from a test of a theory-based approach to speech functionality. The test used a large corpus of claims about speech functionality derived from the recent literature.

1. SPEECH FUNCTIONALITY

The *speech functionality problem* is the question of what speech is good or bad for, or under which conditions to use, or not to use, speech for information representation and exchange - either speech alone or in combination with other modalities. With the rapid spread of speech technologies, the speech functionality problem has become one of real practical importance. The research literature is becoming replete with studies of speech functionality including speech in multimodal systems, such as speech and multimedia [1], speech and graphics [2,3], speech and gesture [4], speech in auditory interfaces [5,6], speech, pen and graphics [7,8,9,10], email vs. voice mail [11]. It seems unlikely, however, that empirical studies will suffice in telling system developers what they need to know in a timely fashion in order to avoid user dissatisfaction or poor system performance due to erroneous choices of modality combinations. This is due to the complexity of the speech functionality problem (Figure 1).

The combinatorics described in Figure 1 is daunting. If possible at all, it would take decades of empirical experimentation to investigate all the possibilities. There are several speech modalities, such as keywords and unrestricted discourse; there is speech as input and speech as output; there are scores of non-speech modalities with which speech might conceivably be combined; and the success of a particular modality choice is subject to an unlimited number of instantiated *domain variables*, including task type (e.g. navigating hypermedia), communicative act (e.g. alarm), user group (the blind), work environment (natural field settings), system type (e.g. personal intelligent assistant), performance parameters (e.g. more efficient), learning parameters (e.g. learning overhead), and cognitive properties (e.g. attention load).

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

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

In other words, it would be useful for developers to be able to rely largely on comprehensible theoretical guidance instead of lengthy experimentation. This paper reports on the results of a recent study of how it might be possible to support developers' reasoning about speech functionality, emphasizing the use of speech in a multimodal context.

2. AN ENCOURAGING RESULT

Given the huge complexity described in Section 1, it is a striking fact that the only constant property of claims about speech functionality, such as "Speech input is useful when the user's hands are occupied", is that the claims involve, often oblique, reference to objective *modality properties,* such as that speech is omnidirectional or is eyes-free. The purpose of *Modality*

Theory [12,13] is to describe the objective properties of all unimodal modalities in acoustics, graphics and haptics. The observation that all speech functionality claims refer to modality properties gave rise to the idea of testing the explanatory power of Modality Theory on a small but well-defined fragment within the scope of the theory, i.e. a set of claims about speech functionality.

Using as data points 120 claims about speech functionality that were systematically gathered from papers dedicated to the issue [14], it was shown that a mere 18 modality properties (Figure 2), were sufficient to justify, support or correct 106 (97%) of the 109 claims that were not flawed in one way or another [15]. The 18 modality properties were taken from Modality Theory and include all the properties that the theory could contribute to the claims analysis. All claims could be categorised as belonging to one of 13 types (Figure 3). Eleven of the 13 types were represented in the data.

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 2. The 18 modality properties used in [15].

By *justification* of a data point is meant that, given a set of modality properties and a claim about 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, modality properties 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 modality properties. It should be noted that, even if a positive claim about speech functionality is justified, this does not necessarily mean that the designer should be using 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.

An interesting point is that most of the 18 modality properties in Figure 2 are *not* (only) about speech. Justification 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.

Claims recommending combined speech input/output (T1), speech output (T3), speech input (T5).

Claims positively comparing combined speech input/output (T2), speech output (T4), speech input (T6) to other modalities.

Conditional claims on the use of speech (T7).

Recommendations against the use of combined speech input/output (T8), speech output (T10), speech input (T12).

Claims negatively comparing combined speech input/output (T9), speech output (T11), speech input (T13) to other modalities.

Figure 3. The 13 claims types used for categorising data points in [15]. T2 and T8 were not represented in the data.

The fact that only 18 modality properties were needed to justify, support or correct nearly all the data was considered an encouraging result. The derived hypothesis is that *knowledge of a small set of modality properties might suffice to evaluate most issues of speech functionality without trial-and-error or recourse to costly empirical investigation.*

3. SPEECH AND MULTIMODALITY

To test the hypothesis mentioned at the end of Section 2, we did a second study of speech functionality claims according to the strict protocol described in [16]. Basically, data selection and analysis verification was done by the second author, whereas data representation and analysis was done by the first author. The objectives were to (a) investigate the extent to which Modality Theory would be capable of providing justification, support or correction to a large selection of claims, possibly by invoking modality properties in addition to those listed in Figure 2; and (b) obtain an indication of the proportion of new modality properties needed. The new study includes a new type of claim which was excluded from [15], cf. Figure 3, i.e. *claims recommending speech in combination with other modalities* (Rsc.). These claims are of particular interest in the present paper.

A set of 153 claims on speech functionality were collected in 23 papers from the literature from 1993 to 1998. The claims, or data points, were represented semi-formally and evaluated from the point of view of Modality Theory. It is important to bear in mind that we are dealing with very complex data which, moreover, have been extracted from their context. The purpose of *data representation* is to express all 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 formalised representation. (c) would probably be impossible; and (a) and (b) would mean producing lengthy renderings of the data, which would defeat the practical aims of the analysis. The data, as rendered, therefore remain partially "messy". Figure 4 shows data point 48 from [7] and its semi-formal representation.

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

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

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

Figure 4. Data point 48.

Figure 4 illustrates how each claim is represented in terms of the modalities involved and the domain variables it instantiates (cf. Figure 1), followed by an evaluation in terms of modality properties (if any), the claims type, such as “Rsc.” (recommendation of speech in combination with other modalities), an (optional) explanatory note, and an evaluation of the claim independently of Modality Theory. The latter is important: it should be held against the modality property approach if a true, or at least reasonable, claim cannot be justified or supported by modality properties, but if a claim is false then no modality property should justify it.

Overall, the new study showed that 143 in 153 claims deserved justification or support or, in the case of false claims, correction. 25 modality properties provided this in 134 (or 94%) cases. The 25 modality properties included the 18 properties listed in Figure 2. Thus, roughly, whereas 18 properties sufficed to justify, support or correct the original set of 120 claims, 25 modality properties can justify, support or correct 120 + 153 = 273 claims. With only 7 new modality properties being needed to handle (most of) the 153 new claims, this suggests that speech functionality could be addressed from a limited set of modality properties.

Of the 153 claims, 40 or 26% were recommendations of speech in combination with other modalities, illustrating the extent to which researchers have turned towards the use of speech in multimodal contexts. The large majority of claims (36) advocate the usefulness of novel interaction paradigms that include speech. One paradigm combines the graphical user interface (GUI) paradigm with speech input and/or speech output. Thus, fourteen claims argue that speech input can often be added to the GUI paradigm for enhancement, efficiency and complementarity rather than replacement. For instance, the user points to some object in graphical output space and, using speech, specifies what should be done to it. A second paradigm represents an alternative to the GUI paradigm on the input side. Thus, nineteen claims argue that speech input can often provide added efficiency and flexibility to an (input) interface in which mouse and keyboard have been replaced by the pen. For instance, the speech-pen combination can be used by mobile users which cannot easily operate with the standard GUI set-up. Speech would not replace the pen but the two together have powerful complementary properties when used for input into graphical output space. Finally, a third paradigm (3 claims) combines speech and other acoustic modalities into an all-acoustic output interface for the blind. In this case, for instance, speech can be used to label acoustic images. In addition to these three alternative paradigms, speech output, in particular, is recommended for special roles within the GUI paradigm, such as for introducing large amounts of text, highlighting key information and, in a more traditional vein, acoustic alarms.

4. CONCLUSION

Without claiming any statistical significance, the observations reported at the end of Section 3 illustrate that we are only beginning to address the powers and limitations of speech in a multimodal context.

The results gained from taking a theory-based approach to speech functionality have encouraged us to develop a hypertext/hypermedia web-based speech functionality design support tool as envisioned in [15]. The tool which is called SMALTO will be announced on the DISC web pages, so keep an eye on http//www.elsnet. org/disc/

REFERENCES

1. Furui, S.: Projects for spoken dialogue systems in a multimedia environment. *Proceedings of the ESCA workshop on Spoken Dialogue Systems*, Vigsø, Denmark, 1995, 9-16.
2. Wyard, P., Appleby, S., Kaneen, E., Williams, S. and Preston, K.: A combined speech and visual interface to the BT business catalogue. *Proceedings of the ESCA workshop on Spoken Dialogue Systems*, Vigsø, Denmark, 1995, 165-168.
3. Nitta, T.: Speech recognition applications in Japan. *Proceedings of ICSLP ‘94*, Yokohama, Japan, 1994, 671-674.
4. Loken-Kim, K., Park, Y., Mizunashi, S., Fais, L. and Morimoto, T.: Verbal gestural behaviours in multimodal spoken language interpreting telecommunications. *Proceedings of Eurospeech ‘95*, Madrid, 1995, 281-284.
5. Mynatt, E. D.: Transforming graphical interfaces into auditory interfaces for blind users. *Human-Computer Interaction*, Vol. 12, No. 1&2, 1997, 7-45.
6. Stevens, R. D., Edwards, A. D. N. and Harling, P.A.: Access to mathematics for visually disabled students through multimodal interaction. *Human-Computer Interaction*, Vol. 12, No. 1&2, 1997, 47-92.

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

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

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

10. Cohen, P. R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L. and Clow, J.: QuickSet: Multimodal interaction for distributed applications. ACM Multimedia 97, Seattle Washington, USA, 1997, 31-40.

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

12. 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.

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

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

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

16. Bernsen, N. O. and Dybkjær, L.: Is speech the right thing for your application? *Proceedings of ICSLP '98*, Sydney, Australia, 1998, 3209-3212.