GUIDELINES FOR MULTIMODAL User Interface Design

In today's pursuit of more transparent, flexible, and efficient human-computer interaction, a growing interest in multimodal interface design has emerged [5]. The goals are twofold: to achieve an interaction closer to natural human-human communication, and to increase the robustness of the interaction by using redundant or complementary information. New interaction paradigms and guidelines are necessary to facilitate the design of multimodal systems from the ground up (see articles by Cohen and McGee and Pieraccini et al. in this section). This article discusses six main categories of guidelines and represents a preliminary effort to establish principles for multimodal interaction design. A more detailed discussion of these guidelines will be available in [6].

Requirements Specification. Critical to the design of any application are the user requirements and system capabilities for the given domain. Here, we provide some general considerations for multimodal system requirements specification.

Design for broadest range of users and contexts of use. Designers should become familiar with users' psychological characteristics (for example, cognitive abilities, motivation), level of experience, domain and task characteristics, cultural background, as well as their physical attributes (for example, age, vision, hearing). An application will be valued and accepted if it can be used by a wide population and in more than one manner. Thus, multimodal designs can aid in extending the range of potential users and uses, such as when redundancy of speech and keypad input enables an application to be used in dark and/or noisy environments. Designers should support the best modality or combination of modalities anticipated in changing environments (for example, private office vs. driving a car).

Address privacy and security issues. Users should be recognized by an interface only according to their explicit preference and not be remembered by default. In situations where users wish to maintain privacy by avoiding speech input or output, multimodal interfaces that use speech should also provide a non-speech mode to prevent others from overhearing private conversations. Non-speech alternatives should also be provided when users enter personal identification numbers, passwords (for example, automatic bank teller), or when they might be uncomfortable if certain private information is overheard by others. For example, to reduce the likelihood of others being aware of a user's mistakes, it may be preferable to provide error messages in a visual form instead of audible speech.

Designing Multimodal Input and Output. The cognitive science literature on intersensory perception and intermodal coordination has provided a foundation for determining multimodal design principles [2, 5, 7]. To optimize human performance in multimodal systems, such principles can be used to direct the design of information presented to users, specifically regarding how to integrate multiple modalities or how to support multiple user inputs (for example, voice and gesture). Here, we provide a brief summary of some general guiding principles essential to the design of effective multimodal interaction.

Maximize human cognitive and physical abilities. Designers need to determine how to support intuitive, streamlined interactions based on users' human information processing abilities (including attention, working memory, and decision making) for example:

- Avoid unnecessarily presenting information in two different modalities in cases where the user must simultaneously attend to both sources to comprehend the material being presented [1, 3]; such redundancy can increase cognitive load at the cost of learning the material [1].
- Maximize the advantages of each modality to reduce user's memory load in certain tasks and situations, as illustrated by these modality combinations [7, 8]:
  - System visual presentation coupled with user manual input for spatial information and parallel processing;
  - System auditory presentation coupled with user speech input for state information, serial processing, attention alerting, or issuing commands.

Integrate modalities in a manner compatible with user preferences, context, and system functionality. Additional modalities should be added to the system only if they improve satisfaction, efficiency, or other aspects of performance for a given user and context. When using multiple modalities:

- Match output to acceptable user input style (for example, if the user is constrained by a set grammar, do not design a virtual agent to use unconstrained natural language);
- Use multimodal cues to improve collaborative speech (for example, a virtual agent's gaze direction or gesture can guide user turn-taking);
- Ensure system output modalities are well synchronized temporally (for example, map-based display and spoken directions, or virtual display and non-speech audio);
- Ensure the current system interaction state is shared across modalities and that appropriate information is displayed in order to support:
  - Users in choosing alternative interaction modalities;
  - Multidevice and distributed interaction;
  - System capture of a user's interaction history.

Adaptivity. Multimodal interfaces should adapt to the needs and abilities of different users, as well as different contexts of use. Dynamic adaptivity enables the interface to degrade gracefully by leveraging complementary and supplementary modalities according to changes in task and context. Individual differences (for example, age, preferences, skill, sensory or motor impairment) can be captured in a user profile and used to determine interface settings such as:

- Allowing gestures to augment or replace speech input in noisy environments, or for users with speech impairments;
- Overcoming bandwidth constraints (for example, local direct manipulation replaces gaze input that must be analyzed remotely);
- Adapting the quantity and method of information presentation to both the user and display device.

Consistency. Presentation and prompts should share common features as much as possible and should refer to a common task including using the same terminology across modalities. Additional guidelines include providing consistent:

- System output independent of varying input modalities (for example, the same keyword provides identical results whether user searches by typing or speaking);
- Interactions of combined modalities across applications (for example, consistently enable shortcuts);
- System-initiated or user-initiated state switching (for example, mode changing), by ensuring the user's interaction choices are seamlessly detected and that the system appropriately provides feedback when it initiates a modality change.

Feedback. Users should be aware of their current

MULTIMODAL *interfaces should adapt to the needs and abilities of different users, as well as different contexts of use. Dynamic adaptivity enables the interface to degrade gracefully by leveraging complementary and supplementary modalities according to changes in task and context.*

connectivity and know which modalities are available to them. They should be made aware of alternative interaction options without being overloaded by lengthy instructions that distract from the task. Specific examples include using descriptive icons (for example, microphone and speech bubbles to denote click-to-talk buttons), and notifying users to begin speaking if speech recognition starts automatically. Also, confirm system interpretations of whole user input after fusion has taken place [4], rather than for each modality in isolation.

Error Prevention/Handling. User errors can be minimized and error handling improved by providing clearly marked exits from a task, modality, or the entire system, and by easily allowing users to undo a previous action or command. To further prevent users from guessing at functionality and making mistakes, designers should provide concise and effective help in the form of task-relevant and easily accessible assistance. Some specific examples include [5]:

- Integrate complementary modalities in order to improve overall robustness during multimodal fusion, thereby enabling the strengths of each to overcome weaknesses in others;
- Give users control over modality selection, so they can use a less error-prone modality for given lexical content;
- If an error occurs, permit users to switch to a different modality;
- Incorporate modalities capable of conveying rich semantic information, rather than just pointing or selection;
- Fuse information from multiple heterogeneous sources of information (that is, cast a broad "information net");
- Develop multimodal processing techniques that target brief or otherwise ambiguous information, and are designed to retain information.

Conclusion

The guiding principles presented here represent initial strategies to aid in the development of principle-driven multimodal interface guidelines. In order to develop both innovative and optimal future multimodal interfaces, additional empirical studies are needed to determine the most intuitive and effective combinations of input and output modalities for different users, applications and usage contexts, as well as how and when to best integrate those modalities. To fully capitalize on the robustness and flexibility of multimodal interfaces, further work also needs to explore new techniques for error handling and adaptive processing, and then to translate these findings into viable and increasingly specific multimodal interface guidelines for the broader community.

References

1. Cooper, G. *Research in to Cognitive Load Theory and Instructional Design at UNSW* (1997); www.arts.unsw.edu.au/education/CLT_NET_Aug_97.HTML.
2. European Telecommunications Standards Institute. Human factors: Guidelines on the multimodality of icons, symbols, and pictograms. (Report No. ETSI EG 202 048 v 1.1.1 (2002-08). ETSI, Sophia Antipolis, France.
3. Kalyuga, S., Chandler, P., and Sweller, J., (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology 13* (1999), 351–371.
4. McGee, D.R., Cohen, P.R., and Oviatt, S. Confirmation in multimodal systems. In *Proceedings of the International Joint Conference of the Association for Computational Linguistics and the International Committee on Computational Linguistics* (Montreal, Quebec, Canada, 1998).
5. Oviatt, S.L. Multimodal interfaces. *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications.* J. Jacko and A. Sears, Eds. Lawrence Erlbaum, Mahwah, NJ, 2003, 286–304.
6. Reeves, L.M., Lai, J. et al. Multimodal interaction design: From a lessons-learned approach to developing principle-driven guidelines. *International J. of HCI*. Forthcoming.
7. Stanney, K.M., Samman, S., Reeves, L.M., Hale, K., Buff, W., Bowers, C., Goldiez, B., Lyons-Nicholson, D., and Lackey, S. A paradigm shift in interactive computing: Deriving multimodal design principles from behavioral and neurological foundations. In review.
8. Wickens, C.D. *Engineering Psychology and Human Performance* (2nd ed). Harper Collins, NY, 1992.

Leah M. Reeves, Jennifer Lai, James A. Larson (jim@larson-tech.com), Sharon Oviatt (oviatt@cse.ogi.edu), T.S. Balaji, Stéphanie Buisine, Penny Collings, Phil Cohen, Ben Kraal, Jean-Claude Martin, Michael McTear, TV Raman, Kay M. Stanney, Hui Su, and QianYing Wang.

This article was originally drafted among the co-authors as part of a CHI'03 workshop on multimodal interaction and interface design principles organized by James Larson and Sharon Oviatt.