PERCEPTUAL INTERFACES IN HUMAN COMPUTER INTERACTION

Dominic W. Massaro
Perceptual Science Laboratory, Department of Psychology,
University of California, Santa Cruz, CA 95064 USA

ABSTRACT

Humans are viewed as efficiently processing multiple sources of information supporting their interactions with the world and with others. This theoretical framework has been successful in predicting for a wide variety of research findings, as well as rationalizing everyday experiences. Existing controversies over perceptual interfaces are considered in terms of this framework. The conclusion is that both inputs and outputs in human computer environments should be characterized by many different sources of information from multiple modalities.

1. A LITTLE HISTORY, FANTASY, AND EVIDENCE

The issue of interfaces in human computer interactions must certainly be as old as machines (humans apparently have been around much longer). As a graduate student, I didn't interact with a machine but hand delivered my IBM computer cards (even though the machine was a CDC 3600) to the receiving desk of the computing center for 24-hour batch processing. Eventually, the computer operators permitted we few geeks--before there were geeks--to enter the intersanctum and deliver the cards directly to the card reader and to pick up our own printouts. Even for experienced batch processors, direct interactions with DEC minicomputers were at first somewhat intimatating. You initiated the computer with the setting of a set of switches, read in paper tape and hammered away at a humongous Teletype. There were no visual or auditory displays other than lights flashing on the cover of the computer and the rattling vibration of the teletype.

We used these minicomputers to control psychology experiments on perception, memory, and decision making. With just 8K of memory, it was straightforward to present short speech sounds as test stimuli, record human responses and response times (with millisecond resolution), and to adapt the difficulty of the test based on the participant's performance [6]. Mathematical model testing using Newtonian search algorithms had to be done offline at the University computing center [6]. The data from the minicomputer experiment had to be retyped for the model testing because although cards and paper tape both had holes, they couldn't read each other's holes.

1.1 Futuristic Education

In contrast to these "primitive" interactions, today we envision an entirely different scenario of humans interacting with computers [2].

Julia's ninth grade conceptual physics class includes a learning module on the nature of time. Julia consults her favorite animated agent, Teacher Molly, and explains her assignment. Saint Augustine appears, introduces himself, and asks: What is time?" He proceeds to describe the puzzle of time. After an interesting discussion, Albert Einstein arrives, and provides some excellent descriptions of the relativity of time, with interesting animation showing the relationship between travel and aging. During the discussion with Dr. Einstein, Dr. Stephen Hawking interrupts to point out the three arrows of time. Excited by these mysteries, Julia asks Teacher Molly how one of the arrows, entropy, can be true in an organized world. Teacher Molly suggests that Julia use the "active worlds" program to simulate a hypothetical world. After several simulations, and some discussions with Teacher Molly (who has the annoying habit of making Julia answer her own questions), Julia understands the tradeoff between entropy and biological evolution. For her science project, Julia incorporates her simulation into a discussion between Albert Einstein and Charles Darwin, which is made available to other students.

Given the extreme overestimation of progress in this field in the past, it would be foolish for me to predict when something like this scenario would be possible. Notwithstanding the question of when this technology will appear, we can at least debate how many resources should be directed at making it occur. We believe that the best informed decision will be grounded in an understanding of human perception, cognition, and motivation. What do we know about these psychological processes that will guide the development and distribution of human computer interaction?

1.2 Optimizing Interfaces

There has been a lack of consensus about the optimal interface for humans and computers. One of the primary disagreements has been over whether voice or text should be the primary communication medium. Text, being digital, was the most natural medium for interfacing us with our computers. We began by typing a keypunch that poked holes in cards with 80 columns and one ASCII character per column. It was easy enough to adapt the Typewriter (the early name for the mechanical devices long before the days of touch typing) for this task, and the QWERTY keyboard has been a ubiquitous appendage for most smart machines. Even smart small appliances without keyboards like digital assistants are successful because they are easily tethered to a computer cum keyboard for data entry into the device. Not only has the keyboard held its privileged status, its appearance is almost always a QWERTY layout. Devotees of alternative keyboard layouts, such as the Dvorak keyboard, have been unsuccessful in replacing QWERTY--even though substantial evidence exists for more optimal keyboard layouts.

0-7803-6536-4/00/$10.00 (c) 2000 IEEE

If text entry was the natural input, it is not surprising that the output display would also be text. A colleague remarked that the standard typewriter warranted a high rating because it provided a highly efficient input/output correspondence. It is perhaps the earliest if not the best example of WYSIWYG (what you see is what you get). Not long after reading rolls of paper on our PDP-8 and PDP-8L, DEC provided monitors on PDP-9, 11, and 12. Unix, of course, was centered around text processing and provided the antithesis of WYSIWYG.

Of course, we had oscilloscopes and were not unaware that visual displays could be used to present analog or graphical information. Physics graduate students at UCSC were illustrating nonlinear dynamical events on these displays in the late 70s [17]. About the same time, psychologists were learning that parallel processing was a key feature of visual perception. Readers easily processed 6 or 8 characters simultaneously in reading text. Similarly, a good portion of visual space could be processed simultaneously without very little loss of information. Only today are human computer interactionists developing visual displays that exploit the power of our visual information processing system [4,13].

Many scientists promised us, however, that the days of text were limited. Look around and observe that our natural communication medium is spoken language. Thus, a goal should be to interface humans and machines via speech. Speech synthesis has been fairly pervasive for at least two decades. In the middle 70's [9], we began using one of the first commercial synthesizers (a Swedish synthesizer that also spoke English; is this why they sound like drunken Swedes?). By the arrival of the first PCs in the early 80's, one could buy an IntexTalker for just $300. I remember how easily I could amuse my audience with a singing rendition of "Happy Birthday." Most of you won't be surprised to hear that speech synthesis still sucks. Evidence for the primitive state of auditory synthetic speech comes from fully animated films in which natural voices are used. Neal Stephenson, in his cyber fiction, transports us to the next century but the interactive, electronic, and animated books are read by real people called ractors rather than via synthetic speech [16].

Given the rapid progress on cheap memory and efficient search routines, concatenative synthesis offers a potentially great improvement over the traditional parameter synthesis [14]. For fairly limited vocabularies and constrained contextual interactions, concatenative synthesis is (not rapidly) approaching the real thing. The main barriers appear to be in the synthesis of prosody, suprasegmental quality, and emotion. However, I hope I live long enough to witness speech synthesis pass the Turing test—the reason being that I think this will mean that I'll live to be a ripe old age indeed.

Given this prelude, one would expect that I would not have optimistic words about speech recognition. True, inexpensive speaker-dependent recognizers are succeeding in the marketplace, and thus must be ready for prime time. However, a recent news report raved about some new speech technology but admitted that its presentation was not completely successful because the demonstrator was not able to make himself understood in order to get to the menu and control the system. Given this state of the art, it is not surprising that the executives of a leading technology firm refuse to do speech demos.

What will lead to significant improvements in speech recognition by machine? A truism is that lots of training data is required to educate the recognition system. Roy's [12] work on coordinating speech and visual context holds great promise for accumulating an unlimited supply of training data without the tedious labeling process. With this abundance of data, it will also be possible to train recognizers on speech from populations and situations that are highly similar to the application. A second enhancement is not to give up on what a linguist might contribute. In this regard, special phonologies could be added to the recognition systems to allow recognition systems to be more robust across speakers and situations. Some success with this method has been demonstrated by improved recognition of non-native speech [3]. Finally, and closest to my heart is the use of sources of information other than the auditory input. Several independent studies have shown that visible speech from the speaker's face improves the performance of speech recognition by machine. In this regard, Petajan [11] provides a helpful tutorial on how the MPEG-4 Face Animation standard can be fruitfully used in visible speech processing and integrated with other sources of information. I expect that other nonintrusive but informative devices can be appended to the user to improve recognition even more.

Speech versus text versus graphs is always a fun debate, one that doesn't seem to tire us. Although being ecumenical is not always a good strategy, in this case, it certainly is. Why give up one for the other when we can have all of them? It is obvious that there will be situations in which any one modality or interface will fail, and no one modality can be completely robust. Furthermore, having multiple interfaces is advantageous even when each one is informative or even sufficient. The reason is that we usually benefit from having available multiple sources of information.

2. PATTERN RECOGNITION

Pattern recognition is central to the human user's interaction with computers. We have learned from empirical and theoretical research that pattern recognition is achieved via a variety of bottom-up and top-down sources of information. A wide variety of results have been described within the framework of a fuzzy logical model of perception (FLMP). Within this model, perceivers are assumed to efficiently use multiple sources of information supporting identification and interpretation. The assumptions central to the model are 1) each source of information is evaluated to give the continuous degree to which that source specifies various alternatives, 2) the sources of information are evaluated in parallel and independently of one another, 3) the sources are integrated multiplicatively to provide an overall degree of support for each alternative, and 4) perceptual identification and interpretation follows the relative degree of support among the alternatives [7,8,10].

There is now overwhelming evidence that persons have continuous rather than simply categorical information in the processing of speech, text, and graphs [7,8]. Furthermore, parallel processing is supported by the findings that the information transmitted by one source is not reduced by the processing of another source. Independence of sources is motivated by the principle of category-conditional independence [8,10]: it isn't possible to predict the evaluation of one source on the basis of the evaluation of another, so the independent

0-7803-6536-4/00/$10.00 (c) 2000 IEEE

evaluation of both sources is necessary and sufficient to make an optimal category judgment. While sources are thus kept separate at evaluation, they are then integrated to achieve perception, recognition, and interpretation. Multiplicative integration yields a measure of total support for a given category identification. This operation, implemented in the FLMP, allows the combination of two imperfect sources of information to yield better performance than would be possible using either source by itself. However, the output of integration is an absolute measure of support; it must be relativized, which is implemented through a decision stage, which divides the support for one category by the summed support for all categories.

These behavioral results and theoretical explanation are substantiated by the neurophysiological results from experiments by Stein and his colleagues in the mature and developing cat and monkey superior colliculus (SC) [15]. The receptive field features that allow a neuron to code modality-specific information are not degraded during a multisensory response. Thus, a neuron's capability to code the directional movement or the size of a modality-specific stimulus will not lose that capability during a multisensory interaction. They observe a principle of synergy that describes the observation that the absolute magnitude of a multisensory response increases with the combination of increasingly effective modality-specific stimuli. The principle of inverse effectiveness describes the observations that the relative magnitude of this synergy decreases as the modality-specific stimuli are more effective when presented alone.

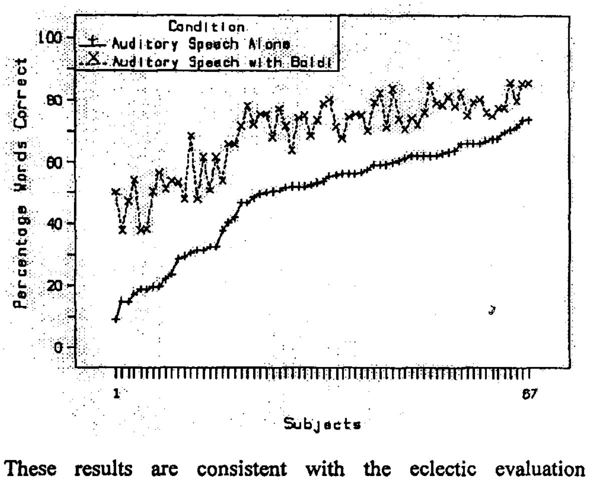
The FLMP turns out to be highly consistent with these neural "principles" by which cross-modal cues are integrated. These neural principles can be partitioned into the following categories: temporal, spatial, synergy, inverse effectiveness, and preservation of receptive field properties. The temporal and spatial principles predict and/or help explain how the effects of stimulus onset asynchronies between cross-modal stimuli and/or their relative spatial positions affect the likelihood that a given multisensory interaction will produce an enhanced or degraded response, and the relative magnitude of that response. These principles parallel the set of principles dictated by the FLMP in speech perception and other pattern recognition situations. Two modalities (e.g., audible and visible speech) are more informative than either one alone, depicting the principle of synergy. There is even the equivalent of the inverse effectiveness principle, derived from the single neuron studies described earlier, contained in the FLMP's predictions. The gain with two modalities becomes proportionately greater as the effectiveness of each source alone becomes weaker. Furthermore, just as in the case of observations at the single neuron level, the synergy created from having two sources increases with their temporal synchrony and spatial overlap [15]. It should be emphasized that the FLMP has been applied successfully across different modalities and different situations, qualifying it as a general law of pattern recognition behavior [8].

3. VALUE OF TALKING FACES IN DIALOG

Many communication environments involve situations that create a noisy auditory channel, which degrades speech perception and recognition. Visible speech from the talker's face (or from a reasonably accurate synthetic talking head) improves intelligibility in these situations [8,14]. If, for example, only roughly half of a degraded auditory message is understood, its pairing with visible speech can allow comprehension to be almost perfect. The strong influence of visible speech is not limited to situations with degraded auditory input. A perceiver's recognition of an auditory-visual syllable reflects the contribution of both sound and sight. For example, if the ambiguous auditory sentence, My bab pop me poo brive, is paired with the visible sentence, My gag kok me koo grive, the perceiver is likely to hear, My dad taught me to drive. Two ambiguous sources of information are combined to create a meaningful interpretation [8,10].

Visible speech helps everyone, not only trained speechreaders and persons with hearing loss. In a series of evaluation experiments, we asked college students to report the words of sentences presented in noise. An example sentence was "Pick up the pencil." On some trials, only the auditory sentence was presented. Figure 1 gives the proportion of words correctly reported under this condition for 67 different subjects. The most noticeable result in the figure is the tremendous individual variability in accuracy of identification of the auditory sentences. When these same sentences were accompanied by our animated head called Baldi (www.mambo.ucsc.edu/psl/pslfan.html), each of these 67 subjects showed a significant performance benefit. Some subjects with particularly poor auditory recognition benefitted the most. Although we did not ask these subjects to speechread the sentences in the absence of auditory speech, we know that their performance would have been only slightly better than chance. The combination of auditory and visual speech has been called super-additive because their combination can lead to accuracy greater than the sum of their separate accuracies.

Figure 1. Percentage of words correctly recalled for auditory sentences presented alone or with Baldi.



These results are consistent with the eclectic evaluation experiments carried out by Schroeter and his research group [14]. They found improved digit recognition when a talking head (TH) was added to digit presentations in a background of airport babble. Users were more satisfied when the TH accompanied an interactive information system, and a TH added to the success of e-commerce scenarios.

There are several reasons why the use of auditory and visual information together is so successful. These include a) robustness of visual speech, b) complementarity of auditory and visual speech, and c) optimal integration of these two sources of information. Speechreading, or the ability to obtain speech information from the face, is robust in that perceivers are fairly good at speech reading even when they are not looking directly at the talker's lips. Furthermore, accuracy is not dramatically reduced when the facial image is blurred (because of poor vision, for example), when the face is viewed from above, below, or in profile, or when there is a large distance between the talker and the viewer.

Complementarity of auditory and visual information simply means that one of the sources is strong when the other is weak. A distinction between two segments robustly conveyed in one modality is relatively ambiguous in the other modality. For example, the place difference between /ba/ and /da/ is easy to see but relatively difficult to hear. On the other hand, the voicing difference between /ba/ and /pa/ is relatively easy to hear but very difficult to discriminate visually. Two complementary sources of information make their combined use much more informative than would be the case if the two sources were non-complementary, or redundant [10].

The final characteristic is that perceivers combine or integrate the auditory and visual sources of information in an optimally efficient manner. There are many possible ways to treat two sources of information: use only the most informative source, average the two sources together, or integrate them in such a fashion in which both sources are used but that the least ambiguous source has the most influence. Perceivers in fact integrate the information available from each modality to perform as efficiently as possible. A wide variety of empirical results has been accurately predicted by the FLMP, which describes an optimally efficient process of combination.

4. IMPLICATIONS

Our experiences have convinced us that several new trends and challenges come to the forefront with technology-driven behavior. Returning to our initial example of an animated agent in education, several new roles for human teachers become apparent. Rather than actively teaching, the technology promotes the teacher to a more interactive role in the classroom. They become much more active, collaborative and effective, since they can watch each student interact with the program they designed, understand individual problems, and assist when necessary. The classroom becomes an interactive learning environment with as many tutors as students, and with the teacher monitoring learning. Within this new learning environment, teachers become less didactic and more collaborative and thus are implicitly fulfilling a goal of reflective rather than standard education [5].

5. ACKNOWLEDGEMENTS

The research was supported by grants from the National Institute of Deafness and Other Communicative Disorders (PHS R01 DC00236), the National Science Foundation (Challenge Grant CDA-9726363; Linguistics Program Grant BCS-9905176), Intel Corporation, and the University of California Digital Media Innovation Program. The author thanks Chris Campbell, Michael Cohen, Ron Cole, and members of the Perceptual Science Laboratory for their contributions.

6. REFERENCES

[1] Cohen, M. M., & Massaro, D. W. "Real-time speech synthesis," *Behavior Research Methods and Instrumentation, 8*, 189-196, 1976.

[2] Cole, R., Massaro, D. W., et al. New tools for interactive speech and language training: Using animated conversational agents in the classrooms of profoundly deaf children. *Proceedings of ESCA/Socrates sponsored Method and Tool Innovations for Speech Science Education workshop.* London: University College London, 1999.

[3] Kawai, G. "*Spoken language processing applied to nonnative language pronunciation learning*" Unpublished Ph.D. dissertation, University of Tokyo, Department of Information and Communication Engineering, 1999.

[4] Lehrer, R. & Chazan, D. (Eds.) *Designing learning environments for developing understanding of geormetry and space.* Mahwah, NJ: Erlbaum, 1999.

[5] Lipman, M. *Thinking in Education.* New York: Cambridge University Press, 1991.

[6] Massaro, D.W. "Consolidation and Interference in the Perceptual Memory System," *Perception and Psychophysics, 7*, 153-156, 1970.

[7] Massaro, D.W. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry.* Hillsdale, NJ: Lawrence Erlbaum Associates, 1987.

[8] Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle.* MIT Press: Cambridge, MA.

[9] Massaro, D. W., & Cohen, M. M., "The Contribution of Fundamental Frequency and Voice Onset Times to the /zi/-/si/ Distinction," *Journal of the Acoustical Society of America, 60*, 704-717, 1976.

[10] Massaro, D.W., & Stork, D.G. "Speech recognition and sensory integration." *American Scientist, 86*, 236-244, 1998.

[11] Petajan, E. Approaches to visual speech processing based on the MPEG-4 Face Animation standard. *Proceedings of IEEE International Conference on Multimedia and Expo, New York*, 2000.

[12] Roy, D. Learning form multimodal observations. *Proceedings of IEEE International Conference on Multimedia and Expo*, New York, 2000.

[13] Shneiderman, B. Direct Manipulation for Comprehensible, Predictable and Controllable User Interfaces *Proceedings of ACM International Workshop on Intelligent User Interfaces '97, ACM*, New York, NY, 33-39, 1997.

[14] Schroeter, J., Ostermann, J., et al. Multimodal speech synthesis. *Proceedings of IEEE International Conference on Multimedia and Expo*, New York, 2000.

[15] Stein, B. E., Laurienti, P. J., Stanford, T. R., & Wallace, M.T. Neural mechanisms for integrating informaiton from multiple senses. Proceedings of IEEE International Conference on Multimedia and Expo, New York, 2000.

[16] Stephenson, N. *The diamond age.* Bantam Books, 1996.

[17] Stewart, B. *The Lorenz Attractor*, 16mm film, Aerial Press, Santa Cruz, 1985.