



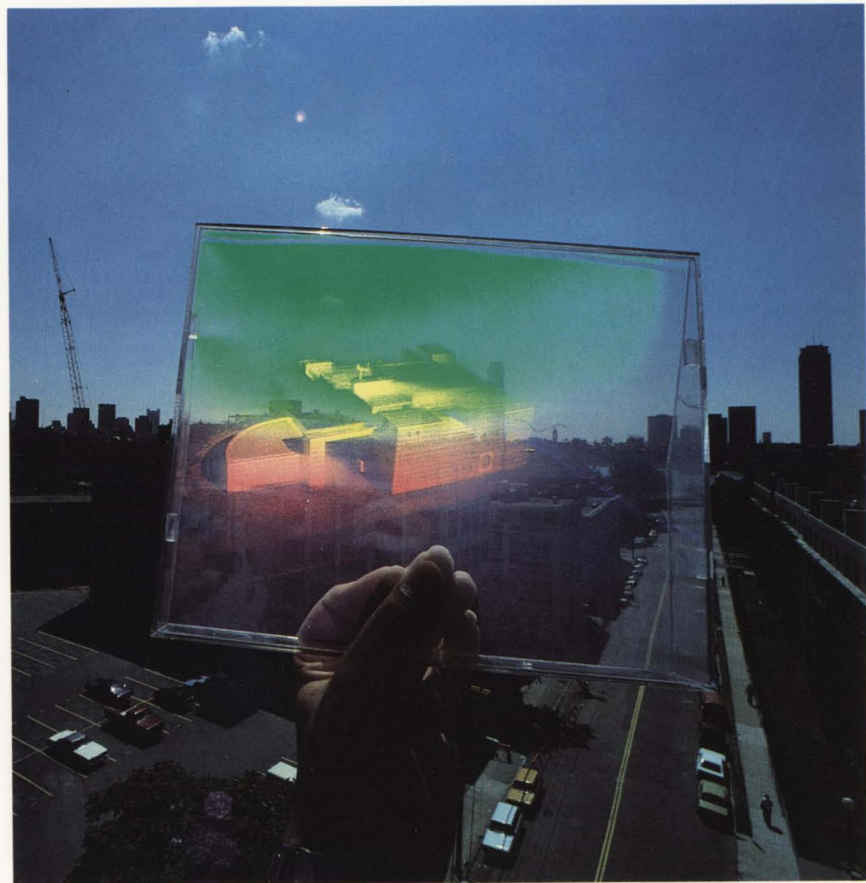
Media Technology Symposium

Massachusetts Institute of Technology

October 3, 1985

Kresge Auditorium

Jointly Sponsored by the Media Laboratory and the Industrial Liaison Program



8:15 am  
Registration  
Kresge Auditorium

8:45 am  
Introduction  
Professor Nicholas P. Negroponte  
Mr. J. Peter Bartl

9:20 am  
Television Past Broadcast  
Mr. Walter R. Bender  
Professor Andrew B. Lippman

10:00 am  
Advanced Television  
Professor William F. Schreiber

10:40 am  
Break

11:00 am  
Intelligent Telephones  
Mr. Christopher M. Schmandt

11:40 am  
Eyes as Output  
Dr. Richard A. Bolt

12:20 pm  
Lunch  
Sala de Puerto Rico, Student Center  
Luncheon Speaker:  
Dr. Jerome B. Wiesner

2:00 pm  
Synthetic Holography  
Professor Stephen A. Benton

2:35 pm  
Realistic Computer Animation  
Professor David L. Zeltzer

3:10 pm  
Break

3:30 pm  
Synthetic Performers  
Professor Barry L. Vercoe

4:05 pm  
Computers and Creativity  
Mr. Marvin Denicoff  
Professor Marvin L. Minsky

5:00 pm  
Reception  
The Wiesner Building Atrium  
20 Ames Street



9:20 am

### Television Past Broadcast

Mr. Walter R. Bender  
Media Laboratory

Professor Andrew B. Lippman  
Media Laboratory

Interactive television represents a change as fundamental to the world of broadcasting as television itself was when introduced to an existing world of broadcast radio. In this emerging field, programmable computing is integrated with the video channel at the points of origination, distribution and reception. As a result, the audience for "broadcasts" may be programmable machines which process the information before presenting it to a human user, rather than the traditional mode in which the human is only able to choose from whatever is being broadcast at the moment. Similarly, the program itself may be automatically derived from a database rather than by a producer. Already the advent of simply programmable storage has resulted in time-

shifting complete programs for the convenience of the viewer. In the future, programs might be better described as dynamic, personal magazines than movies, the viewing of which takes on the characteristics of an individual dialogue with the author. Both new hardware and new software are needed before this vision can be realized. It requires the availability of large-scale storage and powerful processing at the receiver, in addition to the use of the video channel as a joint data and image transmission medium. Software design is more subtle and rests upon innovations in information publishing and the human interface. This presentation will present several approaches by discussing experiments and research in progress.

10:00 am

### Advanced Television

Professor William F. Schreiber  
Research Laboratory of Electronics  
Media Laboratory

The Advanced Television Research Program is sponsored by 10 US television broadcasters and equipment manufacturers who recognize that new technologies will soon enable television systems to be designed with vastly improved image and sound quality. The Japan Broadcasting Company has already proposed such a system, but it requires five times the channel capacity currently used. Other laboratories around the world have demonstrated systems that do sophisticated signal processing at the receiver and/or the transmitter to produce significantly improved quality with only moderate or, in some cases, no expansion of the channel capacity. The more visionary workers believe that it will be possible

to obtain quality equalling or even exceeding that of 35mm motion pictures within the present TV channels. How such a new television system might be brought to the marketplace, and how much, if anything, consumers would be willing to pay for improved quality are important issues that, although more difficult to answer with any degree of confidence, must also be addressed. This presentation will describe the program including a large computer facility that is being assembled to permit simulation of high definition moving images.

11:00 am  
Intelligent Telephones

Mr. Christopher M. Schmandt  
Media Laboratory

Integration of voice recognition, voice synthesis and digital recording can enhance both the message-carrying capability as well as accessibility of intelligent nodes on a telecommunication network. Voice may be used as both a data and a control channel to interface with these nodes. In the latter mode, particular concerns are: successful parsing of recognizer output, modeling dialogues on human conversational behavior and expectations, and correct determination of the intent of the speaker to address either the machine or the remote site being accessed through the machine. These and other issues of machine mediated voice communication will be discussed.

11:40 am  
Eyes as Output

Dr. Richard A. Bolt  
Media Laboratory

To date, the use of eye tracking has been limited to such studies as determining where people look when they read, how pilots scan aircraft cockpit instrument arrays, or which characters kids tend to watch in Sesame Street episodes. Eye movements have yet to be considered seriously as part of the essential information that a computer should have to help infer its user's attention and intents. The laboratory's Human Interface Group is beginning to explore how computer awareness of user eye movements, either alone or in combination with user speech and gesture, can contribute to the richness and naturalness of human/computer dialogue.

2:00 pm  
Synthetic Holography

Professor Stephen A. Benton  
Media Laboratory

In synthetic holography, computer-processed data is translated into fully 3-dimensional holographic images. Applications that are either being addressed currently or are being considered for future projects include processing CAD/CAM, medical and seismic data, in addition to images for use in education and entertainment. These data are converted to holograms by a variety of techniques, including the computation of entire interference fringe systems and the superposition of elemental point holograms. This talk will outline the development of holographic imaging, emphasizing recent work on new processes to optically combine progressions of computed perspective views that are made visible only at

specified angles by synthetic holographic elements. The holograms created through these techniques present images that can be viewed in white light without special glasses, and which can be examined over a wide range of angles to present an impression of a solid 3-dimensional object.



2:35 pm

### Realistic Computer Animation

Professor David L. Zeltzer  
Media Laboratory

Striking realism for certain classes of objects and scenes is achieved with current computer graphics technology. In the future, it should be possible to render similar realism to the behavior, as well as the appearance, of objects and characters in virtual microworlds. Thus, the computer will finally become a medium for story telling, scientific modelling and creative expression. To achieve this, it will be necessary to design computers capable of interactively specifying and controlling the actions of articulated figures ultimately in real-time. This presentation will discuss current research aimed at integrating robotic and knowledge representation concepts with some new techniques to create a powerful animation workstation.

3:30 pm

### Synthetic Performers

Professor Barry L. Vercoe  
Media Laboratory

Technology in the performing arts has so far meant tools, devices and amplifiers—powerful participants with no artistic awareness, whose eyes and ears are merely the buttons, switches and mice of standard interfaces. A new initiative to put the power of technology inside the arts is demonstrated by the Synthetic Performer. Here a computer is made to understand the nuances of chamber music performers so that it can behave as a sensitive and expressive member of an ensemble. It can follow players and conductors, and can participate in music from the Baroque to modern idioms as a skilled and well-rehearsed performer. Composers of the future will routinely write for ensembles of live and synthetic performers and we can already demonstrate what that ultra-expressive world will be like.

4:05 pm

### Computers and Creativity

Mr. Marvin Denicoff  
Media Laboratory

Professor Marvin L. Minsky  
Artificial Intelligence Laboratory  
Media Laboratory

Utilizing the context of theatre, film, and television, this talk will discuss research to enhance the role of modern computation as a person's collaborator in the creative process. Well beyond the already accepted status of machines as administrative and secretarial aids, the presentation will explore such possibilities as: facilitating intelligent interaction via computer understanding of speech, text and drawings; and mutual learning by human and machine with dynamic assignment of responsibilities across machines, and people and electronic

networking as a mechanism to encourage collaboration across geographically separated human artists. Primitive examples will be given of the potential of computers to imitate and eventually to extend the writing, painting and performing styles of creative artists.

## Media Technologies

Chairman:

Professor Nicholas P. Negroponte  
Director, Media Laboratory

Industrial Liaison Program

Representative:

Mr. J. Peter Bartl

MIT's newly created interdisciplinary laboratory, the Media Laboratory, has as its purpose the invention and creative use of new communications, information processing, and storage technologies. Researchers are engaged concurrently in the advancement of emerging disciplines and in the development of computer, video, audio and print media. The driving force for the advancement of these disciplines is seen as lying in sophisticated applications and their regular use in daily life. The purpose of this symposium is to expose the research activities of the laboratory in such general areas of research as the school of the future, home computing, telecommunication, and the arts. Specific projects

include: intelligent telephones, personalized newspapers, television past broadcast, electronic cinema, and sensory-rich human-computer interfaces. The symposium coincides with the opening of the new laboratory facilities and their formal dedication that culminates a six-year effort to create an intellectual and physical environment for what may well prove to be the most important developments in computer science and technology.

### Credits

Program Design  
Jacqueline S. Casey and Sylvia Steiner  
MIT Design Services

Typesetting  
International Phototypesetters, Inc.

Printing  
Daniels Printing Company



---

# Media Technologies

---

Thursday, October 3, 1985  
Kresge Auditorium, MIT

---

**Symposium Chairman:**

**Professor Nicholas P. Negroponte**  
**Director, Media Laboratory**

---



Industrial Liaison Program  
Massachusetts Institute of Technology

## MEDIA TECHNOLOGIES

Chairman:

Professor Nicholas P. Negroponte

Industrial Liaison Program Representative:

Mr. J. Peter Bartl

MIT's newly created interdisciplinary laboratory, the Media Laboratory, has as its purpose the invention and creative use of new communications, information processing, and storage technologies. Researchers are engaged concurrently in the advancement of emerging disciplines and in the development of computer, video, audio and print media. The driving force for the advancement of these disciplines is seen as lying in sophisticated applications and their regular use in daily life. The purpose of this symposium is to expose the research activities of the laboratory in such general areas of research as the school of the future, home computing, telecommunication and the arts. Specific projects include: intelligent telephones, personalized newspapers, television beyond broadcast, electronic cinema, and sensory-rich human-computer interfaces. The symposium coincides with the opening of the new laboratory facilities and their formal dedication that culminates a six-year effort to create an intellectual and physical environment for what may well prove to be the most important developments in computer science and technology.



AGENDA  
October 3, 1985

- 8:15 Registration - Kresge Auditorium
- 8:45 INTRODUCTION  
Professor Nicholas P. Negroponte  
Mr. J. Peter Bartl
- 9:20 TELEVISION PAST BROADCAST  
Professor Andrew B. Lippman
- 10:00 ADVANCED TELEVISION  
Professor William F. Schreiber
- 10:40 Break
- 11:00 INTELLIGENT TELEPHONES  
Mr. Christopher M. Schmandt
- 11:40 EYES AS OUTPUT  
Dr. Richard A. Bolt
- 12:20 Lunch - Sala de Puerto Rico, Student Center  
  
Luncheon Speaker:  
Dr. Jerome B. Wiesner
- 2:00 SYNTHETIC HOLOGRAPHY  
Professor Stephen Benton
- 2:35 REALISTIC COMPUTER ANIMATION  
Professor David Zeltzer
- 3:10 Break
- 3:30 SYNTHETIC PERFORMERS  
Professor Barry L. Vercoe
- 4:05 COMPUTERS AND CREATIVITY  
Dr. Marvin Denicoff  
Professor Marvin L. Minsky
- 5:00 Reception - The Wiesner Building Atrium  
20 Ames Street

9:20

TELEVISION PAST BROADCAST

Professor Andrew B. Lippman  
Department of Architecture  
Media Laboratory

Interactive television represents a change as fundamental to the world of broadcasting as television itself was when introduced to an existing world of broadcast radio. In this emerging field, programmable computing is integrated with the video channel at the points of origination, distribution and reception. As a result, the audience for "broadcasts" may be programmable machines which process the information before presenting it to a human user, rather than the traditional mode in which the human is only able to choose from whatever is being broadcast at the moment. Similarly, the program itself may be automatically derived from a database rather than by a producer. Already the advent of simply programmable storage has resulted in time-shifting complete programs for the convenience of the viewer. In the future, programs might be better described as dynamic, personal magazines than movies, the viewing of which takes on the characteristics of an individual dialogue with the author. Both new hardware and new software are needed before this vision can be realized. It requires the availability of large scale storage and powerful processing at the receiver in addition to the use of the video channel as a joint data and image transmission medium. Software design is more subtle and rests upon innovations in information publishing and the human interface. This presentation will present several approaches by discussing experiments and research in progress.

10:00

ADVANCED TELEVISION

Professor William F. Schreiber  
Department of Electrical Engineering & Computer Science  
Research Laboratory of Electronics  
Media Laboratory

The Advanced Television Research Program is sponsored by 10 U.S. television broadcasters and equipment manufacturers who recognize that new technologies will soon enable television systems to be designed with vastly improved image



and sound quality. The Japan Broadcasting Company has already proposed such a system but it requires five times the channel capacity currently used. Other laboratories around the world have demonstrated systems that do sophisticated signal processing at the receiver and/or the transmitter to produce significantly improved quality with only moderate or, in some cases, no expansion of the channel capacity. The more visionary workers believe that it will be possible to obtain quality equalling or even exceeding that of 35mm motion pictures within the present TV channels. How such a new television system might be brought to the marketplace, and how much, if anything, consumers would be willing to pay for improved quality are important issues that, although more difficult to answer with any degree of confidence, must also be addressed. This presentation will describe the program including a large computer facility that is being assembled to permit simulation of high definition moving images.

11:00  
INTELLIGENT TELEPHONES

*General - museum political*

Mr. Christopher M. Schmandt  
Media Laboratory

Integration of voice recognition, voice synthesis and digital recording can enhance both the message carrying capability as well as accessibility of intelligent nodes on a telecommunication network. Voice may be used as both a data and a control channel to interface with these nodes. In the latter mode, particular concerns are: successful parsing of recognizer output, modelling dialogues on human conversational behavior and expectations, and correct determination of the intent of the speaker to address either the machine or the remote site being accessed through the machine. These and other issues of machine mediated voice communication will be discussed.

11:40

EYES AS OUTPUT

Dr. Richard A. Bolt  
Media Laboratory

To date, the use of eye tracking has been limited to such studies as determining where people look when they read, how pilots scan aircraft cockpit instrument arrays or which characters kids tend to watch in Sesame Street episodes. Eye movements have yet to be considered seriously as part of the essential information that a computer should have to help infer its user's attention and intents. The Laboratory's Human Interface Group is beginning to explore how computer awareness of user eye movements, either alone or in combination with user speech and gesture, can contribute to the richness and naturalness of human/computer dialogue.

2:00

SYNTHETIC HOLOGRAPHY

Professor Stephen Benton  
Department of Architecture  
Media Laboratory

30; Sesimai, independent applics.  
Jeffrey Kunkin, new hire member  
- compute the holograms

In synthetic holography, computer-processed data is translated into fully 3-dimensional holographic images. Applications that are either being addressed currently or are being considered for future projects include processing CAD/CAM, medical and seismic data in addition to images for use in education and entertainment. These data are converted to holograms by a variety of techniques, including the computation of entire interference fringe systems and the superposition of elemental point holograms. This talk will outline the development of holographic imaging, emphasizing recent work on new processes to optically combine progressions of computed perspective views that are made visible only at specified angles by synthetic holographic elements. The holograms created through these techniques present images that can be viewed in white light without special glasses, and which can be examined over a wide range of angles to present an impression of a solid 3-dimensional object.



2:35

## REALISTIC COMPUTER ANIMATION

Professor David Zeltzer  
Department of Architecture  
Media Laboratory

Striking realism for certain classes of objects and scenes is achieved with current computer graphics technology. In the future it should be possible to render similar realism to the behavior, as well as the appearance, of objects and characters in virtual microworlds. Thus, the computer will finally become a medium for story telling, scientific modelling and creative expression. To achieve this, it will be necessary to design computers capable of interactively specifying and controlling the actions of articulated figures ultimately in real time. This presentation will discuss current research aimed at integrating robotic and knowledge representation concepts with some new techniques to create a powerful animation workstation.

3:30

## SYNTHETIC PERFORMERS

Professor Barry L. Vercoe  
Department of Humanities  
Media Laboratory

Technology in the performing arts has so far meant tools, devices and amplifiers—powerful participants with no artistic awareness, whose eyes and ears are merely the buttons, switches and mice of standard interfaces. A new initiative to put the power of technology inside the arts is demonstrated by the Synthetic Performer. Here a computer is made to understand the nuances of chamber music performers so that it can behave as a sensitive and expressive member of an ensemble. It can follow players and conductors, and can participate in music from the Baroque to modern idioms as a skilled and well-rehearsed performer. Composers of the future will routinely write for ensembles of live and synthetic performers and we can already demonstrate what that ultra-expressive world will be like.

4:05  
COMPUTERS AND CREATIVITY



Dr. Marvin Denicoff  
Media Laboratory

Professor Marvin L. Minsky  
Department of Electrical Engineering & Computer Science  
Artificial Intelligence Laboratory

Utilizing the context of theater, film and television, this talk will discuss research to enhance the role of modern computation as man's collaborator in the creative process. Well beyond the already accepted status of machines as administrative and secretarial aids, the presentation will explore such possibilities as: facilitating intelligent interaction via computer understanding of speech, text and drawings; mutual learning by man and machine with dynamic assignment of responsibilities across machines and people and electronic networking as a mechanism to encourage collaboration across geographically separated human artists. Primitive examples will be given of the potential of computers to imitate and eventually to extend the writing, painting and performing styles of creative artists.

Minsky

AI: Computer Prove Math Theorems of Russell's Book

Newell & Simon

: Early chess programs

: Jim Slagle 1961 A on MIT Text

Evolution, Success of Human?



## REGISTRATION INFORMATION

Symposium Attendance is primarily for invited guests and representatives from member companies of the Industrial Liaison Program. There is no registration fee for this symposium. Notice of plans to attend should be received by September 29, 1985. Registration cards included in this program should be sent to:

Constance Marino Bonanno  
Conference Coordinator  
Massachusetts Institute of Technology  
292 Main Street - Room E38-520  
Cambridge, MA 02139

Telephone registrations will also be accepted at (617) 253-0424; telex 921473. Please note that we do not confirm registrations; if you must confirm your symposium registration, please do so no later than one week before the symposium.

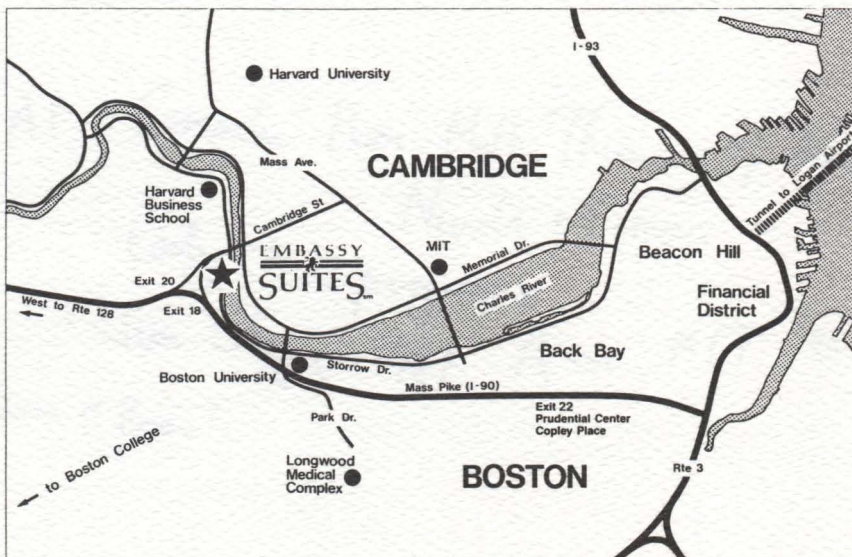
## HOTEL INFORMATION

For your convenience in obtaining accommodations, a block of rooms has been set aside at the Embassy Suites Hotel in Cambridge. Rooms are available the nights of September 29, 30 and October 1-5, 1985. The group rate for suite accommodations is \$105 single and \$125 double. There are a limited number of double occupancy rooms available at \$95 single and \$115 double.

\*To be assured of a room and the special rate, please make your reservation directly with the hotel as early as possible, but no later than August 29, 1985. When making your reservation, please mention the "MIT Media Technology Symposium and Dedication."

For reservations, contact:

**EMBASSY SUITES HOTEL**  
400 Soldiers Field Road  
Boston, Massachusetts 02134  
(617) 783-0090



When exiting from Mass Turnpike (I-90) at exit 18 or 20, follow signs to Cambridge.



Symposium Registration:  
MEDIA TECHNOLOGIES  
October 3, 1985

(Please print or type)

NAME (Dr. Mr. Ms.) \_\_\_\_\_

TITLE \_\_\_\_\_

DIVISION \_\_\_\_\_

COMPANY \_\_\_\_\_

ADDRESS \_\_\_\_\_

\_\_\_\_\_

TELEPHONE \_\_\_\_\_

\* Please indicate if you will attend the reception  
at the Wiesner Building at 5:00:

I will attend       I will not attend

PLACE IN ENVELOPE AND RETURN TO:  
Conference Coordinator, Industrial Liaison Program - MIT,  
Room E38-516, 292 Main Street, Cambridge, MA 02139

Symposium Registration:  
MEDIA TECHNOLOGIES  
October 3, 1985

(Please print or type)

NAME (Dr. Mr. Ms.) \_\_\_\_\_

TITLE \_\_\_\_\_

DIVISION \_\_\_\_\_

COMPANY \_\_\_\_\_

ADDRESS \_\_\_\_\_

\_\_\_\_\_

TELEPHONE \_\_\_\_\_

\* Please indicate if you will attend the reception  
at the Wiesner Building at 5:00:

I will attend       I will not attend

PLACE IN ENVELOPE AND RETURN TO:  
Conference Coordinator, Industrial Liaison Program - MIT,  
Room E38-516, 292 Main Street, Cambridge, MA 02139

Massachusetts Institute of Technology  
Industrial Liaison Program, Room E38-510  
77 Massachusetts Avenue  
Cambridge, Massachusetts 02139

1076 81593



Phoned June 7: Has been occupied; will get back.  
him

May 4 1984

Professor Patrick Purcell  
Architecture Machine Group  
9-522 MIT  
77 Massachusetts Avenue  
Cambridge  
MA 02139

Dear Patrick

Thank you for showing me some of the recent work that has been done in your group. Although I think our public would be interested in practically all the projects, I think it will be realistic to introduce them only to the interactive use of videodisc as with the Picasso or architecture file, the graphical marionette and to the hologram.

Most of our visitors would probably not grasp the full idea of the interactive video disc unless we had some version of a working demonstration in the gallery. I am not sure that the demonstration video disc I saw would, on its own, get the ideas across to the general public.

The graphical marionette would be best displayed on a video tape player, perhaps as part of a longer program. The sequence you showed me lasted under a minute.

We would be very pleased to be able to exhibit a white light hologram of a three-dimensional computer-synthesised scene. We could accompany the hologram with a subset of the series of prints that were used in making up the hologram.

Our opening date is November 12 1984. We will need to know what we can display by the end of June at the latest. As we are in the process of laying out the gallery plan now, the sooner we can fix your contribution, the better.

Many thanks for your help and interest.

Yours sincerely

Dr Oliver Strimpel  
Curator

The  
Computer  
Museum

12 January 1984

Professor Patrick Purcell  
Architecture Machine Group  
9-522 Massachusetts Institute of Technology  
77 Massachusetts Ave  
Cambridge  
MA 02139

Dear Professor Purcell

Further to our telephone conversation, I am enclosing an outline proposal for the gallery "The Computer and the Image" to open at the Computer Museum at the end of this year. As you can see, I would like to have some interactive displays and we have already had some good response from various manufacturers who might be willing to donate or lend equipment.

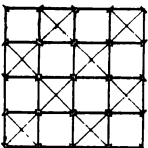
I also enclose two recent quarterly reports of the Museum and also an article from the February 1983 issue of Discover.

I look forward to speaking to you around the end of next week.

Yours sincerely

Oliver Strimpel  
Curator

enclosures





Oct 3 1985 MIT Media Lab Symposium

Andy Lippman - Electronic Celestium / Morris - interactive

Doug...?  
Walter

Electronic newspaper.

Computer edits personally for you.

Political here knob; content knob

Hand copy needed; TV animation instead of pictures

Bill Schroeder : High res TV as good as 35mm film a still photo

Re ARDS, Machom : Alvin Roth 714-856 6945

Terminal

(Instigator of Tenak)

Early use of

California

U of Davis

Leo Beranek re Teleputer

Characteristic generator chip Rob Stutz G being price down.

Tucker - ESP Tallahassee, Fla, early proj TV

# MIT Industrial Liaison Program

## Report

Paper relevant to the symposium:

"MEDIA TECHNOLOGIES"  
October 3, 1985

"Computers and Creativity"  
by  
Professor Marvin L. Minsky

1) "Music, Mind, and Meaning," by Professor Marvin L. Minsky

This paper has been duplicated at the request of the speaker.



Distributed for Internal Use  
by Member Companies Only.  
May Not be Reproduced.

© MIT

---

# Computer Music Journal

Volume 5, Number 3

Fall 1981

---

## Marvin Minsky

Artificial Intelligence Laboratory  
Massachusetts Institute of Technology  
Cambridge, Massachusetts 02139

### Why Do We Like Music?

Why do we like music? Our culture immerses us in it for hours each day, and everyone knows how it touches our emotions, but few think of how music touches other kinds of thought. It is astonishing how little curiosity we have about so pervasive an "environmental" influence. What might we discover if we were to study musical thinking?

Have we the tools for such work? Years ago, when science still feared meaning, the new field of research called *artificial intelligence* (AI) started to supply new ideas about "representation of knowledge" that I'll use here. Are such ideas too alien for anything so subjective and irrational, aesthetic, and emotional as music? Not at all. I think the problems are the same and those distinctions wrongly drawn: only the surface of reason is rational. I don't mean that understanding emotion is easy, only that understanding reason is probably harder. Our culture has a universal myth in which we see emotion as more complex and obscure than intellect. Indeed, emotion might be "deeper" in some sense of prior evolution, but this need not make it harder to understand; in fact, I think today we actually know much more about emotion than about reason.

Certainly we know a bit about the obvious processes of reason—the ways we organize and represent ideas we get. But whence come those ideas that so conveniently fill these envelopes of order? A poverty of language shows how little this concerns us: we "get" ideas; they "come" to us; we are "reminded of" them. I think this shows that ideas come from processes obscured from us and with which our surface thoughts are almost uninvolved. Instead, we are entranced with our emotions, which are so easily observed in others and ourselves. Per-

## Music, Mind, and Meaning

haps the myth persists because emotions (by their nature) draw attention, while the processes of reason (much more intricate and delicate) must be private and work best alone.

The old distinctions among emotion, reason, and aesthetics are like the earth, air, and fire of an ancient alchemy. We will need much better concepts than these for a working psychic chemistry.

Much of what we now know of the mind emerged in this century from other subjects once considered just as personal and inaccessible but which were explored, for example, by Freud in his work on adults' dreams and jokes, and by Piaget in his work on children's thought and play. Why did such work have to wait for modern times? Before that, children seemed too childish and humor much too humorous for science to take them seriously.

Why do we like music? We all are reluctant, with regard to music and art, to examine our sources of pleasure or strength. In part we fear success itself—we fear that understanding might spoil enjoyment. Rightly so: art often loses power when its psychological roots are exposed. No matter; when this happens we will go on, as always, to seek more robust illusions!

I feel that music theory has gotten stuck by trying too long to find universals. Of course, we would like to study Mozart's music the way scientists analyze the spectrum of a distant star. Indeed, we find some almost universal practices in every musical era. But we must view these with suspicion, for they might show no more than what composers then felt *should* be universal. If so, the search for truth in art becomes a travesty in which each era's practice only parodies its predecessor's prejudice. (Imagine formulating "laws" for television screenplays, taking them for natural phenomenon uninfluenced by custom or constraint of commerce.)

The trouble with the search for universal laws of thought is that both memory and thinking interact and grow together. We do not just learn about things, we learn *ways to think* about things; then we learn to think about thinking itself. Before long,

---

This is a revised and updated version of A.I. Memo No. 616. The earlier version will also appear in *Music, Mind, and Brain: The Neuropsychology of Music* edited by Manfred Clynes, and published by Plenum, New York.  
© 1981 by Marvin Minsky

---

our ways of thinking become so complicated that we cannot expect to understand their details in terms of their surface operation, but we might understand the principles that guide their growth. In much of this article I will speculate about how listening to music engages the previously acquired personal knowledge of the listener.

It has become taboo for music theorists to ask why we like what we like: our seekers have forgotten what they are searching for. To be sure, we can't account for tastes, in general, because people have various preferences. But this means only that we have to find the causes of this diversity of tastes, and this in turn means we must see that music theory is not only about music, but about how people process it. To understand any art, we must look below its surface into the psychological details of its creation and absorption.

If explaining minds seems harder than explaining songs, we should remember that sometimes enlarging problems makes them simpler! The theory of the roots of equations seemed hard for centuries within its little world of real numbers, but it suddenly seemed simple once Gauss exposed the larger world of (so-called) complex numbers. Similarly, music should make more sense once seen through listeners' minds.

### **Sonata as Teaching Machine**

Music makes things in our minds, but afterward most of them fade away. What remains? In one old story about Mozart, the wonder child hears a lengthy contrapuntal mass and then writes down the entire score. (I do not believe such tales, for history documents so few of them that they seem to be mere legend, though by that argument Mozart also would seem to be legend.) Most people do not even remember the themes of an evening's concert. Yet, when the tunes are played again, they are recognized. Something must remain in the mind to cause this, and perhaps what we learn is not the music itself but a way of hearing it.

Compare a sonata to a teacher. The teacher gets the pupils' attention, either dramatically or by the quiet trick of speaking softly. Next, the teacher

presents the elements carefully, not introducing too many new ideas or developing them too far, for until the basics are learned the pupils cannot build on them. So, at first, the teacher repeats a lot. Sonatas, too, explain first one idea, then another, and then recapitulate it all. (Music has many forms and there are many ways to teach. I do not say that composers consciously intend to teach at all, yet they are masters at inventing forms for exposition, including those that swarm with more ideas and work our minds much harder.)

Thus *expositions* show the basic stuff—the atoms of impending chemistries and how some simple compounds can be made from those atoms. Then, in *developments*, those now-familiar compounds, made from bits and threads of beat and tone, can clash or merge, contrast or join together. We find things that do not fit into familiar frameworks hard to understand—such things seem meaningless. I prefer to turn that around: a thing has meaning only after we have learned some ways to represent and process what it means, or to understand its parts and how they are put together.

What is the difference between merely knowing (or remembering, or memorizing) and understanding? We all agree that to understand something we must know what it means, and that is about as far as we ever get. I think I know why that happens. A thing or idea seems meaningful only when we have several different ways to represent it—different perspectives and different associations. Then we can turn it around in our minds, so to speak: however it seems at the moment, we can see it another way and we never come to a full stop. In other words, we can *think* about it. If there were only one way to represent this thing or idea, we would not call this representation thinking.

So something has a "meaning" only when it has a few; if we understood something just one way, we would not understand it at all. That is why the seekers of the "real" meanings never find them. This holds true especially for words like *understand*. That is why sonatas start simply, as do the best of talks and texts. The basics are repeated several times before anything larger or more complex is presented. No one remembers word for word all that is said in a lecture or all notes that are played

Fig. 1. Introductory measures of Ludwig van Beethoven's Symphony No. 5 in C Minor.

The musical score for the introductory measures of Beethoven's Symphony No. 5 in C Minor is presented in two systems. The first system includes the woodwinds and percussion: Flutes, Oboes, Clarinets in B $\flat$ , Bassoons, Horns in E $\flat$ , Trumpets in C, and Timpani in C, G. The second system includes the strings: Violin I, Violin II, Viola, Cello, and Bass. The tempo is marked "Allegro con brio ( $\text{♩} = 108$ )". The key signature is C minor. The score shows the first four notes of the first subject, which are repeated in various instruments.

in a piece. Yet if we have understood the lecture or piece once, we now "own" new networks of knowledge about each theme and how it changes and relates to others. No one could remember all of Beethoven's *Fifth Symphony* from a single hearing, but neither could one ever again hear those first four notes as just four notes! Once a tiny scrap of sound, these four notes have become a known thing—a locus in the web of all the other things we know and whose meanings and significances depend on one another (Fig. 1).

Learning to recognize is not the same as memorizing. A mind might build an *agent* that can sense a certain stimulus, yet build no agent that can reproduce it. How could such a mind learn that the first half-subject of Beethoven's *Fifth*—call it *A*—

prefigures the second half—call it *B*? It is simple: an agent *A* that recognizes *A* sends a message to another agent *B*, built to recognize *B*. That message serves to "lower *B*'s threshold" so that after *A* hears *A*, *B* will react to smaller hints of *B* than it would otherwise. As a result, that mind "expects" to hear *B* after *A*; that is, it will discern *B*, given fewer or more subtle cues, and might "complain" if it cannot. Yet that mind cannot reproduce either theme in any generative sense. The point is that interagent messages need not be in surface music languages, but can be in codes that influence certain other agents to behave in different ways.

(Andor Kovach pointed out to me that composers do not dare use this simple, four-note motive any more. So memorable was Beethoven's treatment

14

Musical score for measures 14-26. The score is written for four staves (two systems of two staves each). The notation includes various rhythmic values, accidentals, and dynamic markings. The first system (measures 14-15) features a *p cresc.* marking. The second system (measures 16-17) features a *cresc.* marking. The third system (measures 18-19) features a *cresc.* marking. The fourth system (measures 20-21) features a *cresc.* marking. The fifth system (measures 22-23) features a *cresc.* marking. The sixth system (measures 24-25) features a *cresc.* marking. The seventh system (measure 26) features a *cresc.* marking.

27

Musical score for measures 27-30. The score is written for four staves (two systems of two staves each). The notation includes various rhythmic values, accidentals, and dynamic markings. The first system (measures 27-28) features a *cresc.* marking. The second system (measures 29-30) features a *cresc.* marking.



The image displays two systems of musical notation for a piano piece. The first system covers measures 39 to 54, and the second system covers measures 51 to 54. The notation is arranged in two systems, each with four staves. The first system (measures 39-54) features a complex texture with multiple voices, including a prominent melodic line in the upper right voice and a dense accompaniment in the lower voices. The second system (measures 51-54) includes a section marked 'A' and a dynamic marking of 'p dolce' (piano dolce) in the lower right voice. The score is written in a standard musical notation style with various clefs, time signatures, and dynamic markings.

that now an accidental hint of it can wreck another piece by unintentionally distracting the listener.)

If sonatas are lessons, what are the subjects of those lessons? The answer is in the question! One thing the *Fifth Symphony* taught us is how to hear those first four notes. The surface form is just *descending major third, first tone repeated thrice*. At first, that pattern can be heard two different ways: (1) *fifth and third in minor mode* or (2) *third and first, in major*. But once we have heard the symphony, the latter is unthinkable—a strange constraint to plant in all our heads! Let us see how it is taught.

The *Fifth* declares at once its subject, then its near-identical twin. First comes the theme. Presented in a stark orchestral unison, its minor mode location in tonality is not yet made explicit, nor is its metric frame yet clear: the subject stands alone in time. Next comes its twin. The score itself leaves room to view this transposed counterpart as a complement or as a new beginning. Until now, fermatas have hidden the basic metric frame, a pair of twinned four-measure halves. So far we have only learned to hear those halves as separate wholes.

The next four-measure metric half-frame shows three versions of the subject, one on each ascending pitch of the tonic triad. (Now we are sure the key is minor.) This shows us how the subject can be made to overlap itself, the three short notes packed perfectly inside the long tone's time-space. The second half-frame does the same, with copies of the complement ascending the dominant seventh chord. This fits the halves together in that single, most familiar, frame of harmony. In rhythm, too, the halves are so precisely congruent that there is no room to wonder how to match them—and attach them—into one eight-measure unit.

The next eight-measure frame explains some more melodic points: how to smooth the figure's firmness with passing tones and how to counterpoise the subject's own inversion inside the long note. (I think that this evokes a sort of sinusoidal motion-frame idea that is later used to represent the second subject.) It also illustrates compression of harmonic time; seen earlier, this would obscure the larger rhythmic unit, but now we know enough

to place each metric frame precisely on the after-image of the one before.

Cadence. Silence. Almost. Total.

Now it is the second subject-twin's turn to stand alone in time. The conductor must select a symmetry: he or she can choose to answer prior cadence, to start anew, or to close the brackets opened at the very start. (Can the conductor do all at once and maintain the metric frame?) We hear a long, long unison F (subdominant?) for, underneath that silent surface sound, we hear our minds rehearsing what was heard.

The next frame reveals the theme again, descending now by thirds. (We see that it was the dominant ninth, not subdominant at all. The music fooled us that time, but never will again.) Then *tour de force*: the subject climbs, sounding on every scale degree. This new perspective shows us how to see the four-note theme as an appoggiatura. Then, as it descends on each tonic chord-note, we are made to see it as a fragment of arpeggio. That last descent completes a set of all four possibilities, harmonic and directional. (Is this deliberate didactic thoroughness, or merely the accidental outcome of the other symmetries?) Finally, the theme's melodic range is squeezed to nothing, yet it survives and even gains strength as single tone. It has always seemed to me a mystery of art, the impact of those moments in quartets when texture turns to single line and *forte-piano* shames *sforzando* in perceived intensity. But such acts, which on the surface only cause the structure or intensity to disappear, must make the largest difference underneath. Shortly, I will propose a scheme in which a sudden, searching change awakes a lot of mental *difference-finders*. This very change wakes yet more *difference-finders*, and this awakening wakes still more. That is how sudden silence makes the whole mind come alive.

We are "told" all this in just one minute of the lesson and I have touched but one dimension of its rhetoric. Besides explaining, teachers beg and threaten, calm and scare; use gesture, timbre, quaver, and sometimes even silence. This is vital in music, too. Indeed, in the *Fifth*, it is the start of the subject! Such "lessons" must teach us as much

---

about triads and triplets as mathematicians have learned about angles and sides! Think how much we can learn about minor second intervals from Beethoven's *Grosse Fuge in E-flat, Opus 133*.

### What Use Is Music?

Why on earth should anyone want to learn such things? Geometry is practical—for building pyramids, for instance—but of what use is musical knowledge? Here is one idea. Each child spends endless days in curious ways; we call this *play*. A child stacks and packs all kinds of blocks and boxes, lines them up, and knocks them down. What is that all about? Clearly, the child is learning about space! But how on earth does one learn about time? Can one time fit inside another? Can two of them go side by side? In music, we find out! It is often said that mathematicians are unusually involved in music, but that musicians are not involved in mathematics. Perhaps both mathematicians and musicians like to make simple things more complicated, but mathematics may be too constrained to satisfy that want entirely, while music can be rigorous or free. The way the mathematics game is played, most variations lie outside the rules, while music can insist on perfect canon or tolerate a casual accompaniment. So mathematicians might need music, but musicians might not need mathematics. A simpler theory is that since music engages us at earlier ages, some mathematicians are those missing mathematical musicians.

Most adults have some childlike fascination for making and arranging larger structures out of smaller ones. One kind of musical understanding involves building large mental structures out of smaller, musical parts. Perhaps the drive to build those mental music structures is the same one that makes us try to understand the world. (Or perhaps that drive is just an accidental mutant variant of it; evolution often copies needless extra stuff, and minds so new as ours must contain a lot of that.)

Sometimes, though, we use music as a trick to misdirect our understanding of the world. When thoughts are painful we have no way to make them stop. We can attempt to turn our minds to other matters, but doing this (some claim) just submerges

the bad thoughts. Perhaps the music that some call *background music* can tranquilize by turning under-thoughts from bad to neutral, leaving the surface thoughts free of affect by diverting the unconscious. The structures we assemble in that detached kind of listening might be wholly solipsistic webs of meaninglike cross-references that nowhere touch "reality." In such a self-constructed world, we would need no truth or falsehood, good or evil, pain or joy. Music, in this unpleasant view, would serve as a fine escape from tiresome thoughts.

### Syntactic Theories of Music

Contrast two answers to the question, Why do we like certain tunes?

Because they have certain structural features.  
Because they resemble other tunes we like.

The first answer has to do with the laws and rules that make tunes pleasant. In language, we know some laws for sentences; that is, we know the forms sentences must have to be syntactically acceptable, if not the things they must have to make them sensible or even pleasant to the ear. As to melody, it seems, we only know some features that can help—we know of no absolutely essential features. I do not expect much more to come of a search for a compact set of rules for musical phrases. (The point is not so much what we mean by *rule*, as how large a body of knowledge is involved.)

The second answer has to do with significance outside the tune itself, in the same way that asking, Which sentences are meaningful? takes us outside shared linguistic practice and forces us to look upon each person's private tangled webs of thought. Those private webs feed upon themselves, as in all spheres involving preference: we tend to like things that remind us of the other things we like. For example, some of us like music that resembles the songs, carols, rhymes, and hymns we liked in childhood. All this begs this question: If we like new tunes that are similar to those we already like, where does our liking for music start? I will come back to this later.



The term *resemble* begs a question also: What are the rules of musical resemblance? I am sure that this depends a lot on how melodies are "represented" in each individual mind. In each single mind, some different "mind parts" do this different ways: the same tune seems (at different times) to change its rhythm, mode, or harmony. Beyond that, individuals differ even more. Some listeners squirm to symmetries and shapes that others scarcely hear at all and some fine fugue subjects seem banal to those who sense only a single line. My guess is that our contrapuntal sensors harmonize each fading memory with others that might yet be played; perhaps Bach's mind could do this several ways at once. Even one such process might suffice to help an improviser plan what to try to play next. (To try is sufficient since improvisers, like stage magicians, know enough "vamps" or "ways out" to keep the music going when bold experiments fail.)

How is it possible to improvise or comprehend a complex contrapuntal piece? Simple statistical explanations cannot begin to describe such processes. Much better are the *generative* and *transformational* (e.g., neo-Schenkerian) methods of syntactic analysis, but only for the simplest analytic uses. At best, the very aim of syntax-oriented music theories is misdirected because they aspire to describe the sentences that minds produce without attempting to describe how the sentences are produced. Meaning is much more than sentence structure. We cannot expect to be able to describe the anatomy of the mind unless we understand its embryology. And so (as with most any other very complicated matter), science must start with surface systems of description. But this surface taxonomy, however elegant and comprehensive in itself, must yield in the end to a deeper, causal explanation. To understand how memory and process merge in "listening," we will have to learn to use much more "procedural" descriptions, such as programs that describe how processes proceed.

In science, we always first explain things in terms of what can be observed (earth, water, fire, air). Yet things that come from complicated processes do not necessarily show their natures on the surface. (The steady pressure of a gas conceals those countless, abrupt microimpacts.) To speak of what

such things might mean or represent, we have to speak of how they are made.

We cannot describe how the mind is made without having good ways to describe complicated processes. Before computers, no languages were good for that. Piaget tried algebra and Freud tried diagrams; other psychologists used Markov chains and matrices, but none came to much. Behaviorists, quite properly, had ceased to speak at all. Linguists flocked to formal syntax, and made progress for a time but reached a limit: transformational grammar shows the contents of the registers (so to speak), but has no way to describe what controls them. This makes it hard to say how surface speech relates to underlying designation and intent—a baby-and-bath-water situation. The reason I like ideas from AI research is that there we tend to seek procedural description first, which seems more appropriate for mental matters.

I do not see why so many theorists find this approach disturbing. It is true that the new power derived from this approach has a price: we can say more, with computational description, but prove less. Yet less is lost than many think, for mathematics never could prove much about such complicated things. Theorems often tell us complex truths about the simple things, but only rarely tell us simple truths about the complex ones. To believe otherwise is wishful thinking or "mathematics envy." Many musical problems that resist formal solutions may turn out to be tractable anyway, in future simulations that grow artificial musical semantic networks, perhaps by "raising" simulated infants in traditional musical cultures. It will be exciting when one of these infants first shows a hint of real "talent."

## Space and Tune

When we enter a room, we seem to see it all at once; we are not permitted this illusion when listening to a symphony. "Of course," one might declare, for hearing has to thread a serial path through time, while sight embraces a space all at once. Actually, it takes time to see new scenes, though we are not usually aware of this. That totally compel-

ling sense that we are conscious of seeing everything in the room instantly and immediately is certainly the strangest of our "optical" illusions.

Music, too, immerses us in seemingly stable worlds! How can this be, when there is so little of it present at each moment? I will try to explain this by (1) arguing that hearing music is like viewing scenery and (2) by asserting that when we hear good music our minds react in very much the same way they do when we see things.<sup>1</sup> And make no mistake: I meant to say "good" music! This little theory is not meant to work for any senseless bag of musical tricks, but only for those certain kinds of music that, in their cultural times and places, command attention and approval.

To see the problem in a slightly different way, consider cinema. Contrast a novice's clumsy patched and pasted reels of film with those that transport us to other worlds so artfully composed that our own worlds seem shoddy and malformed. What "hides the seams" to make great films so much less than the sum of their parts—so that we do not see them as mere sequences of scenes? What makes us feel that we are there and part of it when we are in fact immobile in our chairs, helpless to deflect an atom of the projected pattern's predetermined destiny? I will follow this idea a little further, then try to explain why good music is both more and less than sequences of notes.

Our eyes are always flashing sudden flicks of different pictures to our brains, yet none of that saccadic action leads to any sense of change or motion in the world; each thing reposes calmly in its "place"! What makes those objects stay so still while images jump and jerk so? What makes us such innate Copernicans? I will first propose how this illusion works in vision, then in music.

We will find the answer deep within the way the

1. Edward Fredkin suggested to me the theory that listening to music might exercise some innate map-making mechanism in the brain. When I mentioned the puzzle of music's repetitiousness, he compared it to the way rodents explore new places: first they go one way a little, then back to home. They do it again a few times, then go a little farther. They try small digressions, but frequently return to base. Both people and mice explore new territories that way, making mental maps lest they get lost. Music might portray this building process, or even exercise those very parts of the mind.

mind regards itself. When speaking of illusion, we assume that someone is being fooled. "I know those lines are straight," I say, "but they look bent to me." Who are the different I's and me's? We are all convinced that somewhere in each person struts a single, central self; atomic, indivisible. (And secretly we hope that it is also indestructible.)

I believe, instead, that inside each mind work many different agents. (The idea of societies of agents [Minsky 1977; 1980a; 1980b] originated in my work with Seymour Papert.) All we really need to know about agents is this: each agent knows what happens to some others, but little of what happens to the rest. It means little to say, "Eloise was unaware of X" unless we say more about which of her *mind-agents* were uninvolved with X. Thinking consists of making mind-agents work together; the very core of fruitful thought is breaking problems into different kinds of parts and then assigning the parts to the agents that handle them best. (Among our most important agents are those that manage these assignments, for they are the agents that embody what each person knows about what he or she knows. Without these agents we would be helpless, for we would not know what our knowing is for.)

In that division of labor we call *seeing*, I will suppose that a certain mind-agent called *feature-finder* sends messages (about features it finds on the retina) to another agent, *scene-analyzer*. Scene-analyzer draws conclusions from the messages it gets and sends its own, in turn, to other *mind-parts*. For instance, *feature-finder* finds and tells about some scraps of edge and texture; then *scene-analyzer* finds and tells that these might fit some bit of shape.

Perhaps those features come from glimpses of a certain real table leg. But knowing such a thing is not for agents at this level; *scene-analyzer* does not know of any such specific things. All it can do is broadcast something about shape to hosts of other agents who specialize in recognizing special things. (Since special things—like tables, words, or dogs—must be involved with memory and learning, there is at least one such agent for every kind of thing this mind has learned to recognize.) Thus, we can hope, this message reaches *table-maker*, an agent

---

specialized to recognize evidence that a table is in the field of view. After many such stages, descendants of such messages finally reach *space-builder*, an agent that tries to tell of real things in real space.

Now we can see one reason why perception seems so effortless: while messages from scene-analyzer to table-maker are based on evidence that feature-finder supplied, the messages themselves need not say what feature-finder itself did, or how it did it. Partly this is because it would take scene-analyzer too long to explain all that. In any case, the recipients could make no use of all that information since they are not engineers or psychologists, but just little specialized nerve nets.

Only in the past few centuries have painters learned enough technique and trickery to simulate reality. (Once so informed, they often now choose different goals.) Thus *space-builder*, like an ordinary person, knows nothing of how vision works, perspective, foveae, or blind spots. We only learn such things in school: millennia of introspection never led to their suspicion, nor did meditation, transcendental or mundane. The mind holds tightly to its secrets not from stinginess or shame, but simply because it does not know them.

Messages, in this scheme, go various ways. Each motion of the eye or head or body makes feature-finder start anew, and such motions are responses (by muscle-moving agents) to messages that scene-analyzer sends when it needs more details to resolve ambiguities. Scene-analyzer itself responds to messages from "higher up." For instance, *space-builder* may have asked, "Is that a table?" of table-maker, which replies (to itself), "Perhaps, but it should have another leg—there," so it asks scene-analyzer to verify this, and scene-analyzer gets the job done by making *eye-mover* look down and to the left. Nor is *scene-understander* autonomous: its questions to scene-analyzer are responses to requests from others. There need be no first cause in such a network.

When we look up, we are never afraid that the ground has disappeared, though it certainly has "disappeared." This is because *space-builder* remembers all the answers to its questions and never changes any of those answers without reason; mov-

ing our eyes or raising our heads provide no cause to exorcise that floor inside our current spatial model of the room. My paper on *frame-systems* (Minsky 1974) says more about these concepts. Here we only need these few details.

Now, back to our illusions. While feature-finder is not instantaneous, it is very, very fast and a highly parallel pattern matcher. Whatever scene-analyzer asks, feature-finder answers in an eye flick, a mere tenth of a second (or less if we have image buffers). More speed comes from the way in which *space-builder* can often tell itself, via its own high-speed model memory, about what has been seen before. I argue that all this speed is another root of our illusion: *if answers seem to come as soon as questions are asked, they will seem to have been there all along.*

The illusion is enhanced in yet another way by "expectation" or "default." Those agents know good ways to lie and bluff! Aroused by only partial evidence that a table is in view, table-maker supplies *space-builder* with fictitious details about some "typical table" while its servants find out more about the real one! Once so informed, *space-builder* can quickly move and plan ahead, taking some risks but ready to make corrections later. This only works, of course, when prototypes are good and are rightly activated—that is what intelligence is all about.

As for "awareness" of how all such things are done, there simply is not room for that. *Space-builder* is too remote and different to understand how feature-finder does its work of eye fixation. Each part of the mind is unaware of almost all that happens in the others. (That is why we need psychologists; we think we know what happens in our minds because those agents are so facile with "defaults," but we are almost always wrong.) True, each agent needs to know which of its servants can do what, but as to *how*, that information has no place or use inside those tiny minds inside our minds.

How do both music and vision build things in our minds? Eye motions show us real objects; phrases show us musical objects. We "learn" a room with bodily motions; large musical sections show us musical "places." Walks and climbs move



us from room to room; so do transitions between musical sections. Looking back in vision is like recapitulation in music; both give us time, at certain points, to reconfirm or change our conceptions of the whole.

Hearing a theme is like seeing a thing in a room, a section or movement is like a room, and a whole sonata is like an entire building. I do not mean to say that music builds the sorts of things that space-builder does. (That is too naive a comparison of sound and place.) I do mean to say that composers stimulate coherency by engaging the same sorts of interagent coordinations that vision uses to produce its illusion of a stable world using, of course, different agents. I think the same is true of talk or writing, the way these very paragraphs make sense—or sense of sense—if any.

## Composing and Conducting

In seeing, we can move our eyes; lookers can choose where they shall look, and when. In music we must listen *here*; that is, to the part being played now. It is simply no use asking *music-finder* to look *there* because it is not *then*, now.

If composer and conductor choose what part we hear, does not this ruin our analogy? When *music-analyzer* asks its questions, how can *music-finder* answer them unless, miraculously, the music happens to be playing what *music-finder* wants at just that very instant? If so, then how can music paint its scenes unless composers know exactly what the listeners will ask at every moment? How to ensure—when *music-analyzer* wants it now—that precisely that “something” will be playing now?

That is the secret of music; of writing it, playing, and conducting! Music need not, of course, confirm each listener's every expectation; each plot demands some novelty. Whatever the intent, control is required or novelty will turn to nonsense. If allowed to think too much themselves, the listeners will find unanswered questions in any score; about accidents of form and figure, voice and line, temperament and difference-tone.

Composers can have different goals: to calm and soothe, surprise and shock, tell tales, stage scenes,

teach new things, or tear down prior arts. For some such purposes composers must use the known forms and frames or else expect misunderstanding. Of course, when expectations are confirmed too often the style may seem dull; this is our concern in the next section. Yet, just as in language, one often best explains a new idea by using older ones, avoiding jargon or too much lexical innovation. If readers cannot understand the words themselves, the sentences may “be Greek to them.”

This is not a matter of a simple hierarchy, in which each meaning stands on lower-level ones, for example, word, phrase, sentence, paragraph, and chapter. Things never really work that way, and jabberwocky shows how sense comes through though many words are new. In every era some contemporary music changes basic elements yet exploits established larger forms, but innovations that violate too drastically the expectations of the culture cannot meet certain kinds of goals. Of course this will not apply to works whose goals include confusion and revolt, or when composers try to create things that hide or expurgate their own intentionality, but in these instances it may be hard to hold the audience.

Each musical artist must forecast and predirect the listener's fixations to draw attention *here* and distract it from *there*—to force the hearer (again, like a magician) to ask only the questions that the composition is about to answer. Only by establishing such preestablished harmony can music make it seem that something is there.

## Rhythm and Redundancy

A popular song has 100 measures, 1000 beats. What must the martians imagine we mean by those measures and beats, measures and beats! The words themselves reveal an awesome repetitiousness. Why isn't music boring?

Is hearing so like seeing that we need a hundred glances to build each musical image? Some repetitive musical textures might serve to remind us of things that persist through time like wind and stream. But many sounds occur only once: we must hear a pin drop now or seek and search for it; that is

why we have no "ear-lids." Poetry drops pins, or says each thing once or not at all. So does some music.

Then why do we tolerate music's relentless rhythmic pulse or other repetitive architectural features? There is no one answer, for we hear in different ways, on different scales. Some of those ways portray the spans of time directly, but others speak of *musical things*, in worlds where time folds over on itself. And there, I think, is where we use those beats and measures. Music's metric frames are transient templates used for momentary matching. Its rhythms are "synchronization pulses" used to match new phrases against old, the better to contrast them with differences and change. As differences and change are sensed, the rhythmic frames fade from our awareness. Their work is done and the messages of higher-level agents never speak of them; that is why metric music is not boring!

Good music germinates from tiny seeds. How cautiously we handle novelty, sandwiching the new between repeated sections of familiar stuff! The clearest kind of change is near-identity, in thought just as in vision. Slight shifts in view may best reveal an object's form or even show us whether it is there at all.

When we discussed sonatas, we saw how matching different metric frames helps us to sense the musical ingredients. Once frames are matched, we can see how altering a single note at one point will change a major third melodic skip at another point to smooth passing tones; or will make what was *there* a seventh chord into a dominant ninth. Matching lets our minds see different things, from different times, together. This fusion of those matching lines of tone from different measures (like television's separate lines and frames) lets us make those magic musical pictures in our minds.

How do our musical agents do this kind of work for us? We must have organized them into structures that are good at finding differences between frames. Here is a simplified four-level scheme that might work. Many such ideas are current in research on vision (Winston 1975).

*Feature-finders* listen for simple time-events, like notes, or peaks, or pulses.

*Measure-takers* notice certain patterns of time-events like 3/4, 4/4, 6/8.

*Difference-finders* observe that the figure *here* is same as that one *there*, except a perfect fifth above.

*Structure-builders* perceive that three phrases form an almost regular "sequence."

The idea of interconnecting *feature-finders*, *difference-finders*, and *structure-builders* is well exemplified in Winston's work (1975). *Measure-takers* would be kinds of *frames*, as described in "A Framework for Representing Knowledge" (Minsky 1974). First, the *feature-finders* search the sound stream for the simplest sorts of musical significance: entrances and envelopes, the tones themselves, the other little, local things. Then *measure-takers* look for metric patterns in those small events and put them into groups, thus finding beats and postulating rhythmic regularities. Then the *difference-finders* can begin to sense events of musical importance; imitations and inversions, syncopations and suspensions. Once these are found, the *structure-builders* can start work on a larger scale.

The entire four-level *agency* is just one layer of a larger system in which analogous structures are repeated on larger scales. At each scale, another level of order (with its own sorts of things and differences) makes larger-scale descriptions, and thus consumes another order of structural form. As a result, notes become figures, figures turn into phrases, and phrases turn into sequences; and notes become chords, and chords make up progressions, and so on and on. Relations at each level turn to things at the next level above and are thus more easily remembered and compared. This "time-warps" things together, changing tone into tonality, note into composition.

The more regular the rhythm, the easier the matching goes, and the fewer difference agents are excited further on. Thus once it is used for "lining up," the metric structure fades from our attention because it is represented as fixed and constant (like the floor of the room you are in) until some metric alteration makes the *measure-takers* change their minds. *Sic semper* all Alberti basses, um-pah-pahs,

and *ostinati*; they all become imperceptible except when changing. Rhythm has many other functions, to be sure, and agents for those other functions see things different ways. Agents used for dancing do attend to rhythm, while other forms of music demand less steady pulses.

We all experience a phenomenon we might call *persistence of rhythm*, in which our minds maintain the beat through episodes of ambiguity. I presume that this emerges from a basic feature of how agents are usually assembled; at every level, many agents of each kind compete (Minsky 1980b). Thus agents for 3/4, 4/4, and 6/8 compete to find best fits. Once in power, however, each agent "cross-inhibits" its competitors. Once 3/4 takes charge of things, 6/8 will find it hard to "get a hearing" even if the evidence on its side becomes slightly better.

When none of the agents has any solid evidence long enough, agents change at random or take turns. Thus anything gets interesting, in a way, if it is monotonous enough! We all know how, when a word or phrase is repeated often enough it, or we, begin to change as restless searchers start to amplify minutiae and interpret noise as structure. This happens at all levels because when things are regular at one level, the difference agents at the next will fail, to be replaced by other, fresh ones that then re-present the sameness different ways. (Thus meditation, undirected from the higher mental realms, fares well with the most banal of repetitious inputs from below.)

Regularities are hidden while expressive nuances are sensed and emphasized and passed along. Rubato or crescendo, ornament or passing tone, the alterations at each level become the objects for the next. The mystery is solved; the brain is so good at sensing differences that it forgets the things themselves; that is, whenever they are the same. As for liking music, that depends on what remains.

### Sentic Significance

Why do we like any tunes in the first place? Do we simply associate some tunes with pleasant experiences? Should we look back to the tones and patterns of mother's voice or heartbeat? Or could it be that some themes are innately likable? All these

theories could hold truth, and others too, for nothing need have a single cause inside the mind.

Theories about children need not apply to adults because (I suspect) human minds do so much self-revising that things can get detached from their origins. We might end up liking both *Art of Fugue* and *Musical Offering*, mainly because each work's subject illuminates the other, which gives each work a richer network of "significance." Dependent circularity need be no paradox here, for in thinking (unlike logic) two things can support each other in midair. To be sure, such autonomy is precarious; once detached from origins, might one not drift strangely awry? Indeed so, and many people seem quite mad to one another.

In his book *Sentics* (1978), Manfred Clynes, a physiologist and pianist, describes certain specific temporal sensory patterns and claims that each is associated with a certain common emotional state. For example, in his experiments, two particular patterns (that gently rise and fall) are said to suggest states of love and reverence; two others (more abrupt) signify anger and hate. He claims that these and other patterns—he calls them *sentic*—arouse the same effects through different senses—that is, embodied as acoustical intensity, or pitch, or tactile pressure, or even visual motion—and that this is cross-cultural. The time lengths of these sentic shapes, on the order of 1 sec, could correspond to parts of musical phrases.

Clynes studied the "muscular" details of instrumental performances with this in view, and concluded that music can engage emotions through these sentic signals. Of course, more experiments are needed to verify that such signals really have the reported effects. Nevertheless, I would expect to find something of the sort for quite a different reason: namely, to serve in the early social development of children. Sentic signals (if they exist) would be quite useful in helping infants to learn about themselves and others.

All learning theories require brains to somehow impose "values" implicit or explicit in the choice of what to learn to do. Most such theories say that certain special signals, called *reinforcers*, are involved in this. For certain goals it should suffice to use some simple, "primary" physiological stimuli like eating, drinking, relief of physical discomfort.



Human infants must learn social signals, too. The early learning theorists in this century assumed that certain social sounds (for instance, of approval) could become reinforcers by association with innate reinforcers, but evidence for this was never found. If parents could exploit some innate sentic cues, that mystery might be explained.

This might also touch another, deeper problem: that of how an infant forms an image of its own mind. Self-images are important for at least two reasons. First, external reinforcement can only be a part of human learning; the growing infant must eventually learn to learn from within to free itself from its parents. With Freud, I think that children must replace and augment the outside teacher with a self-constructed, inner, parent image. Second, we need a self-model simply to make realistic plans for solving ordinary problems. For example, we must know enough about our own dispositions to be able to assess which plans are feasible. Pure self-commitment does not work; we simply cannot carry out a plan that we will find too boring to complete or too vulnerable to other, competing interests. We need models of our own behavior. How could a baby be smart enough to build such a model?

Innate sentic detectors could help by teaching children about their own affective states. For if distinct signals arouse specific states, the child can associate those signals with those states. Just knowing that such states exist, that is, having symbols for them, is half the battle. If those signals are uniform enough, then from social discourse one can learn some rules about the behavior caused by those states. Thus a child might learn that conciliatory signals can change anger to affection. Given that sort of information, a simple learning machine should be able to construct a "finite-state person-model." This model would be crude at first, but to get started would be half of the job. Once the baby had a crude model of some *other*, it could be copied and adapted in work on the baby's self-model. (This is more normative and constructional than it is descriptive, as Freud hinted, for the self-model dictates more than portrays what it purports to portray.)

With regard to music, it seems possible that we conceal, in the innocent songs and settings of our

children's musical cultures, some lessons about successions of our own affective states. Senticly encrypted, those ballads could encode instructions about conciliation and affection, aggression and retreat; precisely the knowledge of signals and states that we need to get along with others. In later life, more complex music might illustrate more intricate kinds of compromise and conflict, ways to fit goals together to achieve more than one thing at a time. Finally, for grown-ups, our Burgesses and Kubricks fit Beethoven's *Ninths* to *Clockwork Oranges*.

If you find all this farfetched, so do I. But before rejecting it entirely, recall the question, Why do we have music, and let it occupy our lives with no apparent reason? When no idea seems right, the right one must seem wrong.

## Theme and Thing

What is the subject of Beethoven's *Fifth Symphony*? Is it just those first four notes? Does it include the twin, transposed companion too? What of the other variations, augmentations, and inversions? Do they all stem from a single prototype? In this case, yes.

Or do they? For later in the symphony the theme appears in triplet form to serve as countersubject of the scherzo: *three notes and one, three notes and one, three notes and one, still they make four* (Fig. 2). Melody turns into monotone rhythm; meter is converted to two equal beats. Downbeat now falls on an actual note, instead of a silence. With all of those changes, the themes are quite different and yet the same. Neither the form in the allegro nor the scherzo alone is the prototype; separate and equal, they span musical time.

Is there some more abstract idea that they both embody? This is like the problem raised by Wittgenstein (1953) of what words like *game* mean. In my paper on frames (Minsky 1974), I argue that for vision, *chair* can be described by no single prototype; it is better to use several prototypes connected in relational networks of similarities and differences. I doubt that even these would represent musical ideas well; there are better tools in contemporary AI research, such as constraint systems,

Fig. 2. Introductory measures of the third movement of Beethoven's Symphony No. 5 in C Minor.

The image displays a musical score for the introductory measures of the third movement of Beethoven's Symphony No. 5 in C Minor. The score is arranged in two systems. The first system includes staves for Flutes, Oboes, Clarinets in Bb, Bassoons, Horns in Eb, C, Trumpets in C, and Timpani in C, G. The second system includes staves for Violin I, Violin II, Viola, Cello, and Bass. The tempo markings are *Allegro (d. = ee)* and *poco ritard. a tempo*. The score features various musical notations, including dynamics such as *pp* and *f*, and articulation marks like accents and slurs. A measure number '13' is visible at the beginning of the second system.

---

conceptual dependency, frame-systems, and semantic networks. Those are the tools we use today to deal with such problems. (See *Computer Music Journal* 4[2] and 4[3], 1980.)

What is a good theme? Without that bad word *good*, I do not think the question is well formed because anything is a theme if everything is music!

So let us split that question into (1) What mental conditions or processes do pleasant tunes evoke? and (2) What do we mean by *pleasant*? Both questions are hard, but the first is only hard, to answer it will take much thought and experimentation, which is good. The second question is very different. Philosophers and scientists have struggled mightily to understand what pain and pleasure are. I especially like Dennett's (1978) explanation of why that has been so difficult. He argues that pain "works" in different ways at different times, and all those ways have too little in common for the usual definition. I agree, but if pain is no single thing, why do we talk and think as though it were and represent it with such spurious clarity? This is no accident: illusions of this sort have special uses. They play a role connected with a problem facing any society (inside or outside the mind) that learns from its experience. The problem is how to assign the credit and blame, for each accomplishment or failure of the society as a whole, among the myriad agents involved in everything that happens. To the extent that the agents' actions are decided locally, so also must these decisions to credit or blame be made locally.

How, for example, can a mother tell that her child has a need (or that one has been satisfied) before she has learned specific signs for each such need? That could be arranged if, by evolution, signals were combined from many different internal processes concerned with needs and were provided with a single, common, output—an infant's sentic signal of discomfort (or contentment). Such a genetically preestablished harmony would evoke a corresponding central state in the parent. We would feel this as something like the distress we feel when babies cry.

A signal for satisfaction is also needed. Suppose, among the many things a child does, there is one that mother likes, which she demonstrates by mak-

ing approving sounds. The child has just been walking *there*, and holding *this* just so, and thinking *that*, and speaking in some certain way. How can the mind of the child find out which behavior is good? The trouble is, each aspect of the child's behavior must result from little plans the child made before. We cannot reward an act. We can only reward the agency that selected that strategy, the agent who wisely activated the first agent, and so on. Alas for the generation of behaviorists who wastes its mental life by missing this plain and simple principle.

To reward all those agents and processes, we must propagate some message that they all can use to credit what they did, the plans they made, their strategies and computations. These various recipients have so little in common that such a message of approval, to work at all, must be extremely simple. Words like *good* are almost content-free messages that enable tutors, inside or outside a society, to tell the members that one or more of them has satisfied some need, and that tutor need not understand which members did what, or how, or even why.

Words like *satisfy* and *need* have many shifting meanings. Why, then, do we seem to understand them? Because they evoke that same illusion of substantiality that fools us into thinking it tautologous to ask, Why do we like pleasure? This serves a need: the levels of social discourse at which we use such clumsy words as *like*, or *good*, or *that was fun* must coarsely crush together many different meanings or we will never understand others (or ourselves) at all. Hence that precious, essential poverty of word and sign that makes them so hard to define. Thus the word *good* is no symbol that simply means or designates, as *table* does. Instead, it only names this protean injunction: Activate all those (unknown) processes that correlate and sift and sort, in learning, to see what changes (in myself) should now be made. The word *like* is just like *good*, except it is a name we use when we send such structure-building signals to ourselves.

Most of the "uses" of music mentioned in this article—learning about time, fitting things together, getting along with others, and suppressing one's troubles—are very "functional," but overlook



much larger scales of "use." Curt Roads remarked that, "Every world above bare survival is self-constructed; whole cultures are built around common things people come to appreciate." These appreciations, represented by aesthetic agents, play roles in more and more of our decisions: what we think is beautiful gets linked to what we think is important. Perhaps, Roads suggests, when groups of mind-agents cannot agree, they tend to cede decisions to those others more concerned with what, for better or for worse, we call aesthetic form and fitness. By having small effects at many little points, those cumulative preferences for taste and form can shape a world.

That is another reason why we say we like the music we like. Liking is the way certain mind-parts make the others learn the things they need to understand that music. Hence liking (and its relatives) is at the very heart of understanding what we hear. *Affect* and *aesthetic* do not lie in other academic worlds that music theories safely can ignore. Those other worlds are academic self-deceptions that we use to make each theorist's problem seem like someone else's.<sup>2</sup>

2. Many readers of a draft of this article complained about its narrow view of music. What about jazz, "modern" forms, songs with real words, monophonic chant and raga, gong and block, and all those other kinds of sounds? Several readers claimed to be less intellectual, to simply hear and feel and not build buildings in their minds. There simply is not space here to discuss all those things, but:

1. What makes those thinkers who think that music does not make them do so much construction so sure that they know their minds so surely? It is ingenuous to think you "just react" to anything a culture works a thousand years to develop. A mind that thinks it works so simply must have more in its unconscious than it has in its philosophy.
2. Our work here is with hearing music, not with hearing "music"! Anything that we can all agree is music will be fine—that is why I chose Beethoven's *Fifth Symphony*. For what is music? All things played on all instruments? Fiddlesticks. All structures made of sound? That has a hollow ring. The things I said of words like *theme* hold true for words like *music* too: it does not follow that because a word is public the ways it works on minds is also public. Before one embarks on a quest after the grail that holds the essence of all "music," one must see that there is as significant a problem in the meaning of that single sound itself.

## Acknowledgments

I am indebted to conversations and/or improvisations with Maryann Amacher, John Amuedo, Betty Dexter, Harlan Ellison, Edward Fredkin, Bernard Greenberg, Danny Hillis, Douglas Hofstadter, William Kornfeld, Andor Kovach, David Levitt, Tod Machover, Charlotte Minsky, Curt Roads, Gloria Rudisch, Frederic Rzewski, and Stephen Smoliar. This article is in memory of Irving Fine.

## References

- Clynes, M. 1978. *Sentics*. New York: Doubleday.
- Dennett, D. 1978. "Why a Machine Can't Feel Pain." In *Brainstorms: Philosophical Essays on Mind and Psychology*. Montpelier, Vermont: Bradford Books.
- Minsky, M. 1974. "A Framework for Representing Knowledge." AI Memo 306. Cambridge, Massachusetts: M.I.T. Artificial Intelligence Laboratory. Condensed version in P. Winston, ed. 1975. *The Psychology of Computer Vision*. New York: McGraw-Hill, pp. 211–277.
- Minsky, M. 1977. "Plain Talk About Neurodevelopmental Epistemology." In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*. Cambridge, Massachusetts: M.I.T. Artificial Intelligence Laboratory. Condensed in P. Winston and R. Brown, eds. 1979. *Artificial Intelligence*. Cambridge, Massachusetts: MIT Press, pp. 421–450.
- Minsky, M. 1980a. "Jokes and the Logic of the Cognitive Unconscious." AI Memo 603. Cambridge, Massachusetts: M.I.T. Artificial Intelligence Laboratory.
- Minsky, M. 1980b. "K-lines: A Theory of Memory." *Cognitive Science* 4(2): 117–133.
- Roads, C. ed. 1980. *Computer Music Journal* 4(2) and 4(3).
- Winston, P. H. 1975. "Learning Structural Descriptions by Examples." In P. Winston, ed. 1975. *Psychology of Computer Vision*. New York: McGraw-Hill, pp. 157–209.
- Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Oxford University Press.

# MIT Industrial Liaison Program

## Report

Paper relevant to the symposium:

"MEDIA TECHNOLOGIES"  
October 3, 1985

"Computers and Creativity"  
by  
Dr. Marvin Denicoff

1) copies of viewgraphs/slides by Marvin Denicoff

This paper has been duplicated at the request of the speaker.



Distributed for Internal Use  
by Member Companies Only.  
May Not be Reproduced.

© MIT

COMPUTERS  
AND  
CREATIVITY

Marvin Denicoff



## Theme

- **Computer Aiding of Human Creativity.**
- **The Computer as a Creative Organism in its Own Right**

# Context

- Performing Arts

- Theater

- Film

- Television

- Dance

- Music

## Goals

- Aiding Human Creativity
- Qualitative Improvement in Entertainment Products
- Option Development, Costing, and Evaluation
- Cost Reduction in Developing Scripts



√ Evolutionary, Multi-Disciplinary R & D Approach

√ Metrics



## Support Functions

- Human - Machine Interaction
- Smart Library Assistance
- Environment Porting

# Human - Machine Interaction

- Modes of Interactivity

- Type

- Print

- Cursive

- Speech

- Drawing

- Gestures

- Menus

- Color and Font



- √ Facile and Flexible

- √ Idiosyncratic Choice

- √ Optimized

# Smart Library Assistance

- Attributes

- Data Acquisition
- Digitization
- Multi-media Store
- Self Indexing
- Translation
- Fetching
- Bibliog. Prep.
- Abstracting
- Summarization
- Integrating
- Comparing
- Evaluating
- Anticipating
- Knowl. Updating
  
- Annotation - Personalization
  
- Reminding
  
- Alerting
  
- Surprising
  
- Learning & Adaptation - Reality



## Environment Porting

- Transporting - Recreating Ones Creative World

- Notes
- Pictures
- Scrapbooks
- Music
- Films
- Physical Spaces & Contents



√ Dynamic Updating

√ Availability

# Developing A Performing Arts Scenario

- Data Bases
- Collaboration
- Script Preparation
- Utilities

## Data Bases

- Stored Images, Acoustics, Text
  - Stage Configurations
  - Sets --Scenes
  - Characters -- People Types
  - Costumes
  - Voices -- Sounds
- √ Dynamic Acquisition of Data:  
Photos, Drawings, Motion Pictures,  
Text, Voices --Sounds
  - Via Library Search
  - Via Collaborator Inputs
- √ Retrieval Via Index Search and  
Pattern Matching



## Real Time Collaboration

- Participants

- Writers
- Directors
- Actors
- Set Designers
- Scene Designers
- Composers - Choreographers

- Aspects

- Direct Input

- √ Yes -- No

- √ Try this -- Did you consider?

- Data Bank Selection
    - Creation of New Images

- Decision Making

- Learning

- √ Across Collaborators

- √ By the Machine : People Typing and Calibration

# Script Preparation

- Collaborative Process
  - Choice of Participants
  - Instrumentation
  - Controls
- Aspects
  - Script Initiation
    - √ Library Aided
    - √ Inspiration
  - Creating an Electronic Facsimile of the Script
    - √ Characters
    - √ Voices
    - √ Sets
    - √ Costumes
    - √ Animation

# Utilities

- Options
  - Generation
  - Exploration
- Cost - Effectiveness
  - Economic Modeling
  - Trade off Assessment
- Dynamic Interaction
  - Collaborators
  - Critics
  - Audiences
- Product : Electronic Facsimile of Script
  - An Aid to Creativity
  - A Training Aid
  - A Stand Alone Entertainment

# The Creative Computer: Automatic Script Generation

- Script and Character Extension From Standard Plots
- Stylistic Modeling
- Wholly Independent Computer Generation of New Scripts



# MIT Industrial Liaison Program

## Report

Papers relevant to the symposium:

"MEDIA TECHNOLOGIES"  
October 3, 1985

"Television Past Broadcast"  
by  
Andrew B. Lippman  
Walter Bender

- 1) "Imaging and Interactivity," by Andrew B. Lippman
- 2) "Color Word Processing," by Andrew B. Lippman, Walter Bender,  
Gitta Solomon, Mitsuo Saito

These papers have been duplicated at the request of the speakers.



Distributed for Internal Use  
by Member Companies Only.  
May Not be Reproduced.

© MIT

Imaging and Interactivity

1-1

Andrew B. Lippman

Massachusetts Institute of Technology

## ABSTRACT

The technologies, techniques and applications of computer information systems are changing rapidly. At the Massachusetts Institute of Technology, the Media Technology program provides a new approach to research in this area that is a combination of technical disciplines with creative ones.

The Electronic Publishing Group in this laboratory is investigating techniques by which computing can be introduced into traditional information systems both to provide technical improvement in their electronic evolution and to analyze the information on behalf of the user. The goal is to evolve books, movies, and newspapers from one-way broadcast publications to intellectual partnerships, dialogues. This requires the combination of immersive, richly expressive interface technologies such as voice, gesture, autostereoscopic 3-D, and video with a systematic approach to interactive system design. Examples of current research at the Media Laboratory with application to office information systems, broadcast television, and personal photography will illustrate this convergence.

## 1. Introduction

In general, one may characterize the technological evolution of mass media imaging systems as a continuous improvement in either the fidelity or quality of rendition in support of established use. Film, for example, progressed from black and white Nickleodeon "flickers" through the introduction of editing (1906), sound (1924), Technicolor (1936), and Cinemascope (1938), but remains primarily a broadcast narrative medium. Although no one would deny that current movies are vastly different from those of 50 years ago, they are functionally unchanged:

They remain a medium by which a story can be presented identically to a large audience.

The evolution of print, although stretched over a far longer historical interval, shows many parallels, as does television. Both distribute information of a fundamentally visual character in identical form to a temporally and physically distributed audience, and thus operate in a broadcast mode. Stylistic and technical development, with the possible exception of the invention of the index and the table of contents, have made them each more attractive and more efficient but have not altered their fundamental purpose or mode of use. Today, however, the basis for technological development has changed by the introduction of the digital computer. As a digital

signal processor, it represents a continuation of historical trends. Examples include bandwidth compression, word processing, image processing, and signal restoration. As a programmable element in the channel it can be more. It can enter into the communicative process as an active agent, a mediator of content as well as quality. On behalf of the information consumer, it can transform a broadcast into a dialogue, a presentation into an exploration. The result can be movies that are different each time they are viewed, books that are literally written as they are read, and newspapers that combine editorial direction with personal interest to provide timeliness, novelty, and relevance. Thus the "digital revolution" can change our modes of thought as well the details of distribution.

The newly formed Media Laboratory at the Massachusetts Institute of Technology is building a research program where the technical and expressive aspects of communications systems are merged. Some of the ten groups that define the initial research domains are drawn from existing activities that were once distributed philosophically and physically throughout the institute; several have not existed before. A brief description of each is provided in the appendix. This report will concentrate on the work of the Electronic Publishing Group, previously known as the Architecture Machine.

A recurrent theme in this paper will be convergence. The case will be made repeatedly throughout that no single aspect of the research can proceed independently of the others, they reinforce each other and produce a whole that is greater than the sum of the parts. At the lowest level, this implies that both technical capability and systematic application must drive research in new imaging technologies. More broadly, a communications channel is defined by both its characteristic and its use. Perhaps most importantly, the people who do the work must be versed in both scientific and creative styles of thought: They must be drawn from fields of programming, engineering, and the arts.

In terms of display design, key features will be shown to be the (1) rapidity of update, or degree of intergration of the display system with the higher level processing; (2) the capability for color and subtle tone scale versus the spatial resolution; (3) the importance of merging analogue video images with locally generated graphics.

### 1.1. Media and Variability

As an example, consider the evolution of several image communications systems as shown in figure 1. They have been laid out with reference to two independent axes, loosely labeled "quality" and "variability". There is no metric associated with either, they are simply a relative ordering, normalized for each system. As noted before, the historical development of film is primarily a progression along the vertical axis in the direction of higher quality, fidelity, or perceptible "bandwidth".

The horizontal axis is perhaps more interesting and important to the discussion. It is an ordering of the alternatives available to the user of the system. A possible metric for this axis could be the number of primitive elements of a system readily accessible by the user of that system in a given time interval. In a book, for example, the primitive element is the page, and one may turn rapidly to any of the pages in a volume. Broadcast television has progressively included more channels, but there are fewer of them than there are pages in most books.

There is a threshold along the horizontal axis that is arbitrarily placed, labeled the "threshold of Interactivity". As will be shown, there is a clear distinction that may be made between systems where the user has some small measure of control and those that are interactive. The value of this distinction will become clear later.

The significant feature of the illustration is that one can correlate different types of development to different disciplines. In general, practitioners in the field make progress along the vertical direction, and the horizontal direction is the province of computing.

A case in point is high definition television. The primary impetus behind most work is the desire to distribute programs through channels that have more lines, wider aspect ratios, and better color fidelity. It is therefore a linear extension of current TV, with some potential application to cinema and photography. This is the province of most broadcasters and equipment suppliers. It is a form of image processing.

Alternatively, to the extent that teletext and videotext can be associated with television, they are the extensions to the medium that derive from the field of computing

and extend in the horizontal direction only. An unhappy result of this is that most demonstrated systems make such a sacrifice of quality that they are below the standards of even the most spartan commercial program productions. The computer adds programmability only, and contributes nothing to the quality.

The examples presented below can be thought of as progress at a 45 degree angle on this chart. They show a convergence of computer techniques with imaging techniques that ideally improve both the quality and the variability of media, making communications channels both faithful and useful carriers of information. They exploit the potential for the computer and the particular display technology used to both enhance the presentation and to interpret it.

### 1.2. Interactivity

Outside the domain of the mass presentation media, imaging technology is intimately tied to the notion of interactivity. In America, for example, the Association for Computing Machinery links them in the special interest group SIGGRAPH, graphics and interactive systems. Similarly, in medicine, mapping, the graphic arts, and information retrieval, display systems are always provided with interactive controls. It is therefore useful to consider a more formal definition of interactivity that will guide us in the design of new media. As will become evident through the discussion, interactivity imposes demands on display system design in terms of their quality, control, response, and form.

An interactive system is defined as one where each participant is engaged in a mutual discourse characterized by simultaneity, memory, and interruptability. In human terms, interactivity is epitomized by a conversation. When two people converse, each is actively processing information all the time. While one is listening, he is composing his next response, and what he says is based both on what he intends to say and on what he hears. It is full-duplex: the two participants could talk at the same time, in fact, and occasionally do. The ability to interrupt the speaker is a feature of a conversation; without it we would have a lecture.

Note that there is no need for an interaction to have a goal or defined task associated with it. Often a conversation is entered into purely for the sake of talking, and the benefit derives from the interaction itself. Likewise, when a reason for talking initiates the discourse, the twists and turns that arise along the way often lead the conversation in directions not anticipated at the outset and usually result in a quite different conclusion.

In terms of computing systems, we can assert that interactivity is a design goal when the system is to be used in an exploratory and investigative manner, when we approach the system with either a vague or non-existent purpose. Use of the system itself is the goal, and it must perform not as a tool or expedient but as a partner.

A corollary of this definition is the notion of granularity or "primitivity". Simply put, this means that there is an elemental unit from which the interaction is built and which serves as the interruption boundary. The size of this unit must be small enough so that the perceived interrupt is immediate.

Further, it may be true that the elemental unit is a function of the medium of interchange and is independent of content. For example, in human conversation, we may interrupt each other either on word or phrase boundaries regardless of topic. Rarely do we stop in the middle of a word, but equally rarely do we talk until we have finished a complete paragraph. In other media, the primitive element is less well defined and is a subject of research. In cinema it is clearly shorter than the scene but potentially longer than a frame, although there is some work directed at extending interactivity to elements within the frame itself.

It is useful to distinguish between systems that are simply interactive and those that are intelligent. Intelligent systems mimic or evidence what we normally associate with human intelligence: creativity, capability for abstraction, reasoning. Interactive systems need not be intelligent to satisfy the definition, although occasionally it would help. Were one to build a new type of book, for example, it might be reasonable to want that book to be an intellectual partner, deducing the needs of the reader and reacting accordingly. Similarly, a system that abstracts the morning newspaper might benefit from intelligence to both analyze the news and ascertain its potential interest to the reader.

However, this is distinct from interaction. The personally composed journal or newspaper must presuppose intent in order to work, and need not be reactive in their presentation to suit that purpose. They are not necessarily useful when the reason for reading the newspaper is simply to see what's new, or when the book is opened for the first time.

The point is that it is possible to separate system intelligence from interactivity and recognize that each has independent applications and design implications. Neither is a subset of the other, nor a prerequisite. To a great extent, intelligence is more a feature of program design, and interactivity is a characteristic of the interface.

## 2. Examples of The Visual Interface

The following examples are indicative of a style and approach. They are drawn from recently completed and ongoing work directed at incorporating high quality imaging technology into interactive systems and thereby explore new interfaces to information. To a certain extent, they are extrapolations of existing media, as in the case of the news information system "Newspeak", but as well, they are attempts to build a systematic solution to a human requirement. They are visual and computational extensions of human processes, a way for the computer to hold up its end of the conversation.

### 2.1. Newspeak: Interactivity and Exploration

Newspeak is an electronic newspaper that both reads the news for you and helps you explore a dynamic, multi-media data base. It provides parallel information channels such as voice and image, and allows one to peruse the day's events with no more initial commitment than one makes to the printed morning newspaper. The

act of reading becomes an interaction between the user and a personal computer that results in a personalized edition that is suggestive and responsive.

The system includes telephone access to a commercial data base service, NEXIS, as well as television broadcasts and verbal information systems. In operation, the remote data bases are periodically scanned with a set of inquiries composed from an initial profile and modified by continuing use. Television broadcasts that include closed captioning are monitored and recorded for later abstraction.

Accumulated information is processed locally by techniques similar to the search mechanisms existent in NEXIS and is formatted into the front page of a simulated newspaper. As with a printed edition, the front page is larger than one can take in at a glance consisting of headlines, illustrations, and text. It may be viewed on a 525, broadcast compatible graphics display at full scale, in which case only the headlines are clear and the relative area allocated to individual items is evident. The textual content of each individual story is too small to read.

As well, it may be scanned by "sliding" the large scale front page behind a display window. The rest of the newspaper "bleeds" off the edge of the screen, but all the text is readable.

The screen is touch sensitive, and gestures made at either the edges or over the text itself control the exploration. A touch at the edge moves the whole page and reveals a new section of the page; touches within a particular story reveal more about it.

The information is thus laid out in a spatial array that may be taken in all at once or explored section by section. The large view shows at a glance major events of the day and how much information is available on each; within each story lies the detail.

When individual stories are read, additional information that is "behind the front page" may be perused. Touching a single word or phrase highlights its occurrence in other stories, and thus connects one column with the next. A gesture down a column advances through it, and a gesture across paints illustrations and opens the paper to inside pages where additional information is stored.

During the course of use, pictures drawn from local optical storage of both maps and file photographs are inserted on the screen and annotated accordingly. The weather map, for example, is retrieved as an image but overlaid with changing meteorological symbols. Currently, television segments are shown on a separate screen, but they will be inserted as the processing equipment becomes available.

A video facsimile of pages is stored locally on writable optical storage and serves as an electronic "clipping file". This reveals both the content of past issues and is a history of reading style. One may thus put the paper down and return later and pick up at the same point, reviewing an earlier reading.

As a model of interactivity, several features of the display and the programming system are essential. Perhaps primary is the display quality and variability. Text is displayed using a soft fonts algorithm developed at the



Media Laboratory that provides both high density and retains the stylistic character of different type fonts. The text, therefore allows for emphasis and highlighting common in print as an ancillary information carrier, but often lacking in computer retrieval systems. As is the case with human conversation, both the content and the manner of speech is important; here the content and the format of the text are available degrees of freedom.

The dynamics allow the newspaper to present information at a rate sufficient for exploration. A window programming system allows sections of the screen to be updated independently of the rest, and the storage of addition codes along with the bit representation of the character itself allow for rapid highlighting and cross correlation.

Perhaps most important, the newspaper is a merger of already established information presentation modes and computer display. The system layout deliberately models that of a large format daily newspaper because the parallel access provided by the large page allows the user to focus on items of interest directly, without either composing a search strategy in advance or sitting through a sequential presentation. Similarly, associations between events that seems unrelated at first glance are revealed by highlight. One need not have a specific a priori question in mind when the newspaper is opened, and queries are not limited to categories established in advance. The act of reading itself helps determine interest, and the system both sustains that interest and encourages it. The display and the user enter into a dialogue about the news, different for each reader, although drawn from the same ultimate source and colored its editorial slant. It is therefore not a replacement for an edited journal, but it is an extension to it, a conversation with the editor directly.

One should note that the display involved satisfies a few simple but rare criteria: It can display a subset of a memory image that is larger than the screen and scroll over it with no apparent delay; it can portray both video images and computer generated text; it requires no higher resolution than broadcast television yet provides a seemingly high resolution information interface.

## 2.2. Words and Colors

Of all the uses of the computers in everyday life, perhaps the most successful is the word processor. Word processing has so made obsolete by the typewriter that it has become a printer with an optional keyboard. Similarly, word processing has replaced the handwritten note so effectively that some people cry out for the invention of picture post cards with perforated edges. Unfortunately, in the transition to soft copy and assisted editing, the page as an indicator of more than the printed word has been left behind.

Consider a handwritten document. Often it is passed from person to person for comment, overwritten in many colors, erased both completely and incompletely, bent, folded, and soiled. Through this evolution, the original text may be completely lost, yet the page itself is rich in circumstantial information: the color may indicate the author of a edit; the succession of inks provide the sequence, and marginal notes may help reconstruct incidental thoughts that came to mind through the

editing process.

Contrast this with electronic mail. A document circulated through several reviewers and returned to the author may be as distorted as messages passed by children playing the telephone game, with the alterations equally obscure. Similarly, a substitution made on a word processor leaves no reminder of the past, and tells nothing about the manner in which the document was created. Each letter has equal weight, whether it be the result of great calculation or a rough draft.

In general, word processing is a misnomer. Available are printing tools, text processors. Required are word partners: assistants that are suggestive, remember doubts and modifications, and intermediate between author and readers. On-line dictionaries and thesaurus's are a step in the right direction.

A project undertaken by Mitsuo Saito, a visiting affiliate from Toshiba, is integrating color into word processing specifically to provide carriers of meaning other than the text itself. The purpose of the project is not to provide the means to create colorful documents, although that can be done, but to enable an author to return to a partially written document and quite literally pick up where he left off. This is being done on personal computers specifically so that the results may be used throughout the Media Laboratory, and thereby evolve. There are two components to the work: a technical side and an exploratory one.

Technically, color display on PC's suffers from a chicken and egg problem: With no real use for color, they include little capability. Color displays are generally lower resolution than their monochrome counterparts and operate more slowly. A significant proportion of the effort is directed at manipulating the bits rapidly enough to make the editor at least as good as its black and white relative.

A starting feature set has been selected. The system can distinguish between deletions and editorial changes. Deletions are treated normally, but edits leave a tint behind to show both the position and size of the moved or removed text. Similarly, insertions are chromatically distinct from emendations, and text inserted from another part of the document appears in a third color. All colors fade with time, dimming as the text is recalled from storage. User controls also can remove any extraneous color.

Proposed features develop the notion of editorial partner by performing analysis of the text as it is written. Often used phrases and words will be highlighted as an indicator of style. Perhaps alternative will be suggested. Macros that enable automatic control of the editor will be tested through the implementation of a terminal coding scheme that builds an adaptive dictionary of keystrokes both to enable rapid transfer via telephone lines, and to build a library of common actions.

We would like to add a pressure sensitive keyboard, since the pressure may be either a useful control or an indicator. Interestingly enough, a single pressure sensor is sufficient. One per key is not necessary.

### 2.3. Three-Dimensional Display

Two programs are exploring different approaches to three dimensional display and different uses for it. One addresses the design and use of a three dimensional display that combines a varifocal mirror with a sequential liquid crystal color display to provide an autostereoscopic depth image. The second explores non-spatial uses of the third dimension through the construction of a "laminar display" that is an optical superimposition of a set 2-D CRT's into a composite image that appears as a set of closely overlaid sheets. The first is in operation; the second is proposed but not yet constructed.

The connection between these investigations and interactive media is somewhat loose, and intentionally so. They are presented for several reasons. The first is that they are an example of research into a display area that has been historically under-rated largely due to either an incomplete implementation or the presence of inconvenient impedimenta like glasses. The second is that 3-D has traditionally been thought of as a volumetric extension to 2-D, as an additional dimension of space, as linear a X and Y. The proposed laminar display is designed to avoid this use. Part of its intent is to demonstrate that the gold may not be in the use of the third dimension as one of space, but as one of description. Finally, it represents a capability of the interface potentially as important and significant as color, and therefore warrants research for its own sake, without any predetermined bias or immediately apparent application.

The varifocal mirror is one of the few display techniques by which a true three dimensional image can be constructed that remains fixed in space and can be explored by head motion as well as binocular vision. A random scan CRT is viewed by reflection in a mirror that vibrates at the frame repetition rate, and thereby sweeps the display plane through a volume. The image created needs no glasses and remains static as the viewer moves his head through the field of view. In the past it has suffered by the expense of the circuitry needed to generate the points, and by the limitation to black and white imagery. Fuchs has adapted a standard video framestore as a driver for this display through a set of programming techniques. The framestore essentially is a sequential point generator, where each location in memory records the X and Y beam position of the display, and the memory address determines the depth.

Our work has extended the display to color. A liquid crystal "color shutter" in front of the CRT alternates between red and green. On one excursion of the mirror the red image is painted on the screen, and during the reverse motion of the mirror, the green one is painted. Yellow results when the two images coincide. Using a standard, interlaced video framestore for the image generation, essentially, one field contains the red information and the other field the green. The mirror consists of a stretched reflective mylar sheet drawn over a standard bass drum and excited by a 14 inch loudspeaker in the rear of the drum. It oscillates at 30 hertz and the amplitude is such that the forward and backward motion of the mirror are identical and the alternate red and green images overlay exactly.

This is an initial step towards the development of a

high quality display such as might be used in a driving simulator. The hypothesis is that the third dimension is essential in cases where either the image is unfamiliar, and cannot therefore be interpreted as an instance of a familiar vision, or where the image must be used dynamically, with no delay for interpretation allowed. The first case is exemplified by new medical imaging techniques such as positron emission tomography where metabolic structures never before represented are recorded. With no existing 3-D model to which they may be related, a view that is sculptural and allows direct head motion exploration may be helpful. The second case is in the control of remote processes: A car, for example, may be driven via a three dimensional view through the windshield, but potentially not through a 2-D image from the same point.

The laminar display is proposed as a way to explore the use of the third dimension as an explanatory space rather than as a volume. As proposed, it consists of seven separate displays brought into optical convergence through a set of mirrors and lenses. The lenses create a real image of the CRT in or near the plane of a field lens that is also the viewing surface. When the lenses are arranged appropriately, the images may appear slightly separated in depth, with any one of them lying on the surface of the final lens. Alternative arrangements can align them to create extremely high definition displays, or wide aspect ratios.

There are several potential uses for the set of registered images. In cases where a simultaneous set of dynamic events must be correlated or monitored, they may arrayed in depth, and one may look at all of them at once. This may be an improvement over a two dimensional array where in order to see one, the viewer must look away from the other. Liken it to the case of watching a "photo finish" in a horse race. A glance along the finish line readily reveals the first finisher, but a view from above allows one to visually address only one contestant at a time.

### 2.4. Facemaker

As a final example, "Facemaker" functions midway between a computational tool and an interactive system. The system allows one to construct a likeness of person seen only briefly, and was designed as a photographic version of an "Identikit". In an Identikit, a cartoon-like likeness is constructed from an ensemble of sketched facial components. The user picks a facial outline and adds detail by selecting the closest mouth, nose, eyes, and so forth. The result is much like a caricature created by an artist from a verbal description of the component elements.

Facemaker synthesizes a photograph of the target without any user analysis. In operation, the user is shown a succession of faces, each registered at the eyes, and either selects or rejects each by making a binary decision of whether the current face "reminds" him of the target or fails to. When a sufficient number of positive responses has been accumulated, a composite print is made, and it bears a striking resemblance to the target image.

The facial composition is akin to noise cancellation. The features that cause each face to be selected reinforce each other in the final image, while the differences cancel.

A key feature of the final photograph is that the image is a carrier of ambiguity. Were it too soft, it would resemble no one in particular, and were it too sharp, it would imply certainty where none existed.

The prototype operated photographically: a positive selection initiated an exposure that became part of a multiply exposed composite. In the electronic version, the faces were drawn from an optical videodisc that contained 7000 facial images, all registered at the eyes, and were composited dynamically in a video display. The user could terminate the process when the composite was a close replica.

Further, the initial selection was limited visually by scanning a map of available facial types.

By the addition of these two features, a novel imaging idea was extended to become a helpful interactive system.

### 3. Conclusion

The intent of this work is to demonstrate the integral relationship between imaging systems and the programming that lies behind them. It is seen that features like update rate may occasionally outweigh resolution in importance, and programming versatility is a critical design criteria. The work presented is a set of examples, and they should be understood as a group showing the potential for a broad and multi-faceted approach to the design of interactive imaging systems. They exhibit the potential for processing to serve human creative ends and change the character of communications channels, as well as technically improve them.

### References

- Boorstin, D. *The Discoverers*
- Backer, D. and A. Lippman "Future Interactive Graphics: Personal Video", Architecture Machine Group, Massachusetts Institute of Technology.
- Backer, D. and S. Gano "Dynamically Alterable Videodisc Displays", Architecture Machine Group, Massachusetts Institute of Technology.
- Lippman, A. "Computational Videodisc", Architecture Machine Group, Massachusetts Institute of Technology.
- Lippman, A. "High Resolution Display on Standard Raster Scan Systems", Architecture Machine Group, Massachusetts Institute of Technology.
- Monaco, J. *How to Read a Film*, Oxford, 1981.
- Negroponte, N. "Soft Fonts", Architecture Machine Group, Massachusetts Institute of Technology.
- Negroponte, N. "Books Without Pages", Architecture Machine Group, Massachusetts Institute of Technology.
- Schmandt, C. "Soft Typography", Architecture Machine Group, Massachusetts Institute of Technology.

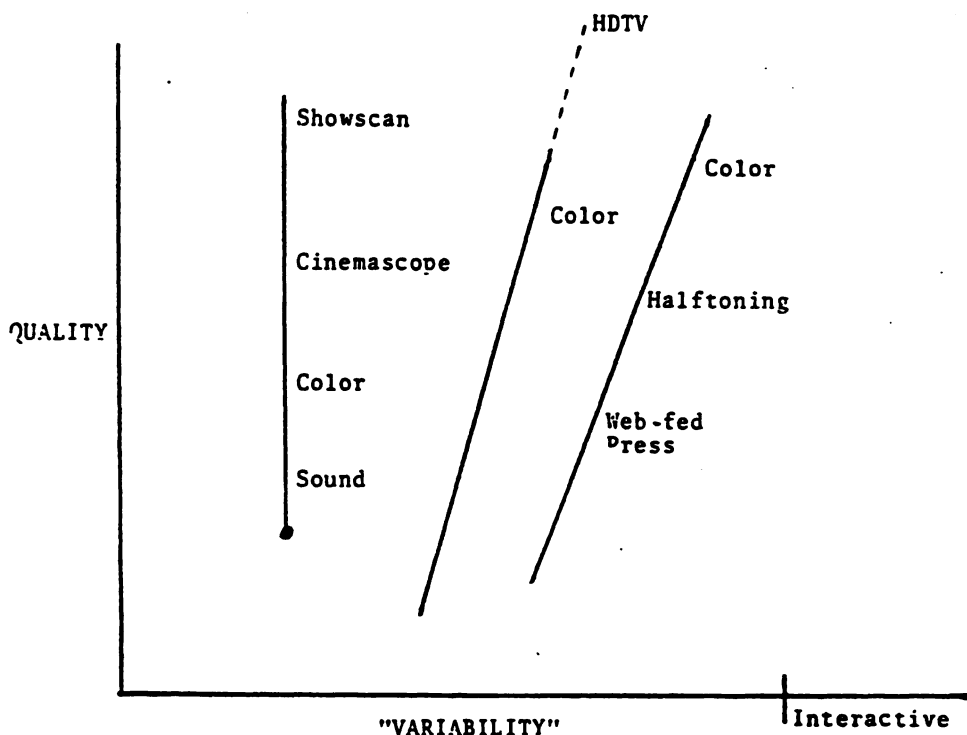


Figure 1: Evolution of media shown as a progression of quality versus variability

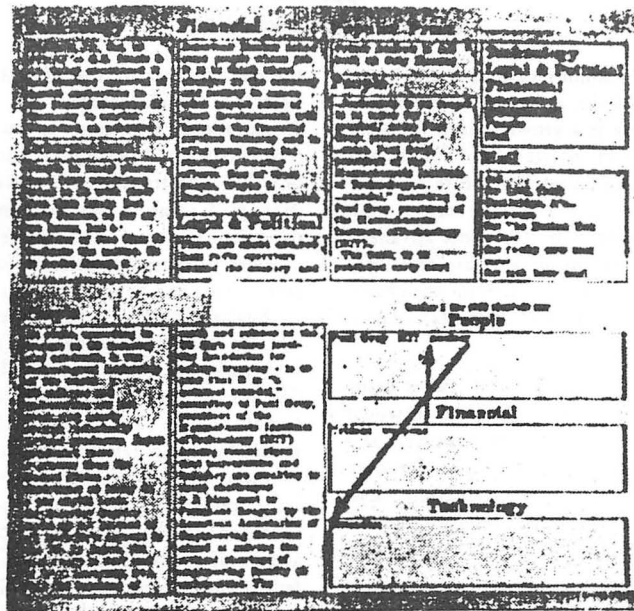


Figure 2:  
The front page of the Electronic Newspaper (Newspeak).

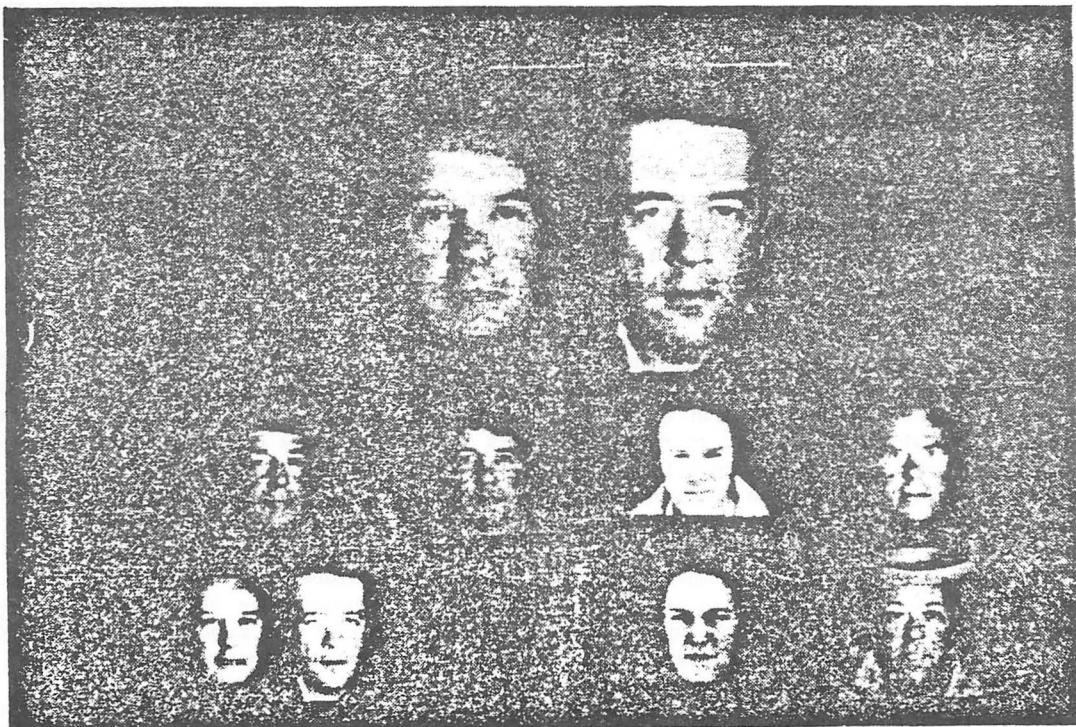


Figure 3:  
Facemacker. The top left picture is a composite of the six below selected to match the features of the top right.



Appendix  
Arts and Media Technology

As information sciences and communications technologies increasingly touch our daily lives, at home, in education, at work, and in leisure, there is a growing need for creative sensibilities at the human interface and for human usage. To this end, the Massachusetts Institute of Technology has committed to build a unique center that would couple artistic thought with the most advanced scientific thinking in order to invent and use creatively new media. The center would address directly the changing needs of education, publishing, and the entertainment industries.

A new facility designed by architect I. M. Pei is being built at the gateway of MIT's East Campus. It would house the Albert and Vera List Art Center as well as a newly established Media Laboratory. This new hundred thousand square foot facility includes production, viewing, and performance spaces specially designed to welcome tomorrow's technologies for sound and image, presentation and interaction.

The Media Laboratory itself is committed to ten groups: learning research, personal computation, electronic publishing, telecommunications, advanced television, spatial imaging, film video, graphics, and computer music. The confluence of these otherwise separate efforts is seen as a dramatic opportunity to bring together expressive, educational, and technological competences into one place; in large measure to create a new kind of person and a new style of thinking, equally fluent in the perceptual, cognitive, and creative processes as in the tradition of scientific inquiry and engineering excellence.

1. Learning research is deeply rooted in the human sciences, particularly epistemology. The computer and its programming are seen as much more than just another teaching technology; instead, as medium for learning exploration and control, with opportunities for at once playful and passionate interaction with increasingly intelligent and interactive computers. A major component of this work would be in third world experiments.

2. Personal computers are viewed as being in their nascent period, currently unwieldy, difficult to use, unpersonalized apparatus. The objective of research in personal computation would be focused at first on the human interface, its sensory richness and its recognition abilities. Speech production are exemplary and ongoing projects.

3. Electronic publishing addresses new delivery systems, combined with the most advanced technologies for storage and manipulation of information. Research would address the profound and fundamental change that turns hitherto monologues into potential conversations, breathing life and animation into a previously static page or frame.

4. Telecommunications has stepped to the forefront of confusion with government actions and technological achievements. Satellites, fiber optics, and deregulations have called into question all the basic tenets of personal access and bandwidth allocation. Direct broadcast satel-

lite and the creative use of nighttime telephone would be objects of initial research.

5. Advanced television is currently devoted to the technologies and perceptual consequences of high definition TV. The transmission and display characteristics are decoupled so as to use local intelligence and computing to enhance sound and image, making the television medium sufficiently enriched to serve simultaneously for true cinema and as computer terminal.

6. Spatial imaging is a research program launched to address synthetic holograms and holographic movies, previously out of reach in the absence of super-computers. The program includes other technologies for three-dimensional capture and display, for such applications as medical imaging, teleconferencing, and entertainment.

7. Film Video has been based in direct cinema, with early research leading to today's hand-held, sound-synchronous small systems. In combination with a regular agenda of production, the film video program is committed to advanced editing technologies, in the face of filmic media moving toward video and video itself moving toward digital image processing.

8. Graphics ranges from typography to computers, from paper to the cathode ray tube. The so-called Visible Language Workshop is devoted to the qualitative aspects of all presentational means, with their creative use illustrated by examples. One charter of this particular group is to bring visual sensibilities to electronic media, sensibilities which have been conspicuously missing to date.

9. Computer music is a tradition of composition and performance with new sounds, hypothetical instruments, and on-line scoring. In conjunction with artistic expression and control, the group has a long standing commitment to advanced signal processing. In addition, music is seen as an important epistemological venue for understanding fundamentals of appreciation itself.

10. Computers, Theatre, and Entertainment is directed at exploiting and extending state-of-the-art techniques to enhance creativity in the arts. A goal of this research is development of advanced methodologies for building electrical facsimilies of theatre productions complete with animated life-like characters, voice, costuming, scenery, and lighting. This work involves fundamental research areas in graphics, speech generation, man-machine interfaces, animation, and artificial intelligence. By marrying arts with modern instruction, and because the inherent facility of computation for providing on-line cooperation and high speed feedback, we anticipate the potential for permitting the exploitation of multiple options and achieve improvement gains in both the creative and cost effectiveness aspects of the performing arts.

The ten groups above would be merged into a single interdisciplinary laboratory with more than just hopes for intellectual co-mingling. The Media Laboratory is overly targeted at a longer term departmental status, creating a true academic pursuit in what we are calling: Media Arts and Sciences.

*This text editor uses display color not as a design element but as a way to convey the author's intent in every stage of a document's evaluation.*

## Color Word Processing

**O**f all the programs run on small computers, the text editing program must rank as one of the most popular. The sheer number of available alternatives attests to this, as does the fact that typewriters are evolving either into printers or into carriers of special-purpose word processing programs complete with dictionaries, spelling correctors, and displays. Soft copy has effectively replaced the written draft; and erasers, editing shears, and glue have become keyboard commands.

Without questioning the value of this transition, we can say that it is clear some things have been lost. Said another way, word processors may do their job too well. The printed version of a first draft, by lacking the smudges left by erasers and annotations and comments, has a finality that may belie its intent. Hard copy and stored text is no longer a working medium but output. Once made, ambiguities, corrections, and asides become indistinguishable components of the text itself; revisions made during suc-

**Table 1.**  
**Editor commands.**

Typographic correction	Editorial manipulation	Editor functions
Positioning commands	Word editing	Formatting
Character	*Delete	Fill
Word	Transpose	Center
Sentence	Replace	Indent
Paragraph	Case	File manipulation
Page	**Restore deletion	*Load
Buffer	Region editing	*Save
Character editing	*Delete	**Save monochromatic
Delete	*Move	Screen
Transpose	*Copy	Redisplay
*Insertion		**Monochromatic display

\*Commands modified to make implicit use of color  
 \*\*Commands added to Color EMACS

cessive readings are invisible. The document looks complete at every stage in its evolution.

To counteract this misleading feeling of completion, we present two variations of available editing programs that use a personal computer color display to preserve the history and train of thought associated with a written document. We distinguish between editor commands that are commonly used for typographical corrections and those that indicate revision. When sections of text are deleted or inserted, they either leave a background trace analogous to an incomplete erasure, or they appear in a new shade. Words left "behind" the apparent text may be recalled by an added editor command that invokes an "audit trail" similar to that used in newspaper production systems. As time progresses, the color variation diminishes, and the new text becomes an integral part of the document, indistinguishable from the original. Other changes are included as well, and we explain them later.

The purpose of these editors is to allow reconstruction of the "train of thought" that went into the creation of a document. We also use the editors to form the basis for a transition from the use of word processing as an efficient creation tool for printed output into a new way of writing and distributing the information itself.

We did not intend to create neon documents where emphasis is indicated by red letters, for example. Rather, we attempted to incorporate nonverbal cues into a word processor that would ultimately allow one to write half of a document then return to it and literally pick up where one left off. Parts that were troublesome to write appear different from those that went smoothly, and earlier thoughts are available for perusal and reinclusion. If the draft is circulated, changes made by each reader are readily apparent, and revisions can be distinguished from notes.

The evolution of text into a soft-copy reading environment is a more complicated and potentially longer range goal. If we assume that the text will be distributed electronically and read with the help of a system similar to that on which it was created, the degrees of freedom available to preserve the author's intent can be a guide for the reader as well. Parenthetical remarks, for example, can be flagged by a tint. Sections of the text where the reader can metaphorically ask for more detail can be highlighted.

In keeping with the notion that the additional display degrees of freedom should not require additional effort on the part of the writer nor should a new set of editing commands be learned, the editor introduces color automatically, as the editing keys are used. The intent of an edit is determined solely through use. Later we will describe a hypothetical enhancement to this process that allows the typing style itself to be recorded.

Three functions have been added to the editors as a result of this work. Two enable and disable the color display, and the third allows use of the audit trail feature. Essentially, text may be recalled by indicating the background tint and "yanking" what was once there from a stack that holds deletions.

As a history and editing mechanism, the editor should modify the color alterations so they will fade as time goes on; work on such a modification is under way. This change will also allow them to indicate similar phrases and constructions used throughout the text. We assume that a change becomes part of the original as time goes on, that more effort should be involved to reconstruct a very old draft versus a more recent one. Similarly, we expect to use directly controlled use of color to add degrees of freedom in revision. For example, when searching for a phrase, rather than simply moving the cursor to the first occur-



rence of the selected phrase, the editor could highlight all occurrences. As an automatic function, this feature is the analog to a spelling checker: misspellings and repetitions can be displayed.

Both editors are variations of EMACS, a full-screen text editor in widespread use. In one case the editor is used on a Sun terminal with a relatively high-performance, NTSC-compatible display system. The display range allows subtle alterations to the text that minimally distinguish it from normal display: The background tint, for example, can be as faint as that left on paper from a good pencil eraser. This display system is also fast enough to make the revised editor as rapid as its black-and-white counterpart. There is no speed penalty imposed by the color during normal text processing.

A similar editor was used on the IBM personal computer with the standard color display. Because of the different environment, we chose a different engineering solution; the display repertoire is comparatively limited. Nevertheless, this editor represents a potentially useful implementation, and it can improve as new display systems are added. Where differences between the two versions are significant, we will note them.

## Editing with EMACS

EMACS is a powerful screen editor that is really an instance of a set of macro commands for text manipulation.<sup>1</sup> While its full power need not concern us here, a few of the characteristics of usual implementations are important.

(1) The editor always operates in insert mode. Text is entered at the cursor location and displaces any text already there. Text is removed only by explicit command.

(2) The editing controls usually take the form of "control" functions. They are invoked by using either the control key or the escape prefix available on most computer keyboards. Usually, a particular command is a one-letter, mnemonic representation of the action; for example, control-b backs the cursor one character.

(3) The editor is designed to allow the simultaneous control of more than one "buffer" of text, and it is possible to write macro commands that operate on several buffers at the same time whether or not they are all on view. The IBM version exploits this to maintain display attributes in a buffer that is a mirror image of the actual text.

**Editing commands.** We divided a major subset of the normal editing commands into operational classes. One class includes operations that we assumed were used primarily for typographical corrections and thus would not alter the color of the displayed or stored text. Another class usually operates on more than one character at a time, and the operations are therefore assumed to be alterations to the content of the document. This second class was modified. The sample command set is shown in Table 1.

**Color EMACS on a Sun Microsystem.** Our first implementation of the color editor is on a Sun Microsystem model 1/100u 68000-based workstation. The workstation has a high-resolution monochromatic bit-mapped display, but this screen is not used by the color editor. An eight-bit Datacube VG-123 frame buffer on the workstation's Multibus is used as a color display terminal. An anti-aliased font package has been implemented for the frame buffer.<sup>2</sup> The proportionally spaced characters use four levels of gray scale and can be displayed at resolutions of up to 120 characters per line. The frame buffer has an 8-bit-in, 24-bit-out color look-up table, which provides subtle control over the hue, value, and chroma of both the foreground and background colors of the text. The color editor is a modification of a commercially available version of EMACS.<sup>3</sup>

There are two parts to this modification of EMACS, the terminal description and the redefinition of a subset of the editor commands. The terminal descriptor defined for the Datacube has provisions for color fonts. Screen updating incorporates color by detection of flags embedded in the text. Nonprintable characters are used to delimit regions of color text, a technique reminiscent of ANSI, a terminal protocol in which cursor positioning is controlled by sequences of embedded characters.

In the color editor, regions of color text are preceded by a character sequence, that is, an initial flag character followed by an attribute/revision field. Following the text is a terminal flag character. Cursor motion has been modified within the terminal description. The characters used to delimit regions of text are invisible to all of the cursor-positioning commands. The cursor motion also accounts for the proportionally spaced character sets. Horizontal screen position is calculated in picture elements rather than in column numbers.

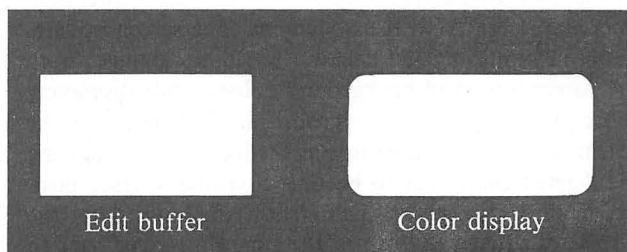


Figure 1. Text format. Typed text is encapsulated to distinguish it from the original document.

Most typed characters are taken by EMACS as text and inserted immediately into the edited document. In color EMACS, character insertion has been redefined (Figures 1 and 2). Typed text is encapsulated to make a distinction between inserted revisions and the original document. An attribute character is inserted to indicate that this is the start of newly inserted text.

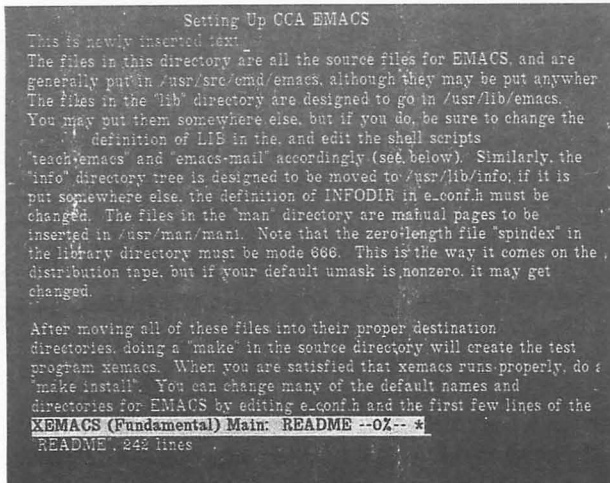


Figure 2. Color EMACS on the Sun. This figure depicts a page of a document in the color editor on the Datacube frame buffer, using antialiased fonts. The white text is from an earlier editing session. The yellow text is a new insertion.

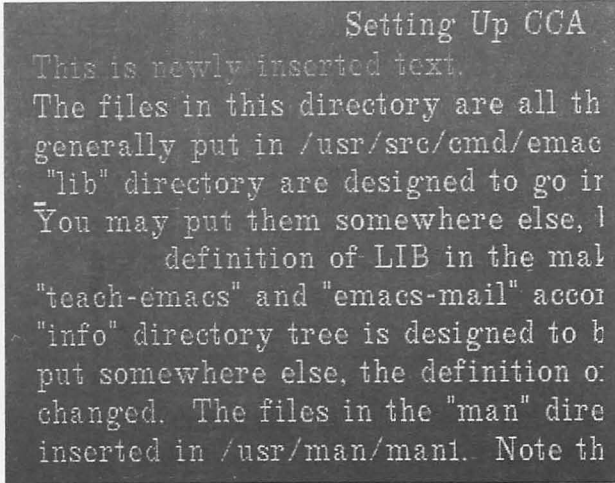


Figure 3. Enlargement of Color EMACS on the Sun. Old edits have decreased saturation.

The revision level is also inserted to maintain a history of revisions. New revisions are displayed with highly saturated colors. Revisions made in the last edit are displayed with less chroma, and the original text is displayed in monochrome (Figure 3). In this implementation only two revision levels are maintained. On exiting the editor, new revisions are made old, while the characters delimiting old revisions are removed. This feature makes those revisions indistinguishable from the original document.

In keeping with the spirit of an edit history, we indicate any deletions from the original by a change in the background color of the screen where the deletion has taken place (Figure 4). This is accomplished by surrounding words to be deleted with the deletion attribute, rather than removing them from the document. The screen update, upon encountering the deletion attribute, changes both the foreground and background colors while displaying subsequent text. The background is darkened and the foreground is made the same as the background, causing the deleted characters to become invisible. Cursor position is also affected by the deletion attribute. After the display of deleted characters, the cursor is moved back to the start of the deleted region. Any text following the deleted region is overlaid on the deleted text.

A consequence of our "hiding" deleted text is that it can be restored to the document simply by changing the encapsulating attribute character. We added a new command to EMACS, which, when invoked over previously deleted text, causes that text to be reinserted into the buffer. The deletion attribute is changed to the restored attribute. Restored text is treated like other revisions; it is displayed in a color that distinguishes it from the original

document. Similarly, text that has been moved from one place to another is displayed in color.

There are several obvious extensions that could be made to this implementation of the color editor. Many more than two revision levels could be maintained. By maintaining a separate revision level for each edit session, the editor is capable of retrieving any state of the document. By gradually decreasing the chroma of old revisions, one can discern at a glance what has been changed recently, while very old edits remain virtually indistinguishable from the original. Also, no provision currently exists to indicate the previous location of copied or moved text. By leaving behind a "where did we move this from" attribute, it would be possible to locate the source of each move. This information is necessary for maintenance of a complete trace of the edits.

There are constraints on the use and extension of this implementation of the color editor. The embedded character scheme becomes awkward as more and different attributes are added. It is especially cumbersome when regions of differing attributes overlap or when color is used on a character-by-character basis. In the former situation the definition of the terminal description becomes overly complex. In the latter, the quantity of attribute characters becomes overbearing. Another limitation of this methodology is imposed by the intertwining of attribute characters with the document being edited. Those characters must be removed before the document may be used beyond the context of the editor.

**Color MINCE on an IBM PC.** A second color editor is being developed on an IBM PC/AT. The PC is con-





phasis, and the second time we enter a sentence the revision is forcefully noted. Similarly "typomatic" functions are occasionally hurried (uneffectively) by harder pressure. We manipulate the keyboard the same way that we might have used a pen: using pressure to control the line weight, to emphasize, to eradicate totally what was there before.

Interestingly enough, it may be sufficient to use only a single pressure sensor on the keyboard, rather than one under each key. Because keys are struck one at a time, the coincidence of the particular keystroke with the pressure reading will reveal the force used on each individual key, with only slight perturbations from the rest of the hand.

Processing that reads the text as it is entered is a way of getting a global view of the document as it scrolls off the screen. For example, often-used constructions and phrases can be highlighted as they are written to reveal style to the author. Similarly, character transpositions and common spelling errors can be flagged during use.

This article directly addresses incorporation of color into word processing systems as a cognitive aid. Most of the specific enhancements are straightforward and automatic; they make the evolution of the document more obvious. We hope they are useful in and of themselves.

More generally, they may also be viewed as ways to integrate a computer into the general writing and reading tasks. Keystroke pressure is a way to measure intent as well as an additional control mechanism; it points the way for future development. The global goal is to make the machine a writing partner, helpful in all aspects of the written word.

Similarly, the converse of this problem is how documents can evolve when they exist as soft copy only. Again, color is a descriptive degree of freedom that can serve the reader as well as the writer. We hope that our work is a step in this direction. ■

## References

1. R. Stallman, "EMACS: The Extensible, Customizable, Self-documenting Display Editor," AI memo 519, MIT, Cambridge, Mass., Jn. 1979.
2. N. Negroponte, "Soft Fonts," *Proc. Society for Information Display*, Vol. 11, May 1980.
3. Z. Steven, *CCA EMACS User's Manual*, Computer Corporation of America, Inc., Cambridge, Mass.
4. MINCE Users Manual, Version 2.62, Mark of the Unicorn, Inc, Cambridge, Mass., 1981.



**Andrew Lippman** has spent the past 18 years at the Massachusetts Institute of Technology as both student and faculty member. Currently, he is associate professor of media technology and the holder of the NEC Career Development Professorship Chair of Computers and Communications. He also directs the electronic publishing group at the newly formed Media Laboratory. He has been the director of MIT's architecture

machine group.

Currently, his work addresses the future of television systems. The scope of this work ranges from technical research in high-definition television systems and image coding to evolutions of broadcast media that assume sophisticated personal computing and graphics generation hardware and software.

Lippman has degrees from MIT in electrical engineering and visual studies.



**Walter Bender** is a principal research scientist at MIT's Media Laboratory. He has been studying and conducting research at MIT since 1978. His work has included software development for the Aspen Movie-Map, numerous experiments in the area of display technology, and the development of News Peek, a news information system. His current work concerns the issues of information systems for the home.

Bender has a BA from Harvard University in visual and environmental studies and a MS from MIT in visual studies.



**Gitta Solomon** is a master's student and research assistant in the architecture machine group at MIT's Media Laboratory. Her interests include computer graphics and interactive computer systems. Solomon received her BA in mathematics at the University of California, Los Angeles.



**Mitsuo Saito** has been with the R&D Center of Toshiba Corporation since 1974, where he is mainly engaged in installing Japanese into the computer. Since March 1984 he has been researching the man-machine interface at MIT's architecture machine group.

The authors' address is MIT Media Laboratory, 20 Ames St., Bldg. E15, Cambridge, MA 02139.

September 30, 1985

CENTER FOR ADVANCED TELEVISION STUDIES CREATED BY US TELEVISION INDUSTRY

The Center for Advanced Television Studies (CATS) was created by ten US companies committed to expanding the development of television broadcasting technology and to promote and sponsor independent research. The member companies, all involved in some way with television broadcasting, are American Broadcasting Companies, Inc., CBS, Inc., Harris Corporation, 3M Company, National Broadcasting Company, Inc., Public Broadcasting Service, RCA Corporation, Tektronix, Inc., and Time, Inc. This new group will contract with independent academic institutions to conduct research on ways in which television systems in the US can be improved and made more effective. Research projects will focus on ways of increasing the efficiency of TV signal transmission and of enhancing picture and sound quality as perceived by the viewer. Results of all research projects will be published and shared among Center members and other interested US companies.

The first academic institution with which CATS has contracted is MIT. MIT has established a television research center called the Advanced Television Research Program (ATRP), designed to serve as a national laboratory for the study and improvement of television science and technology. It is headed by Dr. William F. Schreiber, professor of electrical engineering. For many years, his group has been conducting industrially sponsored research in television, facsimile, and graphic arts. The Laserphoto facsimile system, used by the Associated Press to transmit pictures to newspapers, was developed in this group. That was the first practical, low cost, laser facsimile system. ATRP has been awarded a three (3) year contract totalling 2.7 million dollars, and will use these funds to conduct research into television transmission and display.

CATS Chairman Charles Steinberg, Executive Vice President of Ampex, said the new Center fills a need for a centralized research facility where US companies can exchange ideas and stimulate independent research activities beyond those already being carried out in their own laboratories. "In many other countries," he said, "facilities for television technology research have been established by broadcasting organizations," typically with government support. But in the United States, little effort has been made to rethink and redesign the basic structure of our television transmission system since it was initially designed and approved by the Federal Communications Commission over thirty (30) years ago. The new Center will serve as a resource for such studies, to help the US television industry improve its position of world leadership in TV technology.

Prof. Schreiber said, "The present broadcast television system has been highly successful, both technically and economically. Evolutionary improvements in cameras, picture tubes, and circuitry have brought about better picture quality. However, current technological trends portend changes of a more revolutionary nature which existing systems may well not be able to accommodate. Some of these," he said, "are semiconductor technology, high-definition television, digital television, direct broadcasting from satellites (DBS), cable, fiberoptics, and video discs. A

significant array of such new products and possibilities make it high desirable to conduct research in order to understand the implications of the new technology and to lay the groundwork for future television developments." He listed these objectives for ATRP:

1. To develop the theoretical and empirical bases for the improvement of existing TV systems and the design of those of the future, and for the regulatory policies that will shape their use.
2. To motivate students to undertake careers in the TV industry.
3. To facilitate the continuing education of scientists and engineers already working in the industry, through work at MIT as visiting scientists or students.
4. To establish a resource center to which problems and proposals can be brought for discussion and detailed study.

"We plan to take a very fundamental view of the problem of improved TV systems," he said. "If we really want greatly improved pictures, we have to learn to deliver the information more efficiently. Otherwise, the channel capacity requirements become excessive and uneconomical. This involves more sophisticated signal processing, both at the transmitter and receiver, processing that is expected to become practical as prices of semiconductor components continue to fall. We hope that US industry will develop a leadership role in applying this new technology to TV equipment."

An important component of the research program at MIT will be audience research. The ATRP's initial effort under the funding provided by CATS will be to investigate both the perceptual and technological bases for improved TV systems.

Steinberg said that the center and its proposed activities have been favorably reviewed by the Department of Justice under the department's business review procedures.

For further information:

Charles Steinberg  
Ampex Corporation  
Redwood City, California  
415-367-4769

# MIT Industrial Liaison Program

## Report

Papers relevant to the symposium:

"MEDIA TECHNOLOGIES"  
October 3, 1985

"Advanced Television"  
by  
Professor William F. Schreiber

- 1) "Television's Search for Tomorrow," by William F. Schreiber
- 2) "Psychophysics and the Improvement of Television Image Quality,"  
by William F. Schreiber

This paper has been duplicated at the request of the speaker.



Distributed for Internal Use  
by Member Companies Only.  
May Not be Reproduced.

© MIT



This is a draft of a paper scheduled to be published in *Technology Review* for January 1986. Some minor changes will be made in the final version. Unfortunately, the computer-generated figures that illustrate many of the points made in the paper have not yet been completed.

**Television's Search for Tomorrow**  
by  
**William F. Schreiber**

William F. Schreiber is Professor of Electrical Engineering at MIT and Director of the Advanced Television Research Program. His major professional interest always has been image processing, both theoretical and applied. The opinions expressed herein are those of the author and not of MIT or the sponsors of the Advanced Television Research Program. In a number of cases, they are at variance with sponsors' stated positions.

Premature adoption of an international technical standard for television production could prejudice attempts to develop greatly improved television systems for the future.

There is a strong possibility that the television system that has had such a profound effect on our culture will, within the next five or ten years, change in a number of significant ways. If we so desire, we can have a system with larger, sharper pictures, equal to or better than 35mm theater film. We can have better rendition of motion, improved sound, and worldwide compatibility. If the quality were high enough, an entire new industry -- electronic still photography -- might be created.

Building and selling all the equipment needed for such new systems would give an important stimulus to the world economy. US industry, if it moved fast enough and proved smart enough, might get a good part of the market. However, this alluring future may be derailed by what some, including myself, believe to be the premature adoption of a proposal for a production standard for high-definition TV.

The proposal, which originated in Japan and is supported by some American organizations as well,

calls for programs intended for international exchange to be produced, although not transmitted, using the standards for a high-definition TV system developed under the leadership of Japan Broadcasting Corporation (NHK). The State Department has adopted the proposal as US policy, and supported it in October at the quadrennial standards-setting meeting of the International Radio Consultative Commission (CCIR). Opponents of the standard, particularly among European broadcasters, argue that it would cost a great deal to adopt as long as present broadcasting standards are in use, and that it fails to recognize current rapid progress in TV technology. By accepting the standards now, critics argue, the chances for future innovation are likely to be stifled.

It should be understood that I am talking about the "medium" and not the "message." Of course a TV system that simulated looking at the real world through a wide clear window, rather than through a gauze-covered keyhole, would permit, perhaps even stimulate, a different style of programming. However, it is clear that technology alone cannot correct any of the perceived shortcomings of today's TV programs. Thus to some, efforts to improve the technical quality of the image might seem beside the point. However, since TV is an economically important industry, and one in which the American share has fallen in recent years, the potential for a renaissance through better technology is quite appealing. In addition, working with images is aesthetically satisfying and intellectually challenging; as in many similar cases, developments in the field are likely to be driven to some extent by the engineers and scientists who are attracted to the field.

### Why Now?

Our present television system has proven remarkably durable, especially considering the radical changes in technology that have taken place since its adoption more than thirty years ago. However, a number of contemporary events and developments have combined to provide a setting in which change has become more likely.

*Alternative transmission channels.* Television is a voracious consumer of bandwidth, or channel capacity, one present-day TV channel using four times as much of this limited resource as the entire AM radio band. Many of the proposed improved systems require even more bandwidth, and so cannot be accommodated, with today's technology, in the existing 6 MHz channels. This additional capacity could be provided by cable, which has more than it needs, by direct broadcasting to the home from satellites (DBS), or by improved videocassettes and videodiscs. Cable is already here, DBS can be provided as soon as economically feasible, and magnetic recording is rapidly improving. Many fiberoptics cables have already been laid for common carrier communication systems, and could be run into every

household if the "wired city" ever became a reality.

*Semiconductor development.* The sophisticated TV systems of the future will require complex signal processing and a substantial amount of computer-like memory in the receiver. This will be practical only because of the revolution that has been sweeping the semiconductor industry in recent years. The cheaper and more powerful chips that set the stage for microcomputer development also provide the possibility of vastly increasing the "intelligence" of receivers, a necessity if improved image quality is not to require greatly increased bandwidth. For example, if we wanted to improve motion portrayal by doubling the number of separate pictures transmitted per second, that would double the bandwidth. On the other hand, a receiver with enough computer power might be able to generate these extra frames from those already transmitted, without raising the bandwidth consumption at all.

*Signal processing.* After forty years' effort, we have learned a great deal about processing the signals involved in imaging systems. For example, the ability of Caltech's Jet Propulsion Laboratory to produce spectacular improvement in the quality of photos from space is well known. Virtually every excellent color image we see in a magazine has been digitized and computer-processed for enhancement of quality. Comparable processing of the moving images of television and motion pictures in real time is more challenging, but research workers are making notable progress. The Bell Northern Research Laboratory in Montreal, for example, has demonstrated, by computer simulation, very high quality color images for video conferencing using only *one fiftieth* the normal channel capacity. Image sequences with only slightly restricted motion capability have been produced using only every eighth original frame. A number of laboratories around the world have produced significantly improved image quality from standard broadcast signals by means of signal processing in conjunction with special receivers.

A significant point is that all of this processing is digital. While the existing TV industry uses mostly analog techniques, a certain amount of digital processing has already found its way into TV studios. Many of the elaborate special effects, including all that require continuously variable rotation or change of size, are done digitally. Network coverage of the most recent Olympics and of the 1984 national elections made extensive use of so-called digital video effects (DVE) systems for this purpose. A number of companies have demonstrated digital video tape recorders, although they are not yet in commercial use. Even now, however, virtually every professional video tape recorder (VTR) includes a digital time base corrector that cancels out the effect of nonconstant head and tape speeds.

*Le defi Japonais.* No doubt, the most compelling impetus toward change was the demonstration, in 1981, of the Japanese high-definition television system (HDTV), after a development that took more

than ten years. It was orchestrated by NHK, which coordinated the work of many other laboratories, including Hitachi, Matsushita, and Sony. The results are roughly the equivalent of 35mm motion pictures and are quite striking to anyone accustomed to normal home TV. This significant development, which could not have happened in the US, (for one thing, its organization would probably have been illegal under our antitrust laws) includes high-definition cameras and picture tubes as well as an analog VTR, film scanner, film recorder, and associated circuitry. The cameras, picture tubes, and VTR represent a substantial advancement in the state of the art, and are therefore the subject of universal admiration.

The system tying all these new components together, however, is quite conventional; thus it requires four to five times the channel capacity of existing systems. This requirement, coupled with the noncompatibility of the system with existing receivers, is at the root of much of the controversy. A subsampled version, the ingenious MUSE system, requires a more sophisticated receiver but only about twice the present bandwidth, and it is also noncompatible. The image quality of MUSE has not been fully proved. While it is much higher than NTSC, it almost surely is lower than that of the original HDTV system.

#### How TV Works, and How it Sometimes Fails

*Scanning.* Production of a television signal involves a series of abstractions of the real, three-dimensional world in front of the camera. Light reflected from the scene forms an image in the focal plane of the camera. [SEE BOX] The point-by-point image intensity is sampled by an aperture (physically, an electron beam) that repeatedly scans the focal plane in a regular pattern of closely spaced parallel lines, called a raster. The video signal transmitted to the receiver is proportional to the image intensity sensed by the aperture. At the receiver, a scanning electron beam in the picture tube, synchronized with that at the camera (but delayed, if need be, by the transmission time) traverses the phosphor-coated receiving screen in the same raster pattern. The instantaneous beam intensity is made proportional to the light intensity on the focal plane of the camera, producing the output image by translating the energy of the beam into light.

The pattern of light emitted from the screen depends on the persistence of the phosphor. To avoid carryover of light from one frame to the next, which would blur moving objects, the persistence is always less than a frame time, and usually very much less -- typically less than one half. As a result, the instantaneous screen image actually comprises a small number of very intense scanning lines, moving down the screen. The perception of a continuously illuminated image depends entirely on integration in the eye,

in all present systems.

The image quality obviously depends on the number of complete pictures per second, the number of lines per picture, and the number of resolvable picture elements (pels or pixels) per line. Higher numbers give higher quality but naturally demand higher channel capacity. The required bandwidth of the transmission channel, in Hertz (cycles per second), is about one half the number of pels per second. Thus a 30x30 image, transmitted at ten frames per second, requires 9000 samples per second or 4500 Hz, which is roughly the bandwidth of a radio voice channel. In fact, the great Scottish TV pioneer, James Logie Baird, transmitted such pictures from ships at sea, in color and in 3-d, and at night with infrared illumination, in the 1920's. Of course, the quality of such low resolution images was rather poor.

As time went on, and the relationship of the scanning parameters to the channel capacity became known (Baird was innocent of such matters), image quality steadily improved. Britain used a 405-line, 25 frame per sec (fps) system for the coronation of King George VI in 1937. The French used an 819-line system for some time after World War II. All of Europe went to 625 lines, 25 fps in 1965(?) at the time color was added. An earlier experimental 441-line, 30-fps monochrome system was supplanted by the first American commercial system in 1941, using 525 lines, 30 fps. By comparison, the NHK HDTV system uses 1125 lines, although barely 800 are actually resolved on the screen. This inefficiency of utilization of vertical resolution is common to all existing systems, but can be mostly eliminated by the signal-processing techniques previously mentioned.

The number of pels per frame is a compromise between bandwidth and resolution, or sharpness. The 525-line NTSC compromise yields images with spatial resolution far too low for the large screens we take for granted in the theater. For comparison, office facsimile equipment normally uses about 1000 lines per page with equal horizontal and vertical resolution. Xerox copies are equivalent to about 2000 to 3000 lines, as are good quality printed color images. 35mm motion picture film is equivalent to 1000 to 1500 lines, and amateur slides to 2000 to 3000 lines.

Another difficulty with upgrading TV to theater quality is the aspect ratio -- the ratio of width to height of the screen. Current TV systems have a 4:3 aspect ratio because that was standard for film in 1936 when TV parameters were first chosen. Most films are now made at 1.85:1 or 2.35:1 and it is likely that any new TV system would have a wider picture, such as NHK's 5:3. There is no question but that a large screen and a very wide field of view add greatly to the sense of realism. In fact, systems such as Cinerama and Todd-AO (*Around the World in 80 Days*) achieved a good sense of three-



dimensionality using neither stereo nor polarizing glasses.

The number of frames per second is a compromise between bandwidth on the one hand and motion rendition and flicker on the other. The 25- and 30-fps systems use that frame rate, in the countries that have 50- and 60-Hz power systems respectively, because there once was an advantage in having it nearly identical to a (sub)multiple of the power frequency. Motion pictures use 24 fps. (In about 1928, the rate was raised from the 16 fps of silent movies for the sake of the *sound* quality!)

Since the eye's ability to integrate the wildly fluctuating light from the screen is limited, these rates, if used in TV and motion pictures, all would cause unacceptable flicker. Therefore the display rate must be raised. In film, each frame is repeated two or, more commonly, three times. Repetition is not possible in simple TV systems, so another technique -- interlace -- is used. In all current TV systems, alternate lines of the raster are transmitted on alternate scans of the screen. Thus, in the US system, 60 images (called fields) are displayed each second, each frame having 262 1/2 lines. This gives a large-area flicker rate of 60 per second, almost eliminating the flicker, provided the screen is neither too bright nor too wide. Viewers see more flicker in the 25-fps countries, but generally ignore it. The small-area, line-to-line flicker rate is still 30 per second, however, which is intolerable if the lines can be clearly resolved, as when viewing from close up. For this reason, interlace is usually not used in video display terminals. Flicker is more perceptible at the periphery of the visual field, so that it is often evident at the edges of wide-screen movies, even at 72 flashes per second, when we have our gaze fixed at the center. (It generally disappears if we look directly at the edge of the screen.)

Owing to certain details of the operation of interlace in TV cameras, the picture rate that relates to the rendition of motion is actually the field rate and not the frame rate. Thus, existing TV systems have adequate, although not perfect, motion rendition, rapidly moving objects tending to become blurred, the effective exposure time being 1/60 sec. ABC's "Super Motion" system, developed by Sony, effectively introduced a 1/180 sec shutter into the camera in a partially successful attempt to make the temporal resolution even higher. Flicker at the periphery, which we noted is present in wide-screen film even at 72 flashes per sec., is by itself a barrier to really large TV pictures at 60 fields per sec. In fact, movies portray motion far less smoothly and accurately than TV. Surprisingly, most viewers seem to ignore this problem. Possibly they unconsciously accept it as part of the "film look."

Interlace has problems as well as benefits. The effective vertical resolution is reduced from 525 lines to something less -- no more than 75% and under some conditions only 50%. This is partly because of the physics of camera operation and partly for psychophysical reasons. In any event, if the camera and

display had the full 525 lines of resolution called for by the standards, the interline flicker would be much worse. In addition, vertical motion of the camera produces certain artifacts, including the disappearance of half the lines, a phenomenon that is easy to observe by moving one's finger slowly down the face of the tube. Recently developed noninterlaced systems (called progressive scanning and requiring twice the bandwidth) show a remarkable improvement in "quietness" or "stability" when displayed side-by-side with standard systems. Although we have become accustomed to the shimmery nature of interlaced TV, we can easily see and appreciate the difference.

*Color.* [SEE BOX] Color television requires the transmission of three separate images, made through three different color filters, thus representing three different chromatic aspects of the original scene. This does not, however, require three times the channel capacity, as the resolution of the eye depends markedly on color. For example, in a red/green/blue (RGB) system, the green image might be transmitted at full resolution, the red at half resolution (one quarter on an area basis) and the blue at one fourth resolution (1/16 on an area basis.) Thus the color image would require only about a 30% increase in bandwidth over monochrome. Still better results would be obtained by transforming from RGB form into a different set of three components -- luminance (the intensity aspect of the stimulus, independent of color) and two chrominance signals. Both chrominance components could be reduced in resolution significantly, so that the incremental bandwidth required for color would be quite small. [USE FIGURE HERE]

Our present NTSC system, named after an industry group that proposed it, the National Television System Committee, was established as the US standard in 1953 primarily to achieve compatibility with the then-existing monochrome system. In this very clever system, luminance and chrominance information is transmitted as a single composite signal rather than as three separate components as in the earlier noncompatible field-sequential system. A band-sharing technique is used in which the two chrominance signals are mixed with the luminance signal, requiring them to be separated in the receiver and then converted back into RGB form for display on the picture tube. The luminance signal is much like a standard monochrome signal and can be received on unmodified monochrome receivers. Color receivers can also receive monochrome broadcasts properly.

The NTSC system and its European opposite numbers, PAL and SECAM, have spawned large and profitable industries. However, they are far from perfect. The horizontal color resolution is much lower than the vertical, resulting in horizontal smears of small, brightly colored areas. The luminance and chrominance signals interfere with each other. The color information not completely separated from luminance produces crawling serrations along the edges of large, brightly colored areas. Luminance

components with the same frequency as color components can (and do) appear as false color patterns, especially in areas of high detail content. As a practical matter, the introduction of the color subcarrier effectively reduced the luminance bandwidth, and thus the horizontal sharpness of the final image.

Most special effects require operations on signals in component form, necessitating repeated conversion between that form and the NTSC composite form. This exacerbates the problems due to chrominance/luminance band-sharing with consequent loss of quality. As TV equipment has gotten better, these defects have become more noticeable, leading recently to the introduction of component equipment, including cameras, VTR's and special effects generators, for studio use. Channel 7 in Boston, for example, does its electronic news-gathering production work in component format, translating to the NTSC composite format just once, after playback from a component-style VTR.

### Compatible Improvements to the NTSC System

*At the receiver.* In view of the ample evidence that the receiver of the present system does not achieve the image quality to which its bandwidth consumption entitles it, engineers have expended some effort to make improvements without changing the nature of the transmitted signal. More sophisticated separation of the luminance and chrominance signals using special filters eliminates much of the problem, at least for stationary subjects, and regains some of the lost luminance resolution. The more interesting improvements come from the use of a frame store, a semiconductor memory component that holds one entire TV image. This permits the interpolation of additional scan lines between those transmitted. The simplest such schemes double the number of scan lines per sec, displaying 525 progressively scanned lines in each field of 1/60 sec. This reduces visibility of scan lines and virtually eliminates interline flicker from the displayed image. For stationary images, the interpolation is best done temporally, i.e., by using corresponding lines in successive frames, so as to preserve the vertical resolution, but this blurs moving objects. Some adaptive systems interpolate temporally only in stationary areas. In moving areas, they interpolate vertically, i.e., by using adjacent lines in the same field. [DIAGRAM]

The divorce of scanning standards of picture tubes from those of the camera and channel can be carried further. It is possible to go up to 1050 lines, either 30 fps interlaced or even 60 fps progressively scanned. This greatly reduces the visibility of the line structure and makes possible the use of quality enhancement techniques mentioned previously. Generally, these schemes average out the noise in areas that lack detail, where noise is most visible, and sharpen the image in detailed areas, where the inevitable increase in noise is not easily perceived. Many computer graphics displays are now operating at about these rates, producing images that seem to be printed on paper and pasted on the front of the

picture tube. Such systems, of course, raise the cost of the receiver, but by less than one might expect. The frame memory is the key component. It can be made from 7 or 8 special, high-speed 256K memory chips, at present less than \$100. [CHECK CURRENT PRICES] Predictions of a \$10 frame memory within a few years are common. It is also necessary to improve some of the analog circuitry, but this must be done for any high-definition receiver and will not be very costly in mass production.

*Improvements at the transmitter.* Just as separating the receiver's scanning parameters from those of the transmitted signal has solved some problems, another group of improvements depends on separating the camera's parameters from those of the channel. By using progressive scan in the camera at 60 fps and with at least 525, but preferably 1050 lines, it would be possible to derive a standard signal for transmission that is free from most of the present defects associated with interlace. Technically, these defects are called *aliasing*, or the masquerading of one spectral component for another, representing a different spatial pattern from that actually scanned. Aliasing can be reduced by filtering, which requires more image samples than are to be transmitted. Noise can also be reduced and sharpness enhanced by operating on the high-rate signal, but a limit is placed on the vertical sharpness because it might produce even worse interlace problems on receivers not equipped with frame memories.

In another class of improved systems, exemplified by CBS's DBS proposal, now shelved for economic reasons, the transmission consists of two signals: one can be readily received on existing receivers; the other is an enhancement signal that is added to the first signal to produce high-definition pictures on special receivers. In the CBS system, the camera operates at 1050 lines. Filtering followed by subsampling produces a 525-line picture for the compatible channel. A signal that is essentially the difference between the original and compatible signal is sent on the second channel. In this case, extra information for the sides of the images is also transmitted, so that the final image has a wider aspect ratio. A relatively inexpensive converter is needed in any case to deal with the C-band (?) satellite signal, and this unit would also effect the conversion to NTSC format. This system produces images of quality comparable to those of the NHK system, but has the added feature of near compatibility.

### Beyond NTSC

Further enhancement of quality requires either adding more bandwidth or restructuring the video signal in some noncompatible manner. Several proposals for this kind of improvement involve separate transmission of color components, rather than using the NTSC composite band-sharing technique; they are collectively called multiplexed analog component (MAC) systems. Luminance/chrominance interference is eliminated by transmitting the components in separate frequency bands, or, preferably, in

sequence, time-compressed, on each scan line. Compression and decompression are done digitally at both transmitter and receiver, but require only a few line memories, and not full frame memories. (Of course, frame memories can also be used for additional benefit, if desired, just as in the improved NTSC systems mentioned.) [DIAGRAM] The image quality of typical MAC systems roughly equals that provided by the CCIR digital studio standard, which is very much better than normally achieved by PAL. The "MAC packet" systems espoused by the European Broadcasters' Union (EBU), also make provision for the transmission of multichannel digital audio plus a good deal of auxiliary data. It is anticipated that such signals would be transmitted by DBS. PAL-compatible signals, though of lower quality, could be derived for conventional earth-bound retransmission.

### **Much Better Systems for the Future**

To have much better image quality -- 35mm theater quality or better -- with little or no bandwidth expansion beyond NTSC, we must use the channel capacity more efficiently. Such radically improved systems must, of course, also use much better cameras and displays, and may use some of the techniques already demonstrated for eliminating some of the defects of interlace. Any new system almost certainly would abandon the band-sharing scheme for chrominance transmission.

*Better cameras and displays.* It costs nothing in channel capacity to use improved cameras and displays. In interlaced systems, very high vertical resolution is not useful in these components since it worsens interline flicker. This struck me during a visit to Sony in 1981, where I first saw high-definition picture tubes designed for the NHK-1125 line system. In my innocence, I suggested that NTSC pictures must look really good on such tubes, but was told that, in fact, they looked terrible! In some cases, camera tubes would give higher definition merely if they were operated at a higher line number. In other cases, adaptive enhancement schemes of the kind already used in graphic arts could give better signals. Finally, progress in solid-state cameras will almost certainly eventually provide all the resolution desired. Some invention may be needed to improve or supplant picture tubes, as it is becoming more and more difficult to improve the resolution of the current designs further. In addition, the larger screens that are desirable will have to be made flatter to fit into today's living rooms. Work on small projection systems and on solid-state displays may solve the problem.

*Separation of camera and display parameters from the channel.* In addition to improving their basic resolution, the most important change for the camera and display is to operate them at substantially higher line and frame rates than the channel. The advantages of this approach are the elimination of visible line structure and of its temporal counterpart, flicker, the facilitation of optimum enhancement



(by filtering) before converting to the channel signal, and the simplification of postproduction manipulation, including scan conversion. If carried far enough, these elements would become "transparent," so that the quality of the final image would depend only on the information sent through the channel. Optimum sampling patterns for deriving the channel signal from the camera output must be considered as well as the filtering (interpolation) used to derive the signal for the display from the transmitted information.

*Motion compensation.* It appears possible to get better motion rendition with fewer transmitted frames per second. If the necessary receiver processing can be made practical, this is likely to be the most important avenue to the goal of better quality with little or no bandwidth expansion. The possibility depends on the fact that successive frames of a TV sequence *must* be very similar to give a satisfactory impression of continuous motion. [SEE BOX] If the frame-to-frame motion of each point in the image is known, then intermediate images can be calculated quite accurately from those frames that are transmitted. This is already done in computer graphics and in computer-generated cartoons, where it is called "in-betweening." In these cases, the motion is known *a priori*. In natural scenes, the motion must be calculated from the video information. Noise presents an obstacle to doing this with perfect accuracy. In addition, the reconstruction of the signal for the display from the sparse samples that would be sent through the channel is a formidable computational task that must be performed in each receiver. Difficult problems must be solved to make this technique practical, but the rewards for doing so are correspondingly large.

\*\*\*\*\*

From these considerations, it is possible to discern the shape of a possible TV system of the future. The efficiency of bandwidth utilization, i.e., the image quality obtained in relationship to the channel capacity used, will be significantly improved compared with present systems. It will use excellent cameras and displays operated at very high line and frame rates -- well over 1000 lines and perhaps as high as 100 fps -- and probably progressively scanned. The camera signal will be appropriately filtered and sampled, probably in an offset pattern, giving higher spatial and lower temporal resolution than at present for transmission through the channel. These samples will be adaptively interpolated, using motion information either derived from the samples or transmitted directly (or both), up to a very high line and frame rate for display. There are additional possibilities for achieving greater efficiency including differential coding (only applicable to digital transmission) and multichannel transmission, in which the signal is divided into several channels of different types of spatial and temporal detail, each channel specially tailored to take advantage of specific perceptual characteristics. These latter methods require more

research to prove their validity.

### **Worldwide Standards?**

The present multiplicity of transmission standards -- NTSC, PAL, SECAM, and their many variants -- necessitate conversion in order to use, in one system, programs made in another. The conversion equipment is quite expensive at present and causes some loss of quality, especially for moving images and in cases where the frame rate must be changed. It should be noted that, in spite of this, the entertainment value of converted programs is undiminished. There is no evidence that technical production standards have any influence whatever on the salability of programs. However, if all countries wanted easy international transmission (it is not clear that they do) it would be advantageous to use a common standard.

It is equally important to be able to convert easily between film and video. Much video programming has always been derived from film productions, and many theatrical presentations, especially outside the US, are now originally produced for television. The most difficult conversion problem is the frame rate -- 24 for film, and 25 or 30 fps for video in the 50- and 60-Hz countries. It is not possible to make theatre-quality films from NTSC or PAL because their spatial resolution is too low. Thus all productions intended for both the cinema and TV must now be produced on film, which has become the *de facto* international medium for exchange of programs.

With the advent of HDTV, it has become technically possible to use a new medium, namely high-definition TV, for both film and television production. This would be quite advantageous to producers, since video production is faster and cheaper than film, and editing is easier. Primarily for this reason, several groups have proposed the establishment of a worldwide TV production standard, naturally based on the NHK system, as it is the only system now available. It is not obvious that an HDTV standard for this essentially commercial purpose must be the subject of intergovernmental agreement, especially in the current deregulatory climate. There is much talk about avoiding another situation such as the existence of two competing formats, VHS and Beta, for home video cassettes. In the past most such problems were handled on a voluntary, nongovernmental basis. For example, The Society of Motion Picture and Television Engineers (SMPTE) has successfully coordinated standards for motion picture film. Proponents have also advanced a second argument, which has a more valid claim to the attention of governments. If a large proportion of TV-only producers, worldwide, were to adopt a common standard for production, it would also facilitate interchange of television programs, in the sense that they would not have to be converted at the production level.

Almost any 60-Hz interlaced standard would serve for this second objective, since most technical difficulties are associated with frame-rate conversion. Actually, the main reason why many industry leaders support the proposal is the hope that Europe will accept 60 Hz, rather than enthusiasm for the NHK system, as such. Presumably, this is also the rationale for its adoption by the State Department as the official position of the US.

This fall, the CCIR, the intergovernmental group that sets TV standards, considered such a proposal. Its advantages are clear. Equipment manufacturers, mainly Japanese, would be encouraged to market equipment, and its cost (and perhaps its price) would be lower because of standardization. TV and film producers would save money and gain convenience by going to all-video production methods.

Disadvantages to the proposed adoption are not easy to dismiss. All programs produced in NHK format must be converted for broadcast use, since there is no possibility at all that either the NTSC or PAL systems will be abandoned for regular broadcast use in the foreseeable future. Since the NHK system is 60 Hz, conversion to NTSC would be quite simple. A converter might cost \$25-50K. Conversion to the 50-Hz PAL system would be a little more complicated than present-day PAL-NTSC conversion. Converters are expected to cost \$200-250K and even then it is not assured that the quality of the conversion would be acceptable to meticulous European broadcasters. (In fairness, it should be noted that the motion artifacts associated with present-day 60-50 Hz converters are much less obvious than those always present in 24 fps films.) It is hard to see why they should go through this extra expense with no corresponding benefit.

Even accepting the quality of the conversion and ignoring the extra cost, the formal adoption of the NHK system would carry with it other and more far-reaching problems. If it were to be used for production only, the objections would not be overwhelming. Its production use as a foot in the door to broadcast use that most alarms the opponents, since it is in this application that the deficiencies are most critical. While Western proponents generally say that they have no intention of advocating NHK HDTV for any other purpose, this is not entirely true of the Japanese. For example, Kotaro Wakui, deputy director-general of NHK's network engineering branch, is quoted in *TV Digest* as stating that home video standards should be the same as production standards, and that a home video recorder is under development for the 1125-line system.

The NHK system is entirely analog, and the trend in professional TV equipment in recent years has been toward digital processing. It takes no advantage of modern signal processing ideas discussed above nor of advances in semiconductor technology that make such processing feasible. As a result, its system

design is no more advanced than that of the 32-year old NTSC system. Its efficiency is about the same. What was barely tolerable for the 6-MHz NTSC channel, however, is intolerable for the NHK 20- to 30-MHz channel. This wide bandwidth precludes the use of the system in existing TV channels and makes the associated equipment very expensive. It makes digitization especially costly, particularly the digital video tape recorder and digital special effects equipment, which many expect to figure prominently in future systems. The NHK system is certainly no better than NTSC with respect to motion rendition and flicker, and perhaps even worse, because the field of view is wider. Viewers generally do report less interline flicker with the NHK system than with NTSC, but this can only be true because the line structure is incompletely resolved or because of insufficient vertical resolution in the camera. In either case, it indicates even less efficient use of the channel capacity than NTSC.

If the NHK system is adopted now, the first effect is likely to be substantial capital investment by film/video producers in Japanese HDTV equipment. (It is doubtful that TV-only producers and broadcasters will buy much equipment, since they could not put it to profitable use.) The existence of this equipment and the associated investment will present a significant obstacle to the adoption of a more sophisticated system in the future. I believe that this will foreclose any opportunity for the US television industry to gain a meaningful share of the HDTV market. Admittedly, there is no guarantee that they will get much of the market in any case; but with a better system, on the scene several years from now, partly developed in the US, at least they will have a chance.

Of course, some may say that the Japanese deserve to dominate HDTV, since they alone were willing to make the investment to develop it to the present point. However, this situation is quite unlike that of Japanese cars, which American buyers prefer in devastating proportion because they perceive them to be superior to ours. The objections to the NHK system are not merely chauvinistic. Even in Japan, there is substantial support for compatible improvements in the NTSC system, rather than the use of the much wider bandwidth NHK system. A committee headed by the Ministry of Posts and Telecommunications, already working in this direction, is quoted in *Electronics* as pointing out that the HDTV signal cannot be broadcast in any terrestrial channel, and that sets smaller than 30" (too large for Japanese homes), do not need 1125 lines resolution.

The deficiencies pointed out above in its system design are real. The potential for much better systems is not imaginary. The need to adopt the NHK system at this time is not overwhelming, and the likely results of its adoption, in my opinion, are serious and against the national interest.

#### The Future of Television

Many upheavals may be in store for TV besides higher quality. The film theater may metamorphose into a TV theater fed by a wideband link. Cable and the video cassette recorder have already made great changes in the means by which programs are produced and distributed. There is a real question as to how programs will be sent to homes in the future, and how they will be paid for. Fiberoptic cables or satellite transmission may be used to carry the signals direct to homes. Many proposals have been made for additional services, such as videotex and various interactive informational, transactional, and educational systems that would utilize television, video storage, and computer equipment now in use or that might be installed in the future. There appears to be a role in all this for a greatly improved television system so that at least some of these services could support productions of essentially perfect technical quality.

We are in the midst of a period of extraordinary creativity in every aspect of television system design. Research results indicate that a very attractive system can probably be developed that will require little, if any, increase in bandwidth. The worry expressed by some, that the absence of a quick agreement on a production standard will lead to a proliferation of proposals for new systems, is not at all a cause for concern. It rather demonstrates that knowledge is being acquired at a rapid rate. Adoption of any HDTV system at the present time, for any purpose, is therefore premature. It is likely to discourage investment in TV research and stretch out by many years the time it will take to achieve a truly superior system.

### Glossary

**ATSC.** The Advanced Television Systems Committee. A US industrial group that has been developing voluntary standards for high-definition systems.

**CCIR.** International Consultative Commission on Radio. An intergovernmental standards-setting organization.

**DBS.** A system of transmitting TV programs directly to homes via satellites.

**DVE.** Digital Video Effects. A TV special effects systems that, in current systems, requires first digitizing the signal, then performing the desired process, and finally returning the signal to analog form.



**EBU.** The European Broadcasters' Union. The members are broadcasting authorities, mainly governmental. The EBU is in favor of MAC, but is considering the NHK system as a production standard. It is very concerned about conversion from 60 to 50 Hz.

**HDTV.** High-definition television. 1000 lines or more.

**MAC.** Multiplexed analog components. A color TV transmission system in which three separate signals, usually luminance and two chrominance signals, are time-multiplexed for transmission.

**NHK.** The Japan Broadcasting Corporation. The main broadcaster in Japan with extensive research laboratories. NHK led the development of the 1125-line HDTV system.

**NTSC.** National Television System Committee. An American industry group that proposed monochrome TV standards in 1941 and color standards in 1953. Standard setting in the US is the prerogative of the Federal Communications Commission, which historically has preferred to adopt industry-initiated schemes. In colloquial usage, "NTSC" refers to the current US color TV transmission standards. NTSC is also used in Japan, Latin America, and some other places.

**PAL.** Phase alternation by line. The European color television system, using 625 lines, 25 frames, 50 fields. Used in most 50 Hz countries.

**SECAM.** The French color television system, also used in Eastern Europe. Uses a different method of modulating color information onto the color subcarrier.

**SMPTE.** Society of Motion Picture and Television Engineers. The principal American professional organization for the motion picture and TV industries. Has foreign members and branches. Studies proposals and sets standards for new TV systems.

**UHF.** Ultra high frequency. The spectral band that includes channels 14-69. All receivers sold in the US are required by law to include UHF capability.

**VHF.** Very high frequency. The spectral band that includes channels 2-13.

## BOX ON THE SCANNING PRINCIPLE (Use a diagram or two here.)

The intensity (technically, irradiance) in the focal plane of the TV camera is a function of the space dimensions,  $x$  and  $y$ , as well as the time,  $t$ . I shall call this the *video function*. The TV system produces a version of this function on the surface of the picture tube or other display device. In order to do that, a *video signal* must be derived from the video function for transmission to the receiver. If we visualize the video function in a three-dimensional  $x,y,t$  space, this becomes a problem in sampling theory well understood by modern-day scientists. However, even without the aid of systems analysis, the idea of sampling by scanning was conceived no later than 1847 by Bain, and perhaps much earlier. In fact, a commercial facsimile system using the scanning principle was in service between Paris and Lyon around 1850.

Because of its relationship to the 3-d world, the video function has unique and important statistical properties. In mathematical terms, at each point in  $x,y,t$  space, the video function must be highly correlated. This is so because the function represents the projection, on a surface, of a collection of three-dimensional objects that are moving continuously in time. Thus the image of each point on the surfaces of these objects must trace a continuous path in  $x,y,t$  space, from the time it appears from behind some other object or from beyond the picture boundary, until it disappears from view. The direction of highest correlation is thus the direction of motion, and the intensity along this path, except for noise and perhaps for the effect of shadows, must be slowly and continuously changing. (In artificial intelligence circles, this phenomenon is called optical flow.) If the direction of motion is known, then many points along each flow line can be calculated from a relatively sparse sampling, thus permitting very accurate interpolation of intermediate frames.

## BOX ON COLOR

Newton discovered that the sensation of color depends on the spectral distribution of light, i.e., on the amount of energy at each visible wavelength. However, colors can be matched, or reproduced, by a suitable additive mixture of only three primary lights, the best set being narrow-band red, green, and blue. A color TV tube mixes colors in this way. Paint or dye mixtures, as used in color film and printing, are more complicated

because the physical situation is quite different. Such mixtures are called subtractive. The best set of dye primaries comprises cyan (blue-green), magenta, and yellow, the complementary colors to the primaries used for additive mixtures.

When color standards were first adopted in the US in 1951, a very simple field-sequential system promoted by CBS was chosen. Equal-resolution red, green, and blue images were transmitted in succession by means of rotating color wheels in front of camera and receiver. To fit this signal into the 6-MHz channel, the resolution and frame rate were reduced. The system was incompatible with the then-existing monochrome system, and lasted only two years.

In the NTSC system, the RGB components are transformed, by a change of coordinate system in color space, into luminance and two Cartesian chrominance components. At the transmitter, the latter are modulated onto a subcarrier near the top of the luminance band. The resulting composite signal is sent to the receiver where it is separated (demodulated) into components, several design measures having been taken in an effort to make it possible to perform the demodulation accurately. The amplitude of the subcarrier depends on the amount of color in the picture and is zero for neutral colors. Furthermore, the phase of this subcarrier is chosen so that it cancels out on successive frames. The subcarrier frequency is chosen so that the spectral lines used for color are interleaved with those used for luminance. These measures reduce, but do not eliminate, the interference between the chrominance and luminance signals. Some modern receivers use so-called "comb filters" to improve the separation, but only very expensive three-dimensional filtering at both transmitter and receiver would eliminate luminance/chrominance interference completely. These are the considerations that led to the development of the MAC systems.

# Psychophysics and the Improvement of Television Image Quality

By William F. Schreiber



Reprinted from the *SMPTE Journal*  
August 1984 Issue, Volume 93, Number 8

Copyright © 1984 by the Society of Motion Picture  
and Television Engineers, Inc.

# Psychophysics and the Improvement of Television Image Quality

By William F. Schreiber

Worthwhile improvement in television image quality is obtainable by signal processing at the receiver. However, improvement to the level demonstrated by NHK requires a large bandwidth expansion if only straightforward means are used, such as increasing the line and frame rates. This paper discusses a number of methods for obtaining maximum quality for a given bandwidth. Some of these methods take advantage of visual psychophysics, which is reviewed. Others deal with the special characteristics of TV cameras, displays, and scanning patterns. Quite complicated signal processing, expected to become practical in the next few years, is proposed to improve system performance.

Thirty years' experience with the NTSC and PAL TV systems has demonstrated the general soundness of the original concepts and the appropriateness of the chosen parameters. Despite the stringent constraints of compatibility with the then-existing monochrome system, picture quality has proven acceptable, the hardware sufficiently inexpensive and reliable, and a large industry has arisen based on this technology.

## Introduction

### Motivation

A number of forces have developed for changes in these systems with a view toward improving picture quality. One is the rapid increase in the variety and capability of semiconductor devices, especially memory, and the accompanying decrease in cost. Much more sophisticated signal processing is thus becoming feasible. Frame memories will probably become practical in receivers before the end of the decade. Many other improvements, such as comb filters and digital demodulation, are already practical. Other possibilities arise from digitization of post-production, which promises greater convenience and flexibility for the producer, more

complicated effects, higher signal-to-noise ratio (SNR), and perhaps pre-correction for certain degradations likely to be produced by channel and receiver.

The strongest impetus for improvement has undoubtedly come from the demonstration of the Japanese (NHK) HDTV system.<sup>1</sup> While it is not surprising that better pictures can be obtained with four to five times the bandwidth, impressive technological virtuosity was exhibited by the development of the system components, particularly the camera and picture tubes. The sight of vastly improved images, comparable to 35mm theatre quality, on real TV equipment, has whetted everyone's appetite for more improvements, but preferably with less increase of bandwidth.

The path to the practical application of these potential improvements is hardly clear. There are few channels suitable for the NHK system\* and there is a serious question as to whether, or by what means, a new system ought to be made compatible. Many possibilities for improvement have been demonstrated which do not require so much bandwidth.<sup>2</sup> Digitization for such a system would be considerably more difficult and expensive than in the case of NTSC.

The principal purpose of this paper is the discussion of methods, based on visual psychophysics and signal processing, by which maximum picture

quality can be obtained for whatever channel capacity is provided. It is recognized that there are many other important considerations in the design of new TV systems, such as removing the defects of NTSC, but they are not discussed here.

### TV as Visual Representation

In a sense, the TV system substitutes for directly viewing the original scene; hence its success in that role can be used as a measure of its performance. True "presence" is unattainable with any currently proposed system, not only because of the limited spatial and temporal bandwidth and field of view, but because of the two-dimensional (2-D) representation of a three-dimensional (3-D) scene. A truly serious limitation is the use of a single monocular camera of fixed gaze, perhaps panning to track a (single) moving object, while the viewers are many and are constantly moving their eyes over the scene. Finally, although the large-area color reproduction is often excellent, the dynamic range of a cathode-ray tube is far below that of most outdoor and many indoor scenes. The increased definition and field of view of the NHK system are steps in the right direction. However, its motion rendition is bound to be poorer than at present since the field of view is larger and the frame rate is the same.

### The Potential Contributions of Psychophysics

Psychophysical principles were applied in the development of both the monochrome and color NTSC proposed standards.<sup>3</sup> The relative horizontal and vertical resolution, the frame rate, the use of interlace, and the overall image quality goal were all selected in this manner in 1941 for the monochrome system. In the color deliberations, the primary psychophysical contributions concerned the representation of the color signal as luminance plus lower-resolution chrominance. Nonvisibility of the color

Presented at the Society's 18th Annual Television Conference in Montreal (paper No. 18-3) on February 10, 1984, by William F. Schreiber, Massachusetts Institute of Technology, Cambridge, MA. This article was received February 10, 1984, and also appears in *Television Image Quality*, published 1984, SMPTE. Copyright © 1984 by the Society of Motion Picture and Television Engineers, Inc.

\* Some examples are direct broadcasting from satellites (DBS), cable TV, and fiber optics.



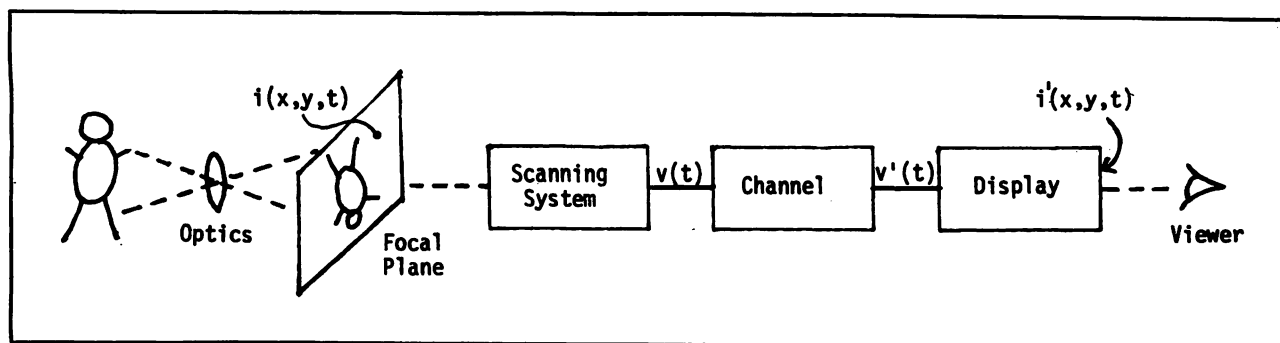


Figure 1. A generalized TV system.

subcarrier was more a hope than a fact, and the desired noninterference between chrominance and luminance never did exist, in general. There was no justification in psychophysics for the gross disparity between vertical and horizontal chrominance resolution. It is true that a number of the color-related problems of NTSC were less visible because of the properties of the then-existing transducers. In any event, virtually all contemporary proposals abandon the nonreversible mixing of chrominance and luminance.

One hope for future improvement rests on the considerable body of evidence that neither the NTSC nor the NHK system makes maximum use of the luminance bandwidth. The sampling theorem states that a certain 3-D bandwidth (the Nyquist bandwidth) should be recoverable "exactly," given the vertical and temporal sampling frequencies and the signal bandwidth. Yet, through a combination of factors, the system throughput, at the upper (3-D) limits of the Nyquist bandwidth, is much less than 100%. Furthermore, simply increasing the response would not increase perceived quality in most cases, since certain defects would become more obvious. A number of demonstrations have shown that much better pictures can be produced from the existing transmitted signal simply by up-converting the line and/or frame rate, thereby decreasing flicker and the visibility of the line structure.<sup>4</sup> Wendland has proposed the use of spatially interlaced sampling to accord more closely with the angular dependence of visual acuity,<sup>5</sup> and Glenn has proposed exploiting spatio-temporal interactions for much the same purpose.<sup>6</sup>

In this paper, we shall first describe the television process from the viewpoint of linear signal transmission theory, the input being the collection

of illuminated objects before the camera and the output being the picture display as perceived by the viewer. We shall then review some psychophysical data that characterize visual response under controlled (and, unfortunately, rather artificial) conditions. With this background, we shall calculate the required channel capacity for a variety of idealized systems, and show that no straightforward system can give greatly improved quality without unreasonable bandwidth expansion. Finally, we shall discuss a number of alternatives to current TV system design that exploit more thoroughly what is known about human vision. Most of these proposals involve signal processing considerably more complex than now used. Thus they may not become economically feasible for a number of years. When they do become practical, however, they promise a much better quality/bandwidth ratio than is now achievable. We shall not discuss the additional improvement that might be attained by statistical coding.

### The TV Chain as a Linear System

#### *A Generalized TV System*

As shown in Fig. 1, light from the scene before the camera is caused to form an image,  $i(x,y,t)$ , in the focal plane. We call this image the "video function." It is a vector for colored images. The purpose of the system is to produce a modified version,  $i'(x,y,t)$ , on the display device for viewing.

The video function is converted to a video signal,  $v(t)$ , by a scanning process operating on the charge image developed by the camera. A simple view of this process is that the signal produced from each point of the focal plane is proportional to the integrated light power that falls on the point between sampling times. The video signal is further processed by the channel

(modulation, filtering, digitization, transmission, etc.), producing a modified signal,  $v'(t)$ , to be applied to the display device. The display process can be thought of as tracing out, on the viewing surface, a scanning pattern (raster) like that in the camera, in which an amount of energy is emitted at each point of the raster proportional to the light energy collected at the corresponding element of the camera focal plane. In practice, the emitted energy is spread out over a time interval, almost always much shorter than one frame time.

This description reveals a significant difference between the original and reproduced video functions. The former is continuous in space and time, while the latter is highly discontinuous. If the output were continuous, the system could be characterized simply by its spatio-temporal frequency response which could then be compared with the corresponding sensitivity of the human visual system (HVS). This space-time discontinuity is the cause of much of the inefficiency in utilization of the channel capacity. Simple-minded elimination of the sampling structure by blurring, the use of long-persistence phosphors, or by viewing from a distance, attenuates important components of the transmitted signal as well as the structure.

#### *The Special Problem of Interlace*

Since 30 frames/sec, progressively scanned, produces totally unacceptable large-area flicker, interlace was introduced early in the history of TV development, doubling the flicker rate to 60 Hz while preserving the full number of lines in the frame. The only condition under which 30-Hz large-area flicker can result with interlace is when the average brightness of odd and even fields is significantly unequal, a rare event.

Interlace has problems as well as

advantages. It was recognized at an early date that with phosphors of persistence short enough not to cause interframe blurring, vertical motion could often produce a display with half the number of scan lines. Horizontal motion ought to produce serrated vertical edges, but usually does not because of camera integration. It was not generally recognized that for viewing distances at which the lines can be clearly resolved, the interline flicker rate is 30 Hz, easily seen as a shimmer. A side-by-side comparison of interlaced and noninterlaced images (the latter requires twice the bandwidth, of course) makes these differences very obvious.

Even at viewing distances at which the line structure cannot be resolved, 30-Hz flicker is clearly visible in interlaced pictures in areas having significant vertical detail. Flicker occurs when odd and even lines are sufficiently different at *any* resolvable spatial frequency. This flicker can be eliminated either by reducing the vertical resolution of the camera and/or the display, or by integrating over a full frame by some temporal averaging device. With any such method, spatio-temporal resolution is reduced.

In the light of these considerations, the subjective effect of interlace has never been fully investigated, since vertical resolution, whose role has only been appreciated recently, was not adequately controlled. Even so, Brown's early study concluded that for a 225-line TV image at 50 fL, viewed at eight times picture height, interlace produced a subjective increase in vertical resolution of only 24% (36% at 40 fL, and a mere 6% at 100 fL) in the line number compared with a progressively scanned image at 60 Hz.<sup>7</sup> A similar NHK study in 1982 showed an increase of 20% for a 1500-line picture viewed at two times picture height.<sup>8</sup> These numbers are so much lower than 100%, which would be obtained if interlace "worked," one wonders why it has been thought to be so effective. For present-day scanning standards, interlace clearly produces artifacts which become more troublesome as the vertical resolution is increased and as the image is more closely viewed, while the vertical resolution is increased only slightly.

#### *Special Properties of the Camera*

In most camera tubes, the target, which integrates the incident light at each point between successive visits of

the scanning beam, is almost completely discharged each field. The integration area, in the vertical direction, thus comprises at least two of the 525 nominal scan lines. A vertical pattern of 262.5 cycles per picture height (cph) is rendered with zero response, and a frequency of even half that is substantially attenuated, to a degree that depends on its phase. Yet the sampling theorem tells us that we ought to be able to use the full bandwidth of 262.5 cph. If, however, the vertical response of the camera is increased, as it readily can be, for example, in laser scanners, we see disturbing interline flicker.

In some modern CCD cameras which have one row of detectors for each scan line, the pairing of two lines of data for each output line is deliberate.<sup>9</sup> In cathode-ray camera tubes, the process is more complex due to the physics of target discharge and the shape of the electron beam.<sup>10</sup> In this case, dark areas are completely discharged by the leading edge of the beam, while bright areas are not fully discharged until passed over by the trailing edge. The resulting geometrical distortion and small-area tone-scale distortion are not very serious. More important is the fact that, as ordinarily operated, the vertical resolution of camera tubes is much less than the horizontal resolution (expressed as lines/mm on the target). This is fortuitous, since higher vertical resolution would make interlace even less acceptable. However, the result is that by employing interlace, we have sacrificed a significant portion of the theoretically available vertical definition, and with it, much of the expected benefit.

#### *The Picture Tube*

Cathode-ray display tubes are essentially linear; thus the integrated light output at each point can be found by convolving an ideal (zero spot diameter) raster with the beam cross section, which is generally Gaussian. With such a shape, the elimination of line structure by defocussing (or by blurring in the eye) also blurs the image. In color tubes, an additional factor is the structure of small phosphor spots. In a 19-in. diagonal shadow-mask tube with 0.31-mm triad spacing, only about four triads are available for each picture element in the NTSC system. This is bound to introduce a great deal of spatial high-frequency noise, to which, fortunately, we are not very sensitive. However, the

channel SNR for AM transmission is uniform with bandwidth and therefore does not take advantage of this phenomenon.

#### *The Channel*

In present-day systems, the purpose of the channel is to reproduce at the picture tube the output of the camera tube with perhaps some minor amount of processing. Of course, noise is invariably added in the process and there may be some loss of bandwidth. As pointed out later, the most probable source of major improvement would be the introduction of substantial signal processing between camera and channel and between channel and display. Since the second processor must be cheap, that is the location of the significant technological challenge.

#### **The Psychophysical Background**

##### *Normal Seeing*

The HVS, presumably as a result of evolution, is well adapted to rapidly deriving a large amount of useful information from the scene before the observer. This scene, 3-D, variously illuminated, and moving, produces slightly different 2-D images on the retinas of the two eyes. The eyes are in constant voluntary motion over the scene, both by head motion and by rotation in their sockets. They also execute small involuntary motions which have been found to have an important, even essential, role in vision.<sup>11</sup>

The retina consists of a matrix of receptors of two kinds — cone cells which exclusively cover the central 2° (the fovea) and whose density decreases away from the axis, and rod cells whose density is maximum 15° from the center. Cones are responsible for the high visual acuity on axis and for color vision at normal (photopic) levels. The rod cells, which are much more sensitive, provide off-axis low-light-level (scotopic) sensitivity but have much lower spatial resolution.

The sensitivity of each cell depends on its state of adaptation and on the excitation of its neighbors. The sensitivity is characterized by both a static (input-output) function and a frequency response. Spatial resolution of point objects is roughly equal to cell spacing, but because of cooperation of retinal receptors, resolution of long parallel lines is much finer than the cell spacing. The discrete nature of the

retinal mosaic is never obvious in normal vision. Much visual processing is carried out on the retina itself, but additional processing occurs at higher levels of the nervous system.<sup>12</sup>

#### Characterization of Visual Response

Learning about vision is a frustrating study, since the vast literature would take years to master, yet data are lacking on many points that are vital to the design of efficient systems. Although all aspects of the visual sense are remarkably interdependent, it is customary to begin by discussing its performance along separate axes.

#### Contrast Sensitivity

By contrast sensitivity, we mean the visual response as a function of luminance, although what is usually measured is the just-noticeable difference between near-equal luminances displayed side by side or one after the other. Obviously, temporal or spatial separation is essential for measuring contrast thresholds, so that it is quite impossible to separate contrast sensitivity completely from these other variables.

With the usual test field (Fig. 2), the observer is allowed to adapt to the surround,  $L$ , and then the smallest discernible  $\Delta L$  is found.<sup>13</sup> The result of this measurement (Fig. 3) shows that  $\Delta L/L$  is nearly constant over five decades. We can thus see over an enormous luminance range, given time to adapt. The constancy of  $\Delta L/L$  is called the Weber-Fechner law, the fraction being as small as 1% under optimum viewing conditions.

In the more normal situation — observing actual scenes or their reproductions — the degree of adaptation is much less. If we now measure  $\Delta L/L$  as a function of the adapting luminance  $L_0$ , using a target such as that of Fig. 4, we find that the operating dynamic range is much smaller. More significant is the appearance of the central patches as a function of the

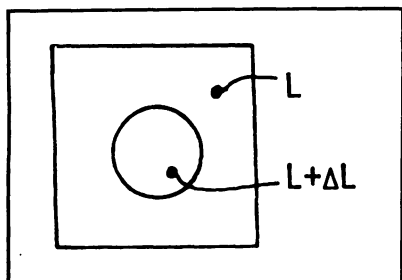


Figure 2. Contrast sensitivity target.

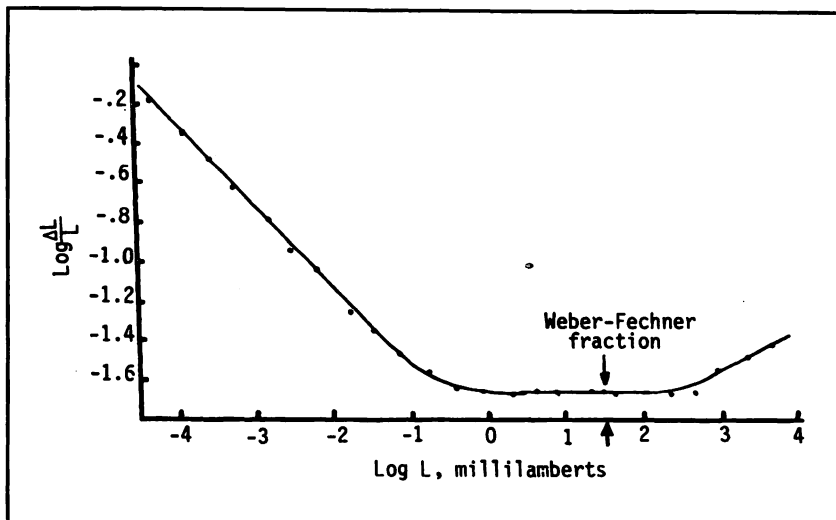


Figure 3. Contrast sensitivity data of Koenig and Brodhun, 1884. (Quoted by Hecht, *J. Gen. Physiol.* 7:421, 1924.)

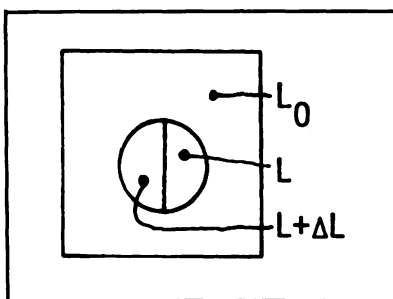


Figure 4. A more realistic contrast sensitivity target.

relative brightness. When the surround is about 100 times brighter than the central area, the latter looks black, no matter what its actual luminance, while in the reverse case it looks white.<sup>14</sup> When the central area is sufficiently intense, it appears to be a light source rather than an illuminated surface.

For picture reproduction, this means that nearly four decades of dynamic range are required to give the visual impression of a real high-contrast scene, such as outdoors on a clear day. This condition is approximated by optical projection from film with good equipment in a perfectly dark room. Under all other conditions, such as TV displays, the dynamic range must be compressed. Although this can be done so as to give pleasing results in terms of brightness and contrast as those terms are normally used, it is very hard to impart realism.

#### Temporal Frequency Response

Flicker and motion rendition are associated with the temporal response factor, so it is of great importance and

has had the attention of psychologists for many years. The "purest" method of measurement (least contaminated by other factors) is to superimpose a sinusoidally fluctuating component on a constant luminance and to use a very wide field with defocused edges. The definitive measurement has been made by Kelly.<sup>15</sup> The most interesting aspect of his results is that over a significant range of temporal frequencies, the HVS is a differentiator (Fig. 5), not an integrator. Flicker in this range is very noticeable. It is quite evident that at 25 or 30 Hz, flicker is almost always present. At 50 or 60 Hz, it is present in very bright images. To avoid flicker in the worst case, which is at the edges of a bright, wide-field display, 80 or 90 Hz might be needed. Note that peripheral flicker is sometimes seen in wide-screen motion pictures, where the flicker rate is usually 72/sec.

#### Spatial Frequency Response

Visual acuity — the ability to see sharply and resolve small details — is one of the most obvious aspects of vision. Although threshold contrast is often measured as a function of spatial frequency using square-wave gratings at various angles, the results are easier to interpret if sine-wave gratings are used.<sup>16</sup> A variety of indirect methods have also been used, in which transient response<sup>17</sup> or response to filtered random noise<sup>18</sup> has been measured. Despite the hazards of applying linear analysis to such a nonlinear system, all the results are similar to those shown in Fig. 6. Remarkably, the spatial characteristic also shows a differentiation region, one of the effects of

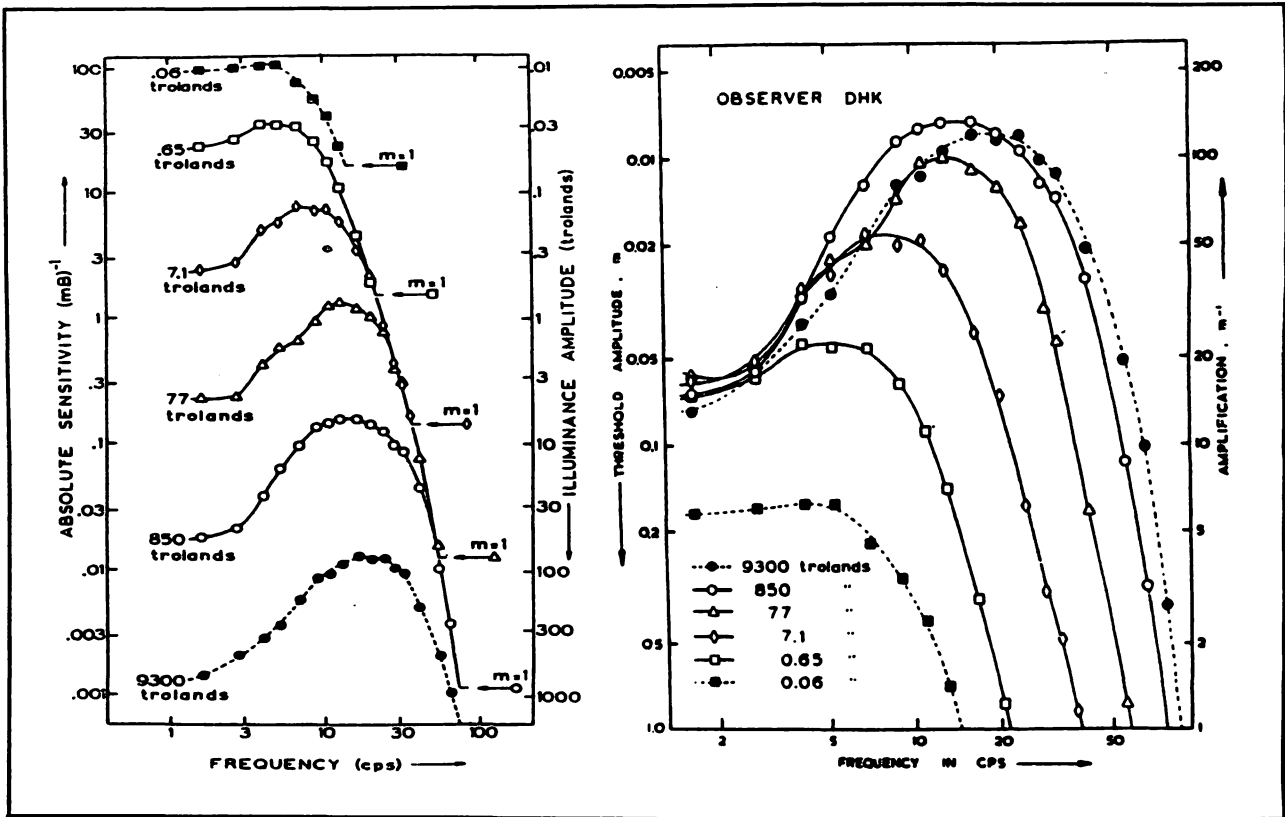


Figure 5. Kelly's temporal data, plotted two ways. (From D. H. Kelly, "Visual Response to Time Dependent Stimuli. I. Amplitude Sensitivity Measurements," *J. Opt. Soc. Am.*, Vol. 51, No. 4, 1961, pp. 422-429.)

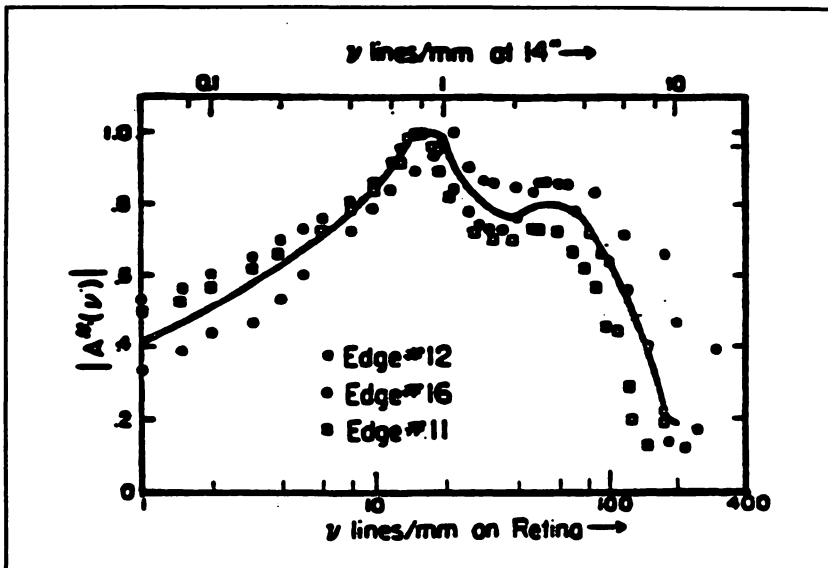


Figure 6. Spatial frequency data of Ref. 17. (E. M. Lowry and J. J. DePalma, "Sine Wave Response of the Visual System," *J. Opt. Soc. Am.*, Vol. 51, No. 10, 1961, p. 474.)

which is to sharpen images significantly. It is thought that this effect is due principally to neural interaction on the retina. Note that there is some response up to 30 or more cycles per degree (cpd). For a 90° display, 2700 cycles, or 5400 picture elements, would be required for absolute invisibility of the scanning structure. For the lines in

an NTSC picture to disappear completely, only a 16° field can be covered.

In the vertical and horizontal directions, the spatial frequency response is almost equal, but at 45° it decreases by a factor of 2 or more. This is the main reason why half-tone patterns are usually at 45°. It is also the basis of

proposals for interleaved sampling.<sup>19</sup> Whether this would be advantageous is hard to say. Baldwin carried out a very careful experiment to determine the effect on picture quality of varying the relative horizontal and vertical resolution.<sup>20</sup> For pictures of 56-in.<sup>2</sup> area with about 36,000 resolvable elements viewed at 30 in., the just-noticeable degree of asymmetry was 2.5:1, despite the equal horizontal and vertical limiting resolution of the eye. Thus it is not obvious that interleaved sampling would improve image quality, although it might make structure less visible.

#### Spatio-Temporal Interactions

Measuring the combined effect of spatial and temporal fluctuations is more difficult, and a wider variety of methods can be used. There is reasonable agreement that peak sensitivity is at about 2 Hz and 2 cpd, with integration at higher frequencies and differentiation at lower frequencies. This has been modelled as the difference between an excitatory and an inhibitory response, an interesting point, but not of direct value to the system designer.<sup>21</sup> There is evidence that the shape is somewhat more complicated,

but the essential result is that the derived passband in 3-D frequency space is not cubical, but more nearly ellipsoidal. No one seems to have repeated Baldwin's experiment for spatio-temporal resolution. Just because the limiting spatial and temporal frequencies have been shown to be inversely related, does not prove, in a system where the signal components are well below the limiting frequencies, that the system bandwidths ought to be so related.

#### Masking

Of great interest to the psychophysicist, and in this case of equal interest and value to the system designer, is the phenomenon of masking. In all sense modalities, response to particular kinds of stimuli is reduced significantly by the presence, in the immediate spatio-temporal neighborhood, of similar stimuli. In the case of large, uniform, slowly changing scenes, we call this phenomenon adaptation. It is of great value because it enables us to see well under a wide variety of conditions.

A similar phenomenon occurs for stimuli of similar spatio-temporal content. Exposure to a spatial grating reduces sensitivity to gratings of similar spatial frequency seen just afterwards or even just before.<sup>22</sup> Exposure to a temporal sinusoid reduces sensitivity to sinusoidal flicker of like frequency.<sup>23</sup> The presence of "activity" (sharp edges or fine detail) reduces noise sensitivity in nearby areas.<sup>24</sup> An example of the latter in the space domain is the much lower visibility of additive random noise in detailed or "busy" image areas and its much higher visibility in relatively blank areas. For this reason SNR, even weighted according to the variation of noise visibility with frequency, is a very poor indicator of image quality. Simple images require a much higher SNR than complicated ones for the same visual quality.

A related phenomenon is the masking of detail in a new scene by the presence of a previous scene. Repeating an earlier experiment by Seyler<sup>25</sup> in our own laboratory, we found that a new scene could be radically defocused and then refocused with a time constant of 0.5 sec, without visible effect. Recently, Glenn has demonstrated that it takes about 0.2 sec to perceive higher spatial frequencies in newly revealed areas.<sup>26</sup> This effect is nature's gift to temporal differential

transmission systems, since it allows new scenes to be built up over a period much longer than a frame.

#### Motion Rendition

Americans believe that it was Edison<sup>†</sup> who discovered that the illusion (sic) of motion could be produced by viewing a rapid sequence of slightly different images. This is the "phi motion" of psychology, in which the successive flashing of two small lights, with the appropriate time and space separation, makes it appear that a light moves from the first position to the second.<sup>27</sup> Should the angular jump be too large or the interval too long, the motion effect is discontinuous, and in some cases, can even be retrograde. We have all seen wheels standing still or even moving backward. This stroboscopic effect, which has its uses, of course, is an example of temporal aliasing. Like other kinds of aliasing, it is but one possible defect that should be traded off against others for optimum image quality. The smoothness of motion is directly related to filling the gaps between successive positions. The degree of temporal bandlimiting required to preclude temporal aliasing absolutely, has the effect of blurring moving objects.<sup>28</sup> Especially in low-frame-rate systems, it may be preferable to show a sequence of sharp still images rather than a continuously moving image so blurred as to be useless.

Careful observation shows that motion is generally smoother in TV than in motion pictures. This is because the TV system actually takes 60 pictures/sec, as compared to 24 for film. In addition, most TV cameras integrate for the full  $1/60$  sec, while all motion-picture cameras use exposure times of less than (and sometimes very much less than)  $1/24$  sec.

Objects that move across the retina while the eyes are fixated elsewhere are blurred by the temporal upper frequency limit of the HVS. The same thing happens in TV cameras, which is harmless unless the observer happens to be tracking the object. In that case the TV (or motion-picture) representation is disappointing. There is thus no way the TV camera can satisfy the entire audience when the scene contains two or more important moving objects.

<sup>†</sup> No doubt other countries have their own favorite inventors of motion pictures.

#### Color

Color is not a principal preoccupation of this paper since colorimetry is quite satisfactory in existing systems. In a new system design not constrained by the requirement of compatibility, however, there are several simple ways of adding color to a monochrome signal. Based on the lower required spatial color resolution, these methods increase the channel capacity by 20% or less.<sup>29</sup> More complicated systems decrease the color penalty even more.<sup>30</sup> At these incremental levels, the color picture, with slightly lower luminance resolution, is usually far superior in perceived quality, almost however measured, to the monochrome picture with slightly higher luminance resolution.<sup>31</sup> Thus, the addition of color can be viewed as a valuable way to *decrease* the total channel capacity for a given subjective quality.

It is also possible, but not yet demonstrated as far as we know, that the required temporal bandwidth for color is less than for luminance, in which case an additional possibility for compression would be available.<sup>32</sup>

#### Performance Goals for TV Systems

##### Perfection

A perfect system can be defined as one that gives a convincing illusion of reality. This probably would not require 3-D reconstruction, as might be done holographically. A very wide field of view is quite effective.<sup>33</sup> Assume that 90° vertically and 180° horizontally would be enough. The question, then, is the required resolution. A frame rate of 100/sec would certainly prevent flicker, but even that would not keep rapidly moving objects in focus. Using 50 cpd as the upper perceptual limit, a raster of about 9000×18,000, or 162 million samples/frame would suffice, for a total rate of 16 billion samples/sec. Of course such a signal would have very high redundancy and could be greatly compressed. Nevertheless, the obstacles to constructing such a system are insurmountable at present.

##### Idealism

An ideal system, for our purpose, can be based on resolution parameters so high that raising them would not materially improve quality. We would, however, accept a more limited field of view and the motion rendition obtainable at 60 frames/sec. For a 45° × 90° field of view and a sampling density of

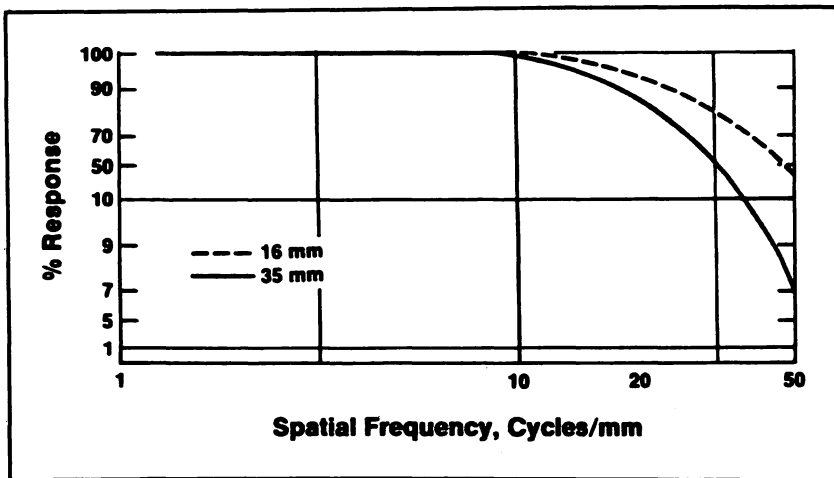


Figure 7. Overall response of film system. (From R. C. Sehlin, et al, *SMPTE Journal*, December, 1983.)

12/mm at normal viewing distance (30 cm), which is considered excellent quality for continuous-tone color prints, we would have a more modest 3000×6000 raster — a mere 18 million samples/frame, or 1 billion samples/sec. To get the full benefit from this resolution, we should probably raise the frame rate somewhat — perhaps to 80/sec, for a rate of 1.25 billion samples/sec. These pictures would give the effect of looking at the real world through a large window, except that if we were to track rapidly moving objects (those that move across the screen in less than 4 sec or so), we would see a definite loss of resolution.

#### Theatre Quality

The term “theatre quality” is very poorly defined, since film is getting better and better,<sup>34</sup> and we now know how to make nearly diffraction-limited optics. For the sake of discussion, for 35mm film with a frame height of 18mm and 30 to 50 line pairs/mm assumed for the effective resolution limit, 1080 to 1800 lines and 1.5 to 4.3 million samples/frame would be required. This is in accordance with the NHK experience.

Equating TV and film quality is not simple and certainly requires careful subjective testing. The spatial frequency response of film and optical

systems tends to fall monotonically, starting at a low spatial frequency (Fig. 7). Television systems have a rather well-defined upper frequency limit, but within the passband we are free to use almost any characteristics we wish. A considerable degree of sharpening is possible and is routinely used in electronic-based graphic arts systems.<sup>35</sup>

#### Vision-Based Design

##### *Spatial Filtering, Sampling, and Interpolation*

Input and output still images are inherently spatially continuous. When represented by an array of numbers, the continuous-discrete and discrete-continuous conversions can have a significant effect on image quality. Since analog TV is sampled only in the vertical direction, this section applies principally to processing designed to give maximum vertical sharpness without artifacts.

A basic problem with discrete imaging systems lies with the sampling theorem, which states that the recoverable signal bandwidth (Nyquist bandwidth) is one-half the sampling frequency. At the transmitter, the bandwidth should therefore be limited to the Nyquist value to prevent aliasing, but there is no obvious mecha-

nism for accomplishing this in a TV camera. Furthermore, if we somehow did implement such a filter, the ringing associated with sharp horizontal edges would be unacceptable. There is an optimum filter and its implementation will be discussed shortly.

At the display, it clearly would be desirable to eliminate the scan lines. They are obtrusive and, due to the masking effect, suppress the high-frequency structure to some extent. Achieving this result by defocussing the more or less Gaussian scanning beam causes noticeable loss of sharpness. Relying on the filter of the HVS produces a similar effect, although with less loss of sharpness. In any event, the effect of the HVS is strongly dependent on the angular subtense of the scan lines at the eye (Fig. 6). When viewing an NTSC picture at 4H, the line structure is 34 cpd. At these spatial frequencies, visual response drops about 18 dB/octave. Thus the line structure is attenuated 18 dB compared to the signal components at the upper end of the Nyquist band. However, to achieve this separation, the signal components are also attenuated substantially. Viewed at 2H, the relative attenuation is only about 12 dB.

Vertical filtering can be done effectively by operating both camera and display at a substantially higher line rate, and interposing processing elements between camera and channel and channel and display (Fig. 8). At both camera and display, this could give enough vertical samples to implement the appropriate digital filter. At the display, such up-conversion would also raise the line rate to a point where the HVS could more easily separate the structure from the image. Incidentally, but perhaps importantly, high-line-rate operation of the camera might well ameliorate the problems discussed above due to the nonlinear target discharge, especially with progressive scanning. The application of similar methods to still pictures has resulted in a channel capacity saving of as much as 40% for the same perceived quality, as compared with simple-minded methods.<sup>36</sup>

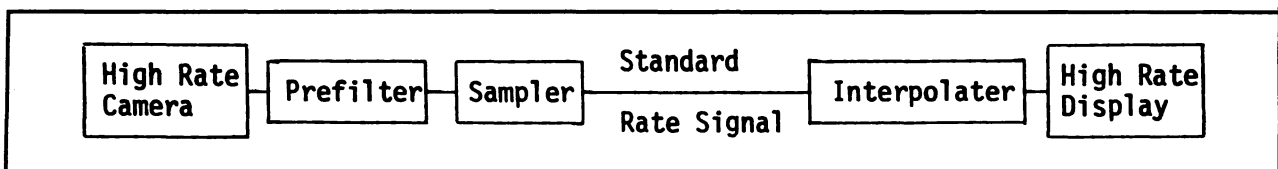


Figure 8. The modified TV chain.



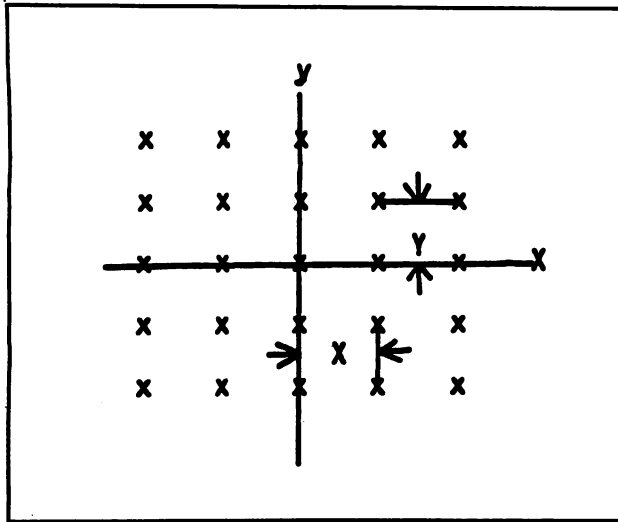


Figure 9. Cartesian sampling.

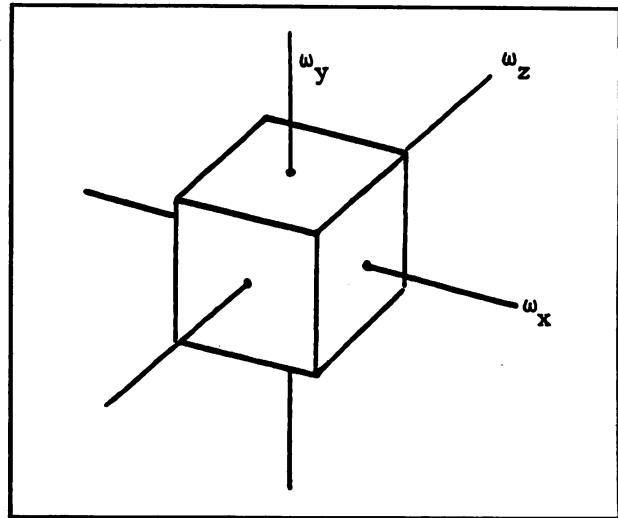


Figure 10. Cartesian spectrum.

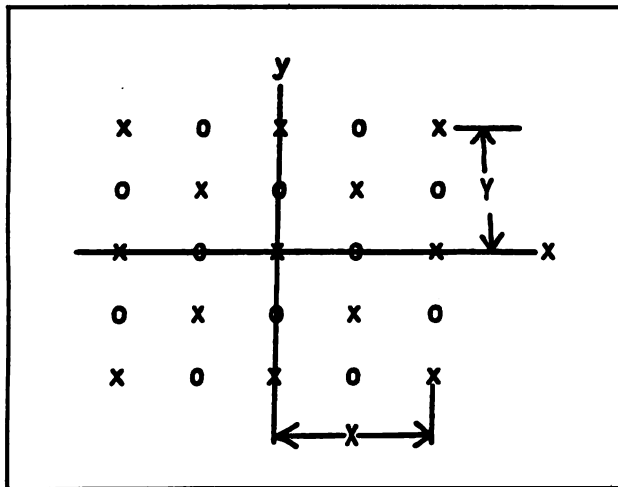


Figure 11. Temporal interleaving: x=even fields; o=odd fields.

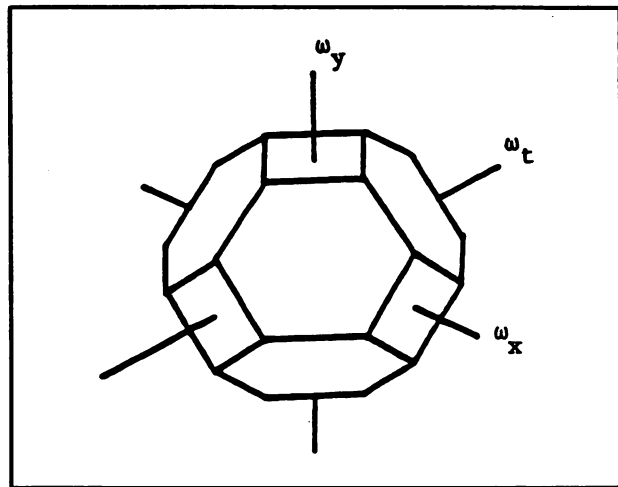


Figure 12. Alias-free bandwidth for temporal/spatial interleaving.

### Temporal Filtering, Sampling, and Interpolation

This situation is analogous to that caused by spatial sampling. In this case, the sampling theorem tells us that 30 frames/sec results in a 15-Hz Nyquist bandwidth. To avoid aliasing (jerky or stroboscopic motion), we should low-pass filter before sampling. Since a camera that integrates perfectly for the frame (or field) time is hardly an ideal filter, we could tailor presampling filters much more accurately if the camera operated at four or five times the frame rate. At the display, the extra samples would have a similar effect, but in addition, as in the spatial case, would raise the rate of the flicker so that it could more easily be separated from the baseband by the HVS.

### Three-Dimensional Processing

The spatial and temporal processing

discussed above could be combined so that the filters of Fig. 8 would be 3-D. Such filters require frame and line stores and can only be implemented digitally. In such a TV system, discrete in all three dimensions, the question of the sampling pattern of the channel signal,<sup>†</sup> as well as the corresponding 3-D Nyquist bandwidth, must be dealt with.

The Cartesian pattern of Fig. 9 gives the Cartesian spectrum of Fig. 10, while the interleaved pattern of Fig. 11 gives the odd-looking spectrum of Fig. 12. (Other 3-D patterns are possible.) This pattern trades off spatial and temporal bandwidth in a manner that probably is better than the Cartesian pattern, although observer tests are necessary to be sure. It has higher spatial response at low temporal

<sup>†</sup> The sampling patterns of the camera tube and display are of little importance since they will not be detected by the viewers.

frequencies and vice versa, and higher vertical and horizontal resolution than diagonal.

Interleaved sampling bears some relationship to present-day interlace, which is used primarily to double the flicker rate. However, we can think of the vertically offset sampling of standard interlace as a means of raising the (time-averaged) vertical resolution for a given vertical scan rate. In this endeavor it mostly fails, for the reasons cited. Interleaved sampling as described here, however, when used in conjunction with appropriate 3-D presampling and interpolation filters, has none of the defects of ordinary interlace. It aims for and gets no "free" expansion of bandwidth. In fact, the volume of the 3-D Nyquist bandwidth is identical for all sampling patterns that have the same number of samples per unit  $(x,y,t)$  volume. What interleaved sampling does do is to change

the shape of the Nyquist bandwidth from Cartesian to one that may be better.

It would be difficult to implement a filter with the response shown in Fig. 12. However, an ellipsoidal impulse response would approximate it and, if Gaussian, would be separable and therefore practical. Since a form of Gaussian filter was found optimum in the studies cited,<sup>20</sup> it is quite likely to work well in this case.

### Multi-Channel Systems

There is some evidence that the HVS treats various spatial frequency components of the visual stimulus so differently that an advantage can be gained by separating the signal into two or more channels and using different transmission parameters for each component. This is quite in accord with widely held theories that the visual system is organized in this manner.<sup>37</sup> A number of systems quantize low and high spatial frequencies differently, using a rather coarse quantization in the high channel.<sup>38</sup> The quantization noise, preferably randomized,<sup>39</sup> tends to be masked by the high-frequency detail.

Glenn has suggested that the high frequencies can be transmitted at a lower frame rate.<sup>6</sup> Although some trade-off between spatial and temporal response is possible by offset sampling, Glenn goes much further, transmitting the highs at only 5 frames/sec. In the case of newly uncovered stationary detail (a new scene, or newly revealed background that emerges from behind a moving foreground object), this seems to work rather well. In the case of detailed objects moving in the scene, the blurring must be much worse than at 30 frames/sec. It is possible that visual acuity in the tracking mode is sufficiently low that this is permissible. Clearly more work needs to be done on this technique, since if successful it would permit substantial saving.

### Conclusion

We have described TV transmission as a problem in the analysis of linear systems. A review of the literature on visual psychophysics as it applies to this formulation has revealed a number of possibilities for the improvement of picture quality in relation to channel bandwidth. These involve 3-D processing at both transmitter and receiver, and, in the latter case, would be practical only with a high degree of circuit integration.

Specific visual problems due to the characteristics of camera and display devices and to the use of interlace have been pointed out. The amelioration of such effects by operating these devices at very high line and frame rates requires rather complicated signal processing, but presents the prospect of significant improvement in the utilization of transmission channel capacity.

### Acknowledgments


The literature in this field is large and growing rapidly, so the list of references should be considered representative and not exhaustive. Of the ideas presented, many have grown out of discussions with colleagues, students, and friends. I have had the assistance of G. Saussy in collecting psychological references. S. Sabri motivated me to write this paper.

### References

1. T. Fujio, et al., "HDTV," NHK Tech. Monograph No. 32, June 1982.
2. K. Lucas, et al., "Direct TV Broadcasts by Satellite," IBA Report No. 116, 1981.
3. D. G. Fink, "TV Standards and Practice" (1941 NTSC), McGraw Hill, 1943; Proc. IRE, Jan. 1954 entire issue (1953 NTSC).
4. C. P. Sandbank and M. E. B. Moffatt, "HDTV and Compatibility with Existing Standards," *SMPTE J.*, May 1983, pp. 552-561.
5. B. Wendland, "Extended Definition TV with High Picture Quality," *SMPTE J.*, October 1983, pp. 1028-1035.
6. W. E. Glenn, et al., "Compatible Transmission of HDTV Using Bandwidth Reduction," IGC paper, unpublished.
7. E. F. Brown, "Low-Resolution TV: Subjective Comparison of Interlaced and Non-Interlaced Pictures," *Bell Sys. Tech. J.*, Vol. 46, January 1967, pp. 199-232.
8. T. Mitsuhashi, "Scanning Specifications and Picture Quality," NHK Tech. Monograph No. 32, June 1982.
9. N. Murata, et al., "Development of a 3-MOS Color Camera," *SMPTE J.*, December 1983, pp. 1270-1273.
10. T. G. Schut, "Resolution Measurements in Camera Tubes," *SMPTE J.*, December 1983, pp. 1287-1293.
11. L. A. Riggs, et al., "The Disappearance of Steadily Fixated Test Objects," *J. Opt. Soc. Am.*, Vol. 43, 1953, pp. 495-501.
12. T. N. Cornsweet, *Visual Perception*, Academic Press, 1970.
13. Data of Koenig and Brodhun (1884) quoted by S. Hecht, *J. General Physiology*, Vol. 7, 1924, p. 421 ff.
14. E. G. Heinemann, "Simultaneous Brightness Induction," *J. Experimental Psychology*, Vol. 50, 1955, pp. 89-96.
15. D. H. Kelly, "Visual Responses to Time Dependent Stimuli. I. Amplitude Sensitivity Measurements," *J. Opt. Soc. Am.*, Vol. 51, No. 4, 1961, pp. 422-429.
16. F. W. Campbell and J. G. Robson, "Application of Fourier Analysis to the Visibility of Gratings," *J. Physiology*, No. 187, 1968.
17. E. M. Lowry and J. J. DePalma, "Sine Wave Response of the Visual System," *J. Opt. Soc. Am.*, Vol. 51, No. 10, 1961, p. 474.
18. O. R. Mitchell, Ph.D. Thesis, MIT EECS Dept., 1972.
19. B. Wendland, "On Picture Scanning for Future HDTV Systems," IBC, 1982.

20. M. W. Baldwin, "The Subjective Sharpness of Simulated TV Pictures," Proc. IRE, October 1940, pp. 458-468.
21. C. A. Burbeck and D. H. Kelly, "Spatio-Temporal Characteristics of Visual Systems," *J. Opt. Soc. Am.*, Vol. 70, No. 9, 1980; and F. L. Van Nes, et al., "Spatio-temporal Modulation Transfer in the Human Eye," *J. Opt. Soc. Am.*, Vol. 57, 1967, pp. 1082-1088.
22. G. Sperling, "Temporal and Spatial Visual Masking," *J. Opt. Soc. Am.*, Vol. 55, 1965, pp. 541-559.
23. O. Braddick, F. W. Campbell, and J. Atkinson, "Channels in Vision: Basic Aspects," in *Handbook of Sensory Physiology*, Vol. VIII, Springer-Verlag, New York, 1978.
24. A. N. Netravali and B. Prasada, "Adaptive Quantization of Picture Signals Using Spatial Masking," Proc. IEEE, Vol. 65, No. 4, April 1977, pp. 536-548.
25. A. J. Seyler, et al., "Detail Perception after Scene Changes in TV," IEEE Trans. Information Theory, Vol. IT-11, January 1965, pp. 31-43.
26. W. E. Glenn, "Compatible Transmission of HDTV Using Bandwidth Reduction," National Association of Broadcasters, Las Vegas, April 12, 1983, videotape demonstration.
27. R. A. Kinchla, et al., "A Theory of Visual Movement Perception," *Psychological Review*, Vol. 76, 1969, pp. 537-558.
28. J. Korein, et al., "Temporal Anti-Aliasing in Computer Generated Animation," *Computer Graphics*, Vol. 17, No. 3, July 1983, pp. 377-388.
29. W. F. Schreiber and R. R. Buckley, "A Two-Channel Picture Coding System: II — Adaptive Companding and Color Coding," IEEE Trans. on Communications, Vol. COM-29, No. 12, December 1981, pp. 1849-1858.
30. R. R. Buckley, Ph.D. Thesis, MIT EECS Dept., 1982.
31. U. Gronemann, Ph.D. Thesis, MIT EE Dept., 1964.
32. H. de Lange, "Research into the Dynamic Nature of the Human Fovea-Cortex System," *J. Opt. Soc. Am.*, Vol. 48, 1958, pp. 777-784.
33. W. C. Shaw, et al., "IMAX and OMNIMAX Theatre Design," *SMPTE J.*, March 1983, pp. 284-290.
34. M. A. Kriss, et al., "Photographic Imaging Technology for HDTV," *SMPTE J.*, August 1983, pp. 804-818.
35. W. F. Schreiber, "Wirephoto Quality Improvement by Unsharp Masking," *Pattern Recognition*, Vol. 2, March 1970, pp. 117-121. See also W. F. Schreiber and D. E. Troxel, U.S. Patent 4,268,861, May 1981.
36. J. R. Ratzel, Sc.D. Thesis, MIT EECS Dept., 1983. See also W. F. Schreiber and D. E. Troxel, "Transformation between Continuous and Discrete Representations of Images: A Perceptual Approach," MIT Report, submitted for publication.
37. L. Maffei, "Spatial Frequency Channels: Neural Mechanisms," in *Handbook of Sensory Physiology*, Vol. VIII, Springer-Verlag, New York, 1978.
38. D. E. Troxel, et al., "A Two-Channel Picture Coding System: I — Real-Time Implementation," IEEE Trans. on Communications, Vol. COM-29, No. 12, December 1981, pp. 1841-1848.
39. L. G. Roberts, "Picture Coding Using Pseudo-Random Noise," IRE Trans. on Information Theory, Vol. IT-8, February 1962, pp. 145-154.

### Bibliography

- E. C. Cartarette, et al., ed., *Handbook of Perception*, Vol. V, "Seeing," Academic Press, 1975.
- T. N. Cornsweet, *Visual Perception*, Academic Press, 1970.
- Handbook of Sensory Physiology*, Springer-Verlag, 1973.
- D. E. Pearson, *Transmission and Display of Visual Information*, Pentech Press, 1975.
- C. G. Mueller, *Sensory Psychology*, Prentice-Hall, 1965.
- G. J. Tonge, "Signal Processing for HDTV," IBA Report E8D, July 1983, USA. 

102687573

# MIT Industrial Liaison Program

## Report

Papers relevant to the symposium:

"MEDIA TECHNOLOGIES"  
October 3, 1985

"Intelligent Telephones"  
by  
Mr. Christopher M. Schmandt

- 1) "A Conversational telephone messaging System," by Chris Schmandt and Barry Arons
- 2) "Voice Interaction in an Integrated Office and Telecommunications Environment," by Christopher Schmandt, Barry Arons, and Charles Simmons
- 3) Copy of slide used during presentation by Christopher Schmandt

These papers have been duplicated at the request of the speaker.



Distributed for Internal Use  
by Member Companies Only.  
May Not be Reproduced.

© MIT

Voice Interaction  
in an  
Integrated Office and Telecommunications Environment

Christopher Schmandt, Barry Arons, and Charles Simmons  
Media Laboratory, Massachusetts Institute of Technology

## Introduction

The *Conversational Desktop* explores the use of speech input/output technologies for machine mediated voice communication in an office and telecommunications environment. Central to this work is an interface design which models several aspects of human conversational behavior. These include the ability to carry on a dialog to resolve ambiguous input, the ability to apply syntactic and acoustical context to the progress of the conversation, and sensitivity on the part of the machine to when *it* is being addressed by human voice. The latter is particularly relevant in an environment in which speech is being used for a variety of purposes, such as audio memos, speaking over a telephone, and alarm functions, in addition to the command channel to control the machine.

Earlier work had demonstrated the utility of dialog based on syntactic analysis as an approach to coping with recognition errors [Schmandt 82], although the parser used was hard coded for the particular application and extensible only in design. The *Phone Slave* [Schmandt 84, Schmandt 85] successfully exploited people's willingness to participate in a computer driven conversation, but it was a fairly passive system which made little use of knowledge of other activities its owner was engaged in. This project attempts to synthesize both of these approaches.

## The Environment

This project is based on the concept of an integrated office workstation which combines the functions of a powerful personal computer and an intelligent telecommunications system. In addition to conventional personal computer applications, this workstation is actually an active node on a digital network. It handles its owner's schedule, travel plans, telephone management and message taking, and event-activated audio memoranda or reminders. As will become clear, the more the workstation is cognizant of its owner's activities, the greater its ability to make correct inferences about its own proper behavior in response to stimuli from the outside world.

As a telecommunications node, we base this work on a vision of point-to-point communication including simultaneous voice and data links; the latter need not be high speed. Thus, nodes are able to engage in joint activity requiring localized databases, such as scheduling meetings between the owners of separate workstations or the intelligent handling of the control signals of a voice telephone connection. When one's workstation is instructed to "*place a call to X,*" it first makes a digital connection to X's workstation to determine whether X will take the call, and, if so, at what location (*telephone number*) to connect the voice link. Similarly, a digital connection to a process on another node is used to negotiate meeting times between remote schedulers.

Our implementation of the *Conversational Desktop* has been on Sun Microsystems workstations using Internet protocol on Ethernet hardware as the data link, and ordinary analog telephone connections as the audio link. With the evolution of digital telephone exchanges and the provision for service protocols such as ISDN, it is reasonable to assume that the voice/data channels will be available in an integrated telephone system in the not too distant future.

Each workstation is equipped with a variety of speech peripherals, including recognition, synthesis, and digital record/playback hardware. The devices actually used are configurable at run-time and the system will run, with reduced capability, with only a subset of them. The main focus of this work, however, is the synergy of these speech technologies, particularly in a context which exploits voice in a range of interrelated and interconnected tasks.

The workstation is designed to be driven entirely by voice, engaging its owner in a conversation interleaved with transactions with remote nodes. The repertoire of available operations at the time of this writing includes: scheduling meetings with individuals or groups, placing outgoing calls, taking incoming voice messages, and recording voice memos related to the above activities. The incoming message taking is based on a conversational *answering machine* described previously as Phone Slave.

### Conversational Aspects

The Desktop is inherently conversational; dialog is used both as a steady flow of feedback as well as a means of resolving ambiguities or errors on the part of the speech recognizer.

The output from a speech recognizer tends to be very noisy, characterized by a combination of insertion, substitution, and undetected word errors. It is necessary to build a *robust* parser to scan the output from the recognizer and build up a data structure describing knowledge of the input which can be used to generate dialog. Standard parsing techniques from the natural language processing community [Winograd 83] are generally inadequate, as they are based on the assumption of correct input (usually typed) in the first place.

The solution employed is a context free grammar and parser based on the Unix YACC (Yet Another Compiler-Compiler) parser generator. Each token is an instance of a syntactic class such as 'command-which-requires-a-date-and-time.' The parser takes output from the recognizer and runs all ordered substrings through YACC, calculating a score at each node of the grammar, skipping those which can be pruned in comparison with the current high score. For example, the string ABC would be parsed for ABC, AB-, -BC, A-C, A--, -B-, --C.

The scoring metric reflects knowledge of the types of recognizer error; in connected speech, errors often come in bursts, as the result of incorrect segmentation decisions. It gives points for number of words recognized, with bonuses for complete sentences and adjacent correct words.

The dialog which ensues is an attempt by the machine to fill in gaps in the parse tree based on the parse with the highest score. Phrasing of the questions is critical for several reasons. The dialog employs echoing techniques [Hayes 83] to implicitly confirm prior communication. For example, "*Schedule a meeting with Walter at <mumble>*" would trigger a query of "*When do you wish to meet with Walter?*". These questions are geared toward eliciting single word responses wherever possible, as they are much more likely to be recognized.

Another aspect of the conversational ability of the system is its method of taking incoming phone messages. Callers are greeted by a recorded voice which asks a series of questions, recording each one, while an adaptive pause detection algorithm triggers the next query. The answers to questions such as "*Who's calling please?*", "*What's this in reference to?*"

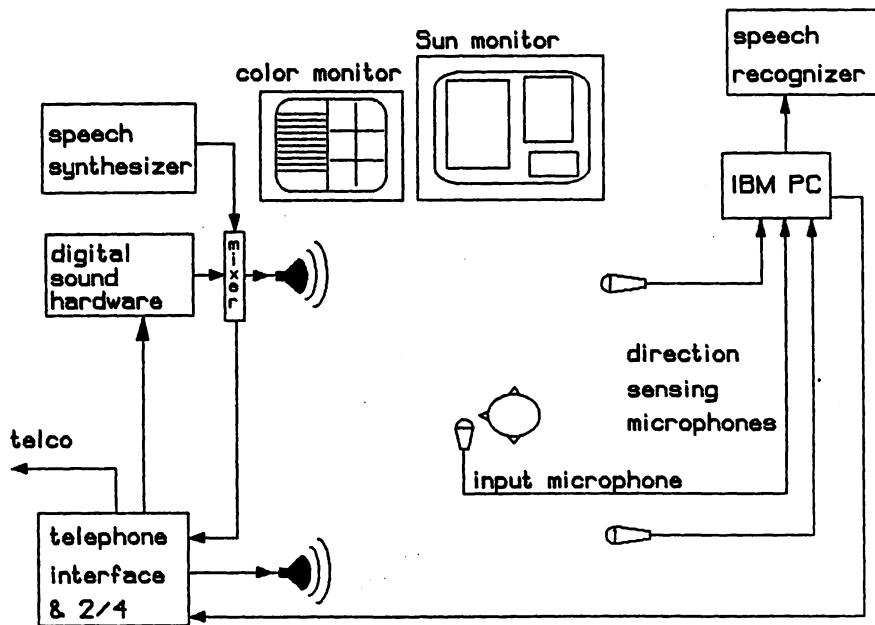


and "At what number can you be reached" are recorded into individual sound files. This sequence of recordings provides a context or handle on the content of the audio data. Even without recognizing any of the words in the answer to the "Who's calling please?" question, the machine knows it is appropriate to play this segment when its owner asks "Who left messages?"

## Addressability

Through a variety of cues, especially eye contact, a person in a small group can determine when the speaker is addressing them in particular. We wish to apply similar techniques to speech recognition so that the computer can determine when *it* is being addressed, as opposed to the telephone or someone else in the office.

To facilitate this, we have assigned spatial orientation to the system; the computer is assigned the direction of the owner's right, and the telephone the left. The system display, which shows calendar entries and phone message status, along with the loud speakers through which the Desktop talks, are both situated on the right side of the office. The 'telephone' is a hands-free arrangement using the head mounted microphone for input and a speaker for output to the left side of the office.



Audio paths in the Conversational Desktop, showing the spatial orientation of devices.

A pair of microphones placed behind the user determines the direction towards which the person is speaking. The microphone receiving the *minimum* signal when speech is detected in the head-mounted (recognizer) mike is in the opposite direction of the voice addressing. Microphones were placed to the rear to take advantage of the greater direction sensitivity based on the radiational characteristics of the human head [Flanagan 60]. Hardware and software running on an IBM PC communicate this information to the Sun Workstation.

The same hardware also controls noise-free ramped switches which direct audio to various

devices in the room. When the computer speaks, either by playing a recording or through the text-to-speech synthesizer, input to the recognizer is disabled to prevent spurious recognition. Recognition output is parsed only when speech is being transmitted in the direction of the recognizer. While speaking on the phone, the owner may have a private conversation with his Desktop by turning to the right; as soon as speech is detected in this direction, audio input to the telephone connection is temporarily disabled.

## Context

The direction-sensing microphones are also used to detect background noise (defined as signal present with no speech on the owner's headmounted microphone) which alerts the system to the presence of other humans in the office. This *background speech present* signal is used for a class of operations characterized by knowledge of the acoustical context of events occurring in the domain of the Desktop system.

When it is time to play an audio reminder, for example, the system first checks this signal and can *postpone* the reminder until a time when the owner is alone in his office. In general, the system follows the rule of not interrupting when the owner is engaged in some detectable activity; future work will attempt to prioritize current and alarm events. It will not, for example, interrupt the owner for a phone call during other activity; instead, it automatically takes a message.

The more the system knows about its owner's activities, the more it can take advantage of context to understand speech input and guide its activities. Several examples relate to telephone conversations. While engaged in a call, the command "*Schedule both of us a meeting*" refers to the owner plus the party on the other end of the connection; the system knows who this is, as it originated the call in the first place.

When the owner tells the Desktop "*I am going out to lunch*", it knows to automatically replace the current outgoing answering machine message with the "out to lunch" one.

In a similar vein, system activities may be triggered by external events. Audio reminders can be recorded by a sequence like "*When I talk to Barry remind me to ...*". Although the system then performs no recognition on the body of the reminder, it knows enough *about* it to automatically remind the owner, by playing the speech file, when asked to "*Place a call to Barry*". The same reminder also gets played as part of the process of accepting an incoming call from Barry. Reminders may be triggered by a telephone connection, a meeting about to occur, or a directive such as "*I am going home*."

## Future Work

Currently we are engaged in expanding the capability of the Desktop in several ways.

The first is added functionality, e.g., placing airline reservations, checking the state of the weather before the owner sets off to bicycle home, etc. There is a wide range of electronic databases which the Desktop can access over digital telephone connections to automatically update local knowledge about the state of the world and potentially alert the owner.

The second is personalization. Several aspects of this are already in place. For example,

each node's scheduler is driven by a set of owner specific preferences; one person may set up preferences to avoid morning meetings, while another node may attempt to avoid any meetings after five in the evening.

A more difficult challenge is to modulate preferences by some sense of the importance of the calling party. For example, I would never meet a student before 10 AM, but maybe would come in early for an important visitor. Similarly, some incoming calls should interrupt many of my activities, but most calls should (in my own preference) never interrupt a conversation with another person in the office.

## Acknowledgement

This work has been supported by NTT, the Nippon Telegraph and Telephone Corporation.

## References

- [Flanagan 60] Flanagan, J. L.  
Analog Measurements of Sound Radiation from the Mouth.  
*J. Acoust. Soc. Am.* 32(12), 1960.
- [Hayes 83] Hayes, P. and Reddy, R.  
Steps Toward Graceful Interaction in Spoken and Written Man-Machine  
Communications.  
*Int'l J. Man-Machine Studies* 19:231-284, 1983.
- [Schmandt 82] Schmandt, C. and Hulteen, E.  
The Intelligent Voice Interactive Interface.  
In *Human Factors in Computer Systems*. NBS/ACM, 1982.
- [Schmandt 84] Schmandt, C. and Arons, B.  
A Conversational Telephone Messaging System.  
*IEEE Trans. on Consumer Electr.* CE-30(3):xxi-xxiv, 1984.
- [Schmandt 85] Schmandt, C. and Arons, B.  
Phone Slave: A Graphical Telecommunications Interface.  
*Proc. of the Soc. for Information Display* 26(1), 1985.  
In Publication.
- [Winograd 83] Winograd, T.  
*Language as a Cognitive Process - Syntax*.  
Addison-Wesley, 1983.

**Christopher Schmandt**

Principal Research Scientist

Mr. Schmandt received his B.S. in Computer Science and his M.S. in computer graphics from MIT. He has continued his work as a Principal Research Scientist at the Architecture Machine Group, a component of the Media Laboratory. His research interests there are focused on interactive systems and human-interface issues, with emphasis on voice interaction and telecommunications.

**Barry Arons**

Research Associate

Mr. Arons received his B.S.C.E. and M.S. in computer graphics and interactive systems from MIT. His research interests include speech input/output, raster graphics, and interactive video.

**Charles Simmons**

Undergraduate Researcher

Mr. Simmons is completing his senior year at MIT. The topic of his Bachelor's thesis is the design and implementation of the speech direction sensing hardware and software.

The authors can be contacted at:

Media Laboratory

Massachusetts Institute of Technology

20 Ames Street, Room E15-327

Cambridge, MA 02139

## A CONVERSATIONAL TELEPHONE MESSAGING SYSTEM

Chris Schmandt and Barry Arons  
Architecture Machine Group  
Massachusetts Institute of Technology

The *Phone Slave* is a personal, integrated telecommunications management system, combining diverse message functions in a single user interface on a small general purpose computer. This paper will focus on the audio components of that interface (a related publication emphasizes the graphical interface [1]).

The Phone Slave is an intelligent answering machine, conversing with callers to format messages and relaying personal greetings to identified parties. Its owner can access these voice messages as well as electronic mail via speech recognition or Touch-Tones over the phone network. Access to both incoming and outgoing messages, an on-line directory, and autodial features are also provided by a touch-sensitive color monitor.

### Conversational Answering Machine

When the machine answers the phone, it quickly engages the caller in a conversation, playing back stored speech segments, while recording the caller's responses. By asking questions such as "Who's calling?" and "At what number can you be reached?", the machine formats message segments for its owner. These questions are phrased both for completeness of the message, and to enable response to queries such as "Who called?" by the owner.

**machine:** Hello, Chris' telephone speaking, who's calling please?

**caller:** This is Barry Arons.

**m:** What is this in reference to?

**c:** Papers for the CES show.

**m:** He's not available at the moment, but he left this message: <Chris' voice>I'M HAVING CHILI DOGS WITH ANDY, I SHOULD BE BACK IN A FEW MINUTES. At what number can he reach you?

**c:** 258-6681

**m:** When will you be there?

**c:** I'll be here all night.

**m:** Can I take a longer message?

**c:** We should start writing a longer paper for Transactions on Consumer Electronics.

**m:** I'll be sure he gets that. Thanks for calling. Goodbye.

**c:** Bye.

A well timed series of questions guide the caller, who may not be prepared to simply recite a message to a tape recorder. The conversation can take a number of branches to exchange the maximum amount of information between both parties.

An adaptive pause detection scheme is used to determine when the caller has finished answering each question; the 'end of answer' timeout is lengthened if a caller speaks slowly. It is important for playback that the reply to each question be short and to the point. Each response has an associated maximum length; if the caller exceeds it, the machine interrupts, in a louder voice, asks the caller to be precise and repeats the question.

### Caller Identification

The answer to the first question, "Who's calling?", is processed by a speech recognizer while recording (figure 1). If a match on the voice pattern of a frequent caller is obtained, the conversation branches, with the caller being greeted by name and playing a personal recording. A familiar caller can recover from recognition error by entering a unique identifier with Touch-Tones.

This branch of the conversation tree asks whether the caller can be reached at his usual number, informs him if his last message has been heard by the owner, and if not says "If you'd like to leave a message, I'll record it now, otherwise hang up and I'll tell him you called" (figure 2).

The machine encourages participation by providing a variety of options in message type and responding personally to all callers. Most important is the prospect of a specific message with greater content than the generic outgoing "I can't answer my phone right now." A dialog may occur through a series of calls, although the parties never connect directly.

### Message Retrieval

As messages are recorded in distinct audio segments, the machine may playback individual responses, or a series of responses to indicate who left message, or the entire content of a single message (figure 3). Local access is by a touch-sensitive display (figure 4), with remote access by speech recognition or DTMF tones.

The owner may access all message components remotely over a phone connection, leave a new personal reply for any caller, or request the time of a call or the caller's phone number from the directory. Messages are grouped such that all calls from the same person will be heard sequentially.

### Unified Electronic Mail

Electronic mail messages are integrated with voice messages, and may be viewed on the screen, or heard over the phone with a text-to-speech synthesizer. On the prototype system in use in our laboratory, it is quite common to receive both forms from the same person, and they are grouped appropriately for easy access.

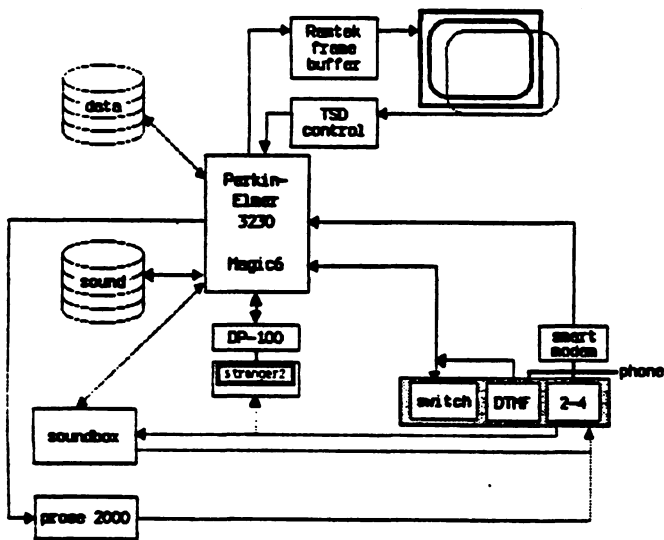


Figure 1: Hardware configuration used in Phone Slave.

As synthetic speech may be difficult to understand, a 'repeat' command replays from the previous sentence at a slower rate. As speech is slow relative to text display on a terminal, most header information is withheld unless requested. A voice reply may be made, in which case mail is sent informing the original sender that a voice message awaits, giving the phone number and an access code.

### Reference

1. Christopher Schmandt and Barry Arons. Phone Slave: A Graphical Telecommunication Interface. Digest of Technical Papers, SID International Symposium, 1984.

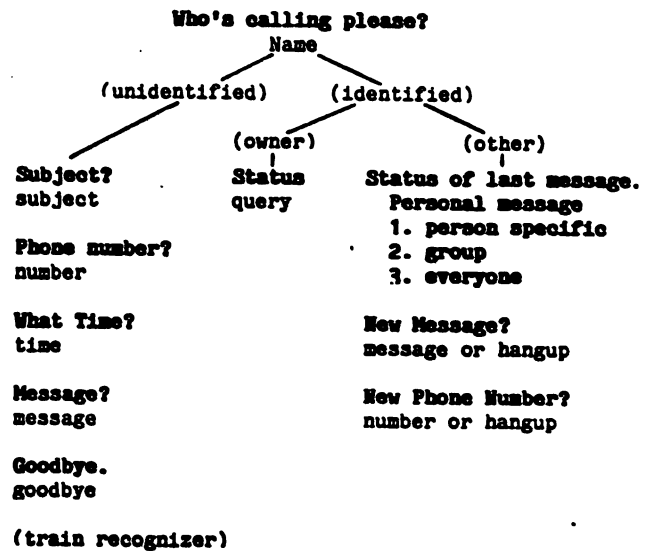


Figure 2: Tree of possible conversations.

	Name	Subject	Phone	Time	Message
Wednesday	Chris				
Wednesday		CES Paper		10:05 AM	
Monday				4:15 PM	
Tuesday	Barry	Thesis Proposal		6:50 PM	



Figure 3: Message screen with text and voice messages.

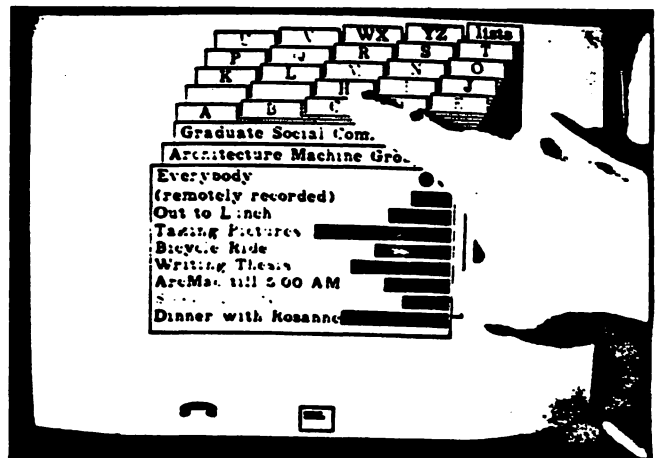
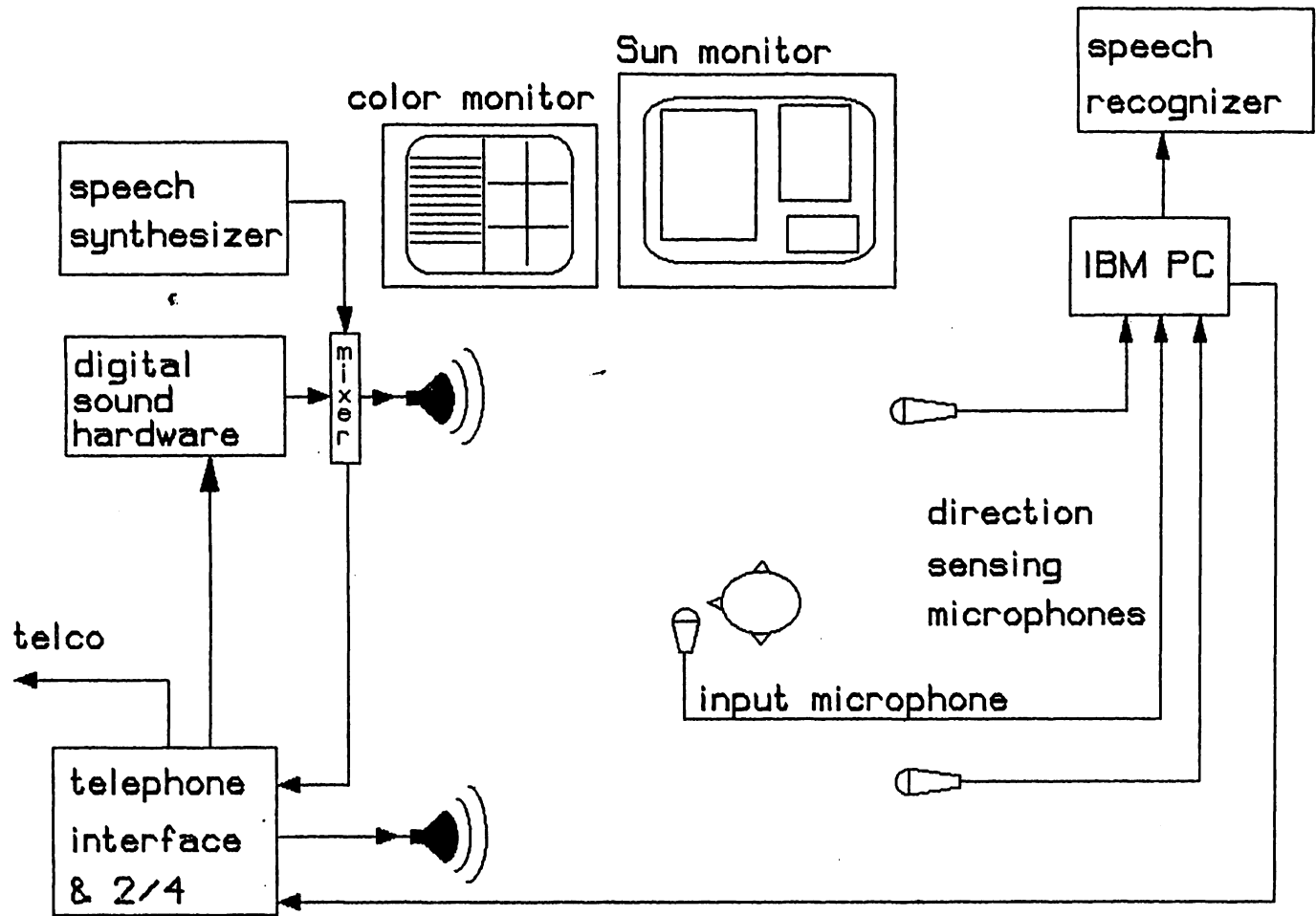


Figure 4: Touch screen access to information.





# MIT Industrial Liaison Program

## Report

Paper relevant to the symposium:

"MEDIA TECHNOLOGIES"  
October 3, 1985

"Eyes as Output"  
by  
Dr. Richard A. Bolt

1) "Conversing with Computers," by Richard A. Bolt

This paper has been duplicated at the request of the speaker.



Distributed for Internal Use  
by Member Companies Only.  
May Not be Reproduced.

© MIT

Computers are learning  
how to carry on a normal conversation  
by capturing the richness of dialogue, gestures, and  
glances. The interface between people and  
computers will be much friendlier  
as a result.

# Conversing with Computers

BY RICHARD A. BOLT

**W**HEN I see an advertisement heralding a certain computer or program as “conversational,” I just don’t believe it. To me, conversation is speaking back and forth, pointing out this or that—a lively dialogue that involves glancing about and following the other person’s glances as well as using gestures to describe, indicate, and emphasize. Whether the topic is trivial or weighty, conversation means a strong sense of another’s presence in a setting we both share.

If you have used a computer at all lately, you know why I’m skeptical about claims that they are “conversational” in this rich sense. Yet I believe that such a relationship between people and computers will come about, and in fact I have been trying to help make it happen. For the past eight years I have been working in M.I.T.’s Architecture Machine Group laboratory, founded in 1968 on the basis of Nicholas Negroponte’s optimism about the future of human-machine communication. My role is to combine a background

in computers with insights from cognitive psychology to improve the interface where people and computers meet.

The rationale for making computers truly conversational is, to borrow a phrase from Professor Negroponte, to provide “supreme usability.” This means making the computer as easy and as interesting to talk to as another person, for the novice or occasional user as well as the computer veteran. Despite vaunted claims of “user friendliness,” even today’s most advanced computer systems—in industry, homes, or the military—are too often excruciatingly difficult and frustrating to use. Of course, many research groups are actively following a variety of paths to make computers friendlier. But at the risk of seeming immodest, what sets our laboratory apart is the zest with which we are bringing many disciplines to bear on the problem.

Our goal is not simply a utilitarian one of enabling a manufacturer, for example, to produce so many more widgets per unit of time. Rather, we

PHOTOGRAPH: WALTER BENDER AND RICHARD BOLT

*Somewhere along the way we've lost the expectation that interacting with computers can be as natural as carrying on a conversation.*

---

want to enhance the quality of the human-computer interface in its own right. We are convinced that using a computer should be a pleasurable, even exhilarating, experience. Whatever else computers are, they should be fun.

This is in many ways a radical idea. Many people have long since adopted the notion that computers are for "getting the job done." They can be improved, certainly, but they will basically remain tools that you put up with for the sake of the task at hand. Somewhere along the way we've lost the expectation that interacting with computers can be as natural as . . . carrying on a conversation. We've been forced, largely by commercial vendors, to be thankful for small advances toward user friendliness. But our group believes we shouldn't mistake microsteps for real progress. Our sights are set on the root meaning of "to converse": to *keep company with*. It is this conviviality that is now so egregiously absent from our experience with computers yet so vital to our sense of ourselves.

### Put-That-There

Personal computers are now on the market that let you point at items on display, either by putting your finger on a touch-sensitive screen or by manipulating a "mouse," a small device on a cord that controls a cursor on the display screen. There are computers that let you speak to them, comparing your message with stored samples of your speech. Some such recognizers handle only discrete speech—that is, phrases spoken one word at a time. Others can handle normal speech in which words are strung together with no pauses in between.

But computers that let you use both gesture and speech are not yet offered by commercial vendors. The expressive power of speech and gesture combined clearly exceeds that of either alone—hence the importance of using both to communicate with computers. We have made a start toward this goal.

For example, several years ago I originated an exercise called "Put-That-There." The setting for Put-That-There is a special room that becomes a computer terminal you enter rather than sit at. Called the Media Room, it is about the size of a personal office: eight feet high, ten feet wide, and roughly thirteen feet long. The front wall is entirely display screen on which images can be created by a projector located behind the screen (see page 39).

The user sits at the center of the room wearing a microphone wired to an automatic speech recognizer. The recognizer is of the "connected-speech" type and has a vocabulary of 120 words. The user also wears a wristwatch band to which a small plastic cube is attached. This cube works in concert with a similar but larger cube mounted on a pedestal close to the chair. Both cubes generate magnetic fields when activated. From the relative position of the two magnetic fields, an associated computer calculates the position and orientation in space of the smaller cube, making it effectively a wrist-borne "pointer."

In Put-That-There, you create objects on the display screen simply by talking and pointing. You can name them, change their color and shape, move them about, and delete them. For example, you might say, "Create a large green circle . . . there"—while pointing with your raised arm at some spot on the screen. The system responds by putting up a large green circle at the spot you indicated. Next you might change the color of the circle by pointing at it and saying, "Make that red."

Instead of manipulating simple shapes on a plain background, you might manage color-coded ships against a map of the Caribbean. You can say, "Put a large blue tanker (pointing) there." Or "create a yellow freighter southeast of Haiti." Then you name the ship: "Call that (pointing to the freighter) the *Flying Cloud*."

The information you give by either hand movements or speech need not be perfect. Put-That-There requires only that gestures and speech, when considered together, converge upon what the user intended. Suppose you point to some spot on the map and issue the command, "Put the green freighter there"—but you mumble a bit on the phrase "green freighter," so the match with the prerecorded phrase is only marginal. A speech recognizer would ordinarily declare it a miss. Your pointing, too, is a bit wobbly, hitting the freighter but sweeping by some other ship's image as well. In other words, the data taken separately from either mode are ambiguous. Combined, though, the evidence gains strength. "Green freighter" is a plausible interpretation of what you intended to say, as you did in fact point toward the freighter, though you hit other items as well. The system, able to weigh the evidence from both modes, draws the reasonable conclusion that you probably did intend to move the green freighter.

This ability to use pronouns and adverbs instead



of names and descriptions—and to point at the objects—reduces the usual problems of automatic speech recognition. It's like being able to identify one clown out of a crowd of them, not by saying "the tall clown with the purple nose and the green jacket," but by pointing and exclaiming "him!"

Vendors of automatic speech recognizers typically claim that their products are at least 99 percent accurate. But such levels of accuracy are achieved only under optimal testing conditions, including low noise levels, trained speakers, and specially selected vocabularies. Performance under more realistic conditions may reach only about 65 percent. Obviously what matters is the *effective* level of communication—not the narrow issue of speech recognition but the broader and more fruitful one of speech interpretation.

The Put-That-There system takes the situation into account when interpreting speech. Suppose you say, "Create a blue freighter there," and the speech recognizer misses the command "create." On the basis of syntax alone, the missing word could be "create" or it could be "move." The system resolves that uncertainty by knowing whether a blue freighter already exists. If one doesn't, the system properly infers that the missed word must be a create command, not a move command, and so it puts a blue freighter at the indicated spot.

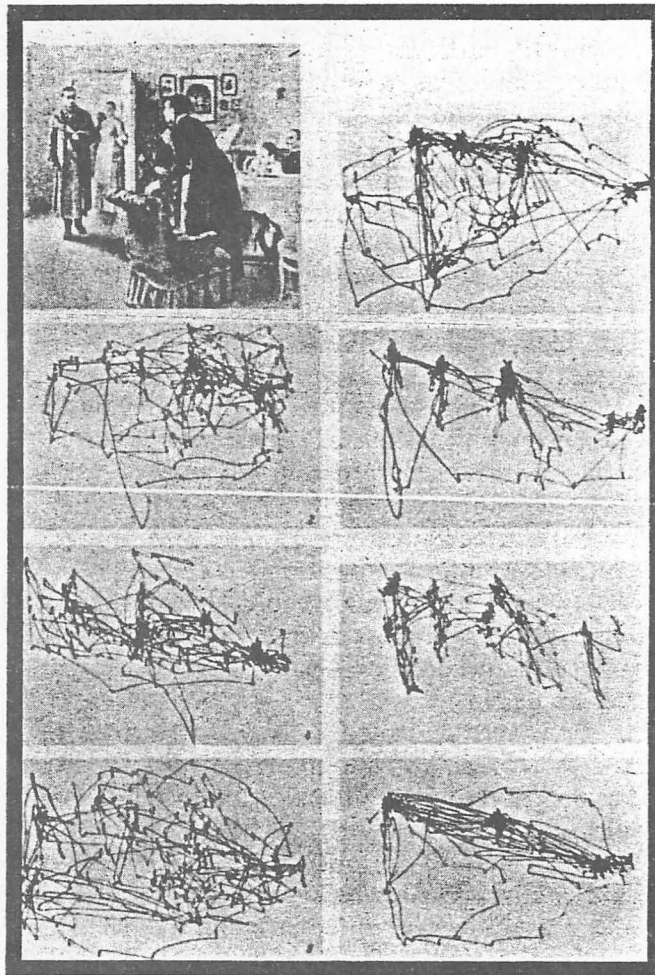
Chris Schmandt and Eric Hulteen, who programmed Put-That-There, did their utmost to avoid making the user repeat words unnecessarily. Suppose the user says, "Move *Flying Cloud* northwest of Haiti." If the word "northwest" is missed, the system simply asks, in its synthesized voice, "What direction?" It does not demand that the entire command be repeated. Computer visionary Alan Kay, in an interview in *Psychology Today*, characterized interacting with the Put-That-There system as being "like dealing with a friendly, slightly deaf butler. . . . From the standpoint of your expectations, you are willing to deal with it." Indeed, our aim is to make the user feel that despite inevitable lapses in word recognition, the system is doing its best to understand the user's intent. Achieving such user confidence is a big factor in any system's acceptance.

### Telling Glances

Developmental psychologists tell us that one of the first things a child learns is to follow the mother's

A computer can determine a user's interests by tracking eye movement. In his classic studies, Alfred Yarbus asked people to examine repeatedly a painting (upper left). Before each trial, he gave the observer

ward. Here, the observer's "looking patterns" differed markedly with each question. (From *Eye Movements and Vision*, Alfred Yarbus, 1967. Used with permission of Plenum Publishing Corp.)



line of gaze as she speaks about things. For example, the child hears the sound "kitty" while following the mother's gaze to that lively, furry creature. Similarly, we can tell where someone's visual attention is directed by watching the eyes. By contrast, we cannot directly determine "where" a person is listening; people's ears don't angle about like a rabbit's do. But interestingly enough, psychologists have found evidence for a link between eye position and auditory attention: we tend to listen in the direction we are looking.

This conversational strategy of watching where the other person is looking—as a way of tapping into what that person is paying attention to—can help a computer. By monitoring which part of the display screen the user is looking at, the machine can better understand what the user is interested in and is trying to communicate.

In a complementary sense, the computer's graphic

*I envision a "self-disclosing" computer system that need not tell you how to operate it or what services it can provide.*

---

display externalizes what the computer is offering to talk about—putting it “out there,” visibly, for the user to react to. This tends to focus the user’s speech upon what is “on the table” for discussion, which in turn greatly simplifies speech recognition by putting implicit limits upon the range of vocabulary likely to be used. At the same time, the display provokes the user’s looking, pointing, and speaking: “What’s this? Where does that come from?”

In his influential book *The Nature of Managerial Work*, Henry Mintzberg of McGill University characterizes the managerial world of the modern executive as one of brevity, variety, and fragmentation. Executives spend little time on any one activity and must deal with a great number and variety of problems in the course of a day. They must be in up-to-the-minute touch with rapidly changing situations yet not succumb to “information overload.” We can mimic electronically this salvo of events that confront executives—events demanding various degrees of attention and decision—to see how they might be better managed.

In the exercise called “Gaze-Orchestrated Dynamic Windows,” we simulate the spirit of such a volatile situation in graphics on the Media Room’s display screen. We do this by creating a composite of many television episodes—a collection of up to 40 moving images playing simultaneously. Some episodes appear just as others are ending. The stereo soundtracks of all the episodes are also piped in, creating a kind of “cocktail party” mélange of voices and sounds. We then exploit the selective visual attention of our hypothetical manager to orchestrate this complex, dynamic display—that is, we let the user’s eyes help sort out what he or she wants.

The user sitting in the Media Room wears glasses that have a miniature eye-tracking system mounted in the frame. The device, developed by the Denver Research Institute, shines a tiny beam of infrared light onto the cornea and traces the reflection. This provides a constant measure of the position of the wearer’s eye with respect to the glasses. We also mount on the frame a small location-sensing cube—the same kind worn on the wrist in Put-That-There—to detect the position of the frame within the room. The measurements of eye position in the glasses and frame position in the room combine to reveal where the user is looking on the display screen.

Whenever the system finds that the user is looking steadily at a particular image, it turns off the sound-

tracks of all other episodes. If the user persists in looking at this episode for a few seconds or so, the system “zooms in” to fill the screen with that image. To recover the many deleted images, the user simply moves a joystick mounted on the arm of the chair. The net effect is to allow users to filter out all but the information of immediate interest, while enabling them to return at will to the more complex environment. We think this offers a way to exploit people’s natural processes of selective visual attention to help them focus on the most relevant events in the midst of near-overwhelming complexity.

### Computers as Good Hosts

Computers that know where on a display the user is looking and can capture speech and gesture suggest the possibility of “self-disclosing” systems: computers that tell about themselves. I don’t mean simply that the computer would be “manual-free,” although that would be a blessing unto itself. The reams of printed material that accompany some personal computers sometimes seem to weigh more—both physically and intellectually—than the hardware. Rather, I envision a self-disclosing system that need not tell you how to operate it at all or what services it can provide.

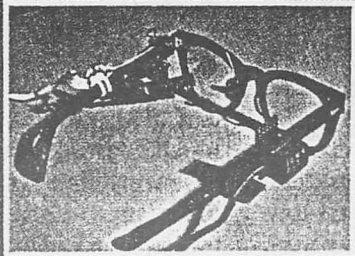
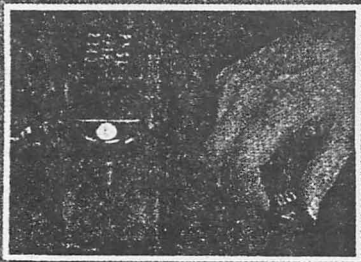
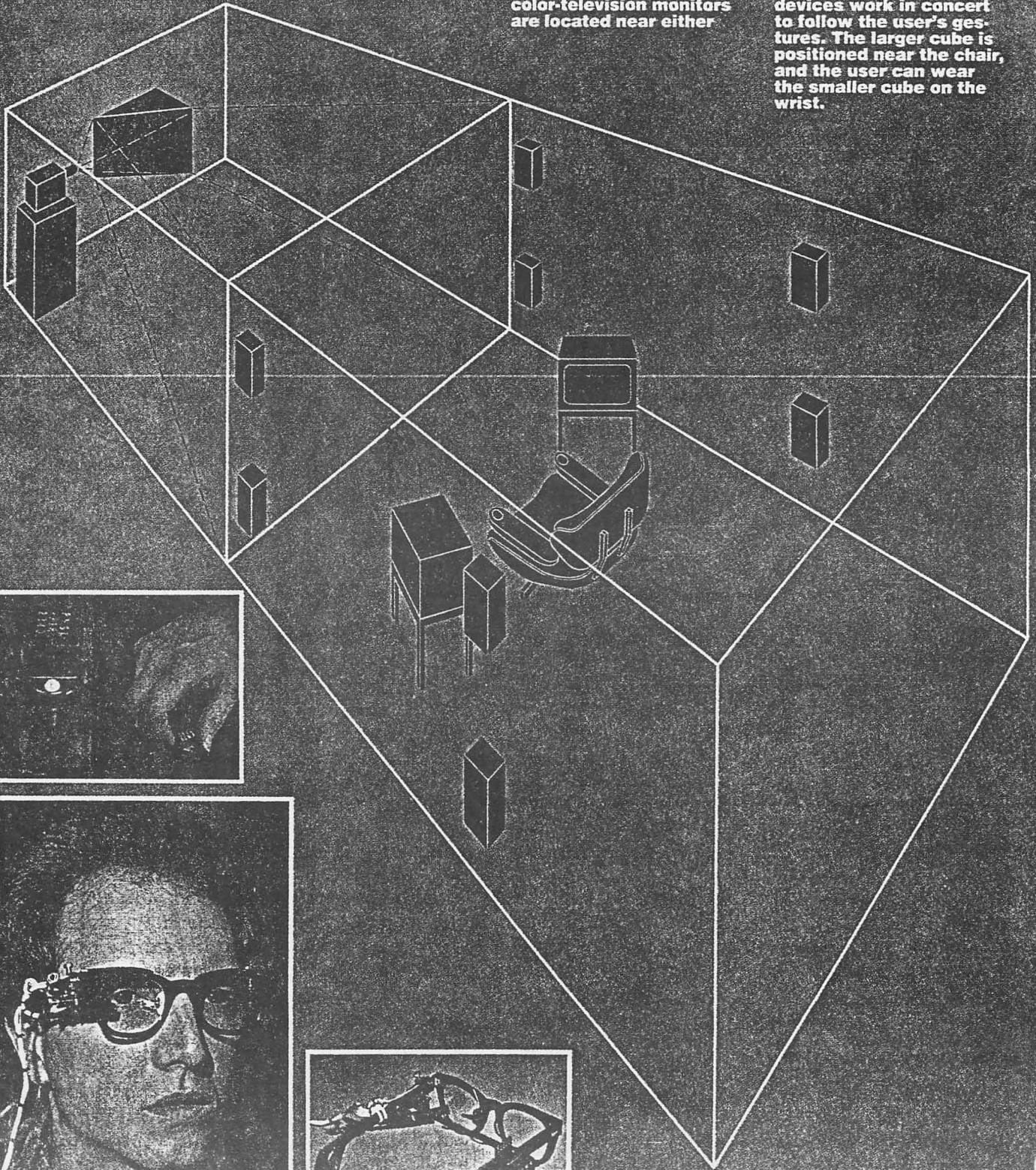
This computer would be instrumented to respond to your presence and normal behavior. It would have a full-color graphics display and eye-tracking capability for determining where on its display you are looking. You would be able to point at it and speak to it, and the computer would communicate back with synthesized or recorded speech as well as text on its display. It would disclose its contents—its information base—according to the interests you exhibit through your actions, and would do so at a speed that matches your own.

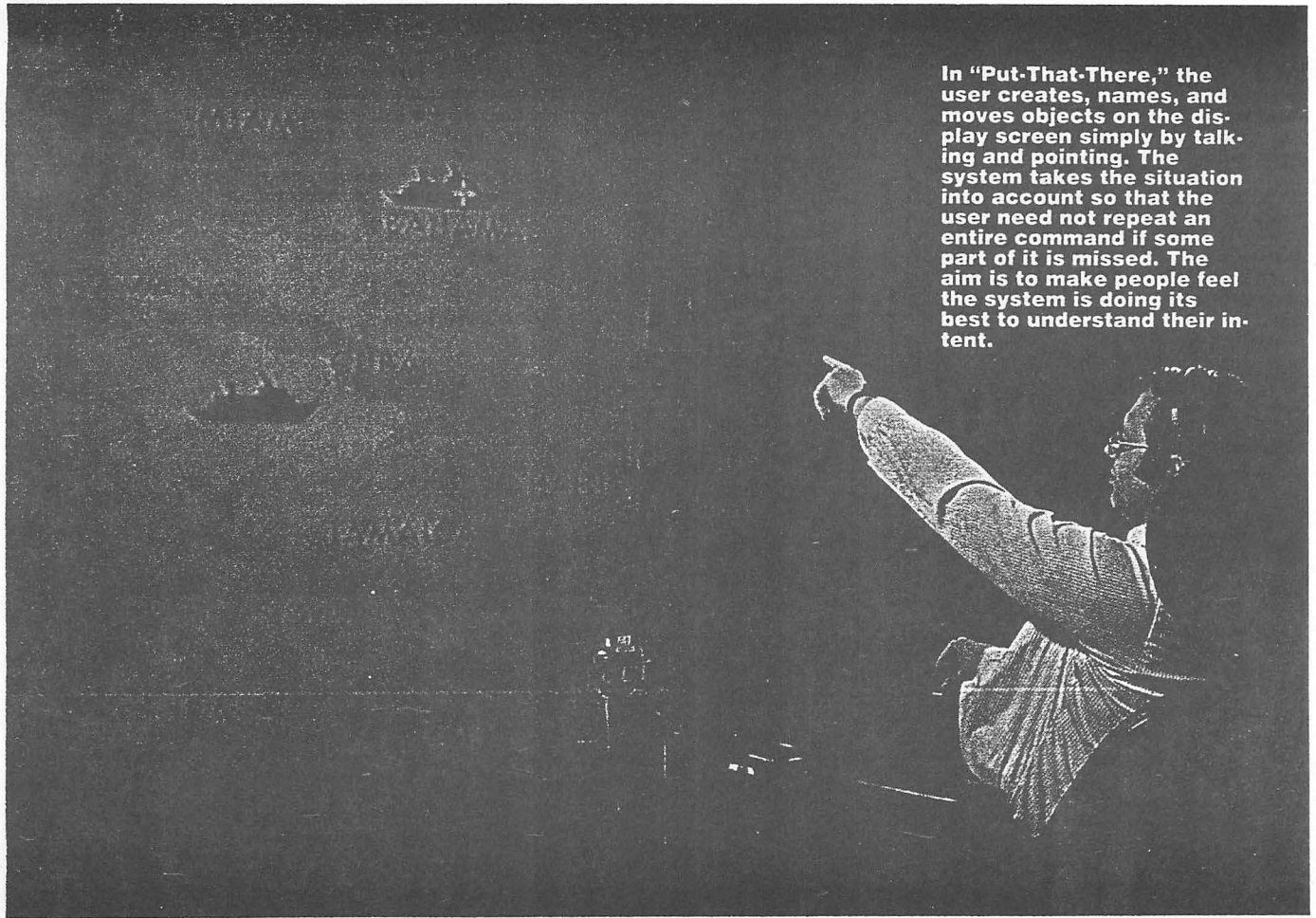
This system would interact with its user just as one person would interact with another. Suppose you breed guinea pigs and are showing someone the badges and ribbons your cavies have won, which are displayed in your trophy room. As a good host you try not to dominate things but remain alert to the cues your guest gives off: what is he looking at, what does he say, what message is his “body language” sending? Suppose the guest, surveying a wall full of ribbons, asks, “What are those?” You respond: “Those are the ones we won over the years at the Westfield Fair and Exposition.” A certain trophy



The Media Room at M.I.T.'s Architecture Machine Group laboratory serves as a computer terminal that you enter (see diagram). The user sits facing a large display screen on which images are projected from behind. Touch-sensitive color-television monitors are located near either

arm of the chair, and eight speakers provide "wraparound" sound. In some exercises, the user wears glasses equipped with an eye-tracking device that monitors where on the screen he is looking (lower left). Two "position-sensing" devices work in concert to follow the user's gestures. The larger cube is positioned near the chair, and the user can wear the smaller cube on the wrist.





In "Put-That-There," the user creates, names, and moves objects on the display screen simply by talking and pointing. The system takes the situation into account so that the user need not repeat an entire command if some part of it is missed. The aim is to make people feel the system is doing its best to understand their intent.

catches your guest's eye. You take it from the shelf and offer it for closer inspection. Your guest reads the inscription and inspects the detail. Then his eyes turn elsewhere, a cue for you to relieve him of the trophy and go on to something else. Next your guest's eyes range over a set of distinctive badges. "What are those?" he asks—the very question he uttered a moment before. But it has become a *different* question, changed by the direction in which he is looking.

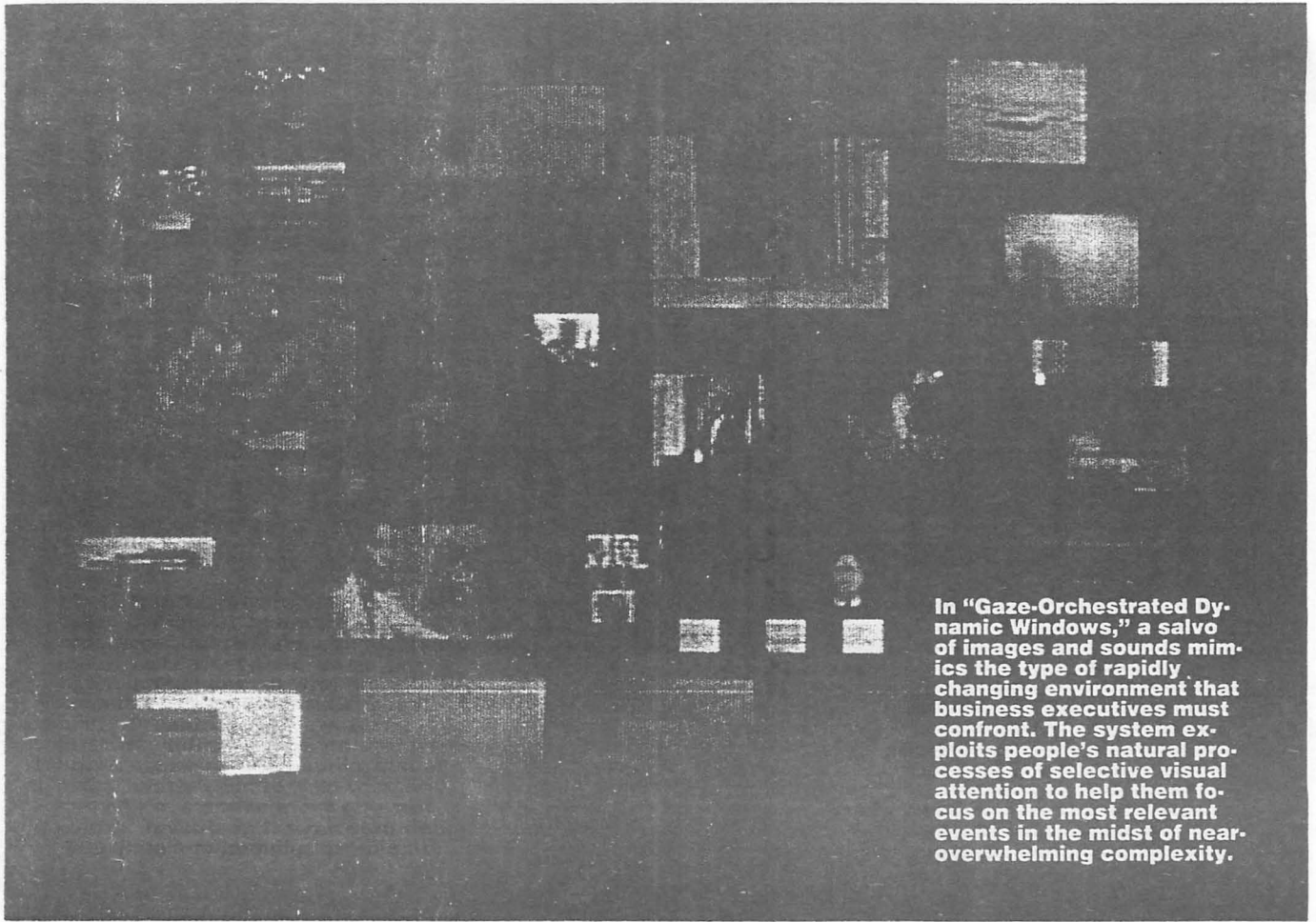
Consider now that our self-disclosing computer is showing such a collection of items. The computer's display sets the topic, just as did the real-life artifacts of the trophy room. The computer—like the human—can determine where its guest is looking, listen to his speech, detect his gestures. The system can then zoom in upon the item or area of interest to discuss it.

Such a computer can act very skillfully in deciding what and how much to tell about the subject at hand, again by taking its cues from the user's eyes. There is evidence that the eyes reveal the pattern of an observer's curiosity along very specific lines. In his classic eye-tracking studies in the Soviet Union during the 1960s, Alfred Yarbus asked people to examine a copy of the famous nineteenth-century

painting by Ilya Repin entitled "They Did Not Expect Him." The scene is of a young man just returned from political exile to the midst of his startled family. Before looking at the picture for three minutes, each observer was asked one of a number of questions: What are the ages of the people? What are the material circumstances of the family? What was the family doing before the young man arrived? The observers' "looking patterns" differed markedly depending upon their goals as set by the question.

Of course, the computer must be able to tell the difference between a protracted stare stemming from the fact that the viewer is puzzled by something and blank staring stemming from saturation. But that is not likely to be difficult: researchers studying eye movements have found that eye patterns when interest is saturated differ from those when curiosity is live. Furthermore, pupil diameter varies with tension, interest, and suspense. These cues, coupled with what viewers say, can help a machine make reasonable inferences about when to move on. The machine can also gauge the effectiveness of its presentation by checking whether viewers look in the right places when it tells them about a display. If viewers fail to look at a relevant spot, the computer can reemphasize or recast its explanation.





In "Gaze-Orchestrated Dynamic Windows," a salvo of images and sounds mimics the type of rapidly changing environment that business executives must confront. The system exploits people's natural processes of selective visual attention to help them focus on the most relevant events in the midst of near-overwhelming complexity.

## Give-and-Take

This kind of interchange between human and computer is a process of mutual self-disclosure. The computer is disclosing itself—or, more specifically, its contents—on the basis of the user's disclosing of himself or herself explicitly in speech and gesture and implicitly by eye movements. It is an ongoing two-way process.

What drives this conversation? What keeps it going and gives it direction? It is primarily the *curiosity* of the user as it interacts with the built-in *reactivity* of the system. For example, psychologists have shown that people are attracted to things that are visually rich and complex. People are also driven by so-called epistemic curiosity: we want to know the why and what of things. In its response to these "simple" levels of curiosity, the system obligingly tries to follow whatever leads its users give. The computer's "personality" is that of the convivial yet reserved person: it responds in a lively way but doesn't push itself on the user.

The system can readily sense the difference between a politely curious inquirer and a "serious" one—the difference, for example, between the tourist who strolls through Westminster Abbey taking it

all in and the scholar who studies it. Both might spend the same time looking, but their ways of observing provide a clue to how organized and comprehensive a guide's exposition must be—technical for the scholar, lucid but light for the tourist.

When a user has an even more compelling reason to communicate with the system, the computer can respond in kind. For instance, if you are a graduate student who must know all about colonial architecture in New England by final exams next Wednesday, you can instill motivation in the system by asking it, "Tell me about colonial architecture." Now the computer goes directly into a kind of superordinate "teaching" mode, in sharp contrast to its previously relaxed, reserved posture. The system may go slowly now and then to let you digest information when it senses that your rate of uptake has faltered. But it assumes more responsibility for shaping the exchange than in the more leisurely mode. All the while, whether the overall mode is relaxed or purposeful, the moment-to-moment conversational initiative drifts to and fro between human and computer, much as the "attack" passes back and forth between two fencers. Thus, the actions of the user and the system are mutually determined, either party by turn driving or being driven.

*The hardest job will be fostering a view of the computer interface as a comfortable place where people and machines truly "keep company."*

---

Users' sophistication—their style of looking and the questions they ask—can also help determine how the computer will respond. For instance, novices and expert chess players look at board positions differently. And of two people interested in antiques, one may have a "trained eye" for detecting the fine points of a collectible while the other examines the item in a less disciplined way. Such differences in looking style, though subtle, can enable a computer to infer what should be shown and said next.

### Keeping Company

In person-to-person conversation, we speak not to some disembodied spirit but to someone right before us. One of the benefits of this direct presence is that we can look the person in the eye—for example, we can shift our gaze from an object under discussion to the eyes of the person with whom we're speaking. Beyond the sense of engagement this creates, eye contact can signal that we wish to shift the discourse to a personal level: to talk "to" the person rather than "with" them about external matters. Given that conversationality is a positive value in human-computer communications, how might we establish the ability to look the machine "in the eye"?

Patrick Purcell and some of his students at M.I.T. have been experimenting with a kind of computer "persona." Next to the Media Room's display screen is a video monitor that bears the face of a person—in this case, Professor Purcell himself. When you speak to the system—for example, to command it to display the art and architecture slides stored on its videodisc—this image speaks back to you. You ask: "I'd like to see some examples of Romanesque." The face on the monitor responds with synthesized speech, "Early or late?" The lips on the image are synchronized to the phonetics of the synthesized speech in a convincingly lifelike way.

This exercise was originally done somewhat tongue-in-cheek, with the modest goal of providing a more cordial, congenial channel for output messages than the line of print or the disembodied voice. The persona doesn't—yet—track the user's eyes and thus can't tell whether he or she is paying attention. With eye-tracking powers added, the persona could become a specific spot for the user to look at to establish "eye contact" with the machine.

Many of these technologies to support friendlier and richer human-computer interactions already ex-

ist, at varying levels of refinement. Devices that recognize and synthesize speech have improved markedly in recent years. Reliable touch-sensitive screens that enable users to point at data on display are on the market. Body-sensing equipment for capturing various levels of gesture is developing apace. For example, researchers in our laboratory are working on a glove that will allow the computer to discern a user's hand and finger movements. What has not existed is the appreciation of such technologies as essential parts of a computer's instrumentation.

Eye tracking offers a good example. Unobtrusive tracking systems that can be situated as much as six feet from the user have been available for several years. But they cost about \$100,000, which prohibits them from serious consideration as system components. Part of the high cost is due to the proverbial chicken-and-egg dilemma. The systems are currently sold one at a time to research laboratories as measurement tools, not in large quantities to become part of well-equipped computer systems. Growing demand and high-volume production would help lower costs. But the fact that eye tracking can become a system component has not yet really struck home. There is a vague sense that tracking may be useful, but no real conviction.

This situation could change with new developments in digital cameras, microprocessors, and tracking technology. As tracking systems become more compact, they will become embedded in terminals as an integral part of new systems. In the long run, the price of a tabletop or personal computer will less and less reflect the memory and processing elements. Rather, it will reflect the costs of the electromechanical accoutrements that capture the user's intentions.

More difficult than developing the hardware for capturing multiple human modes of conversation, however, will be creating the machine intelligence to interpret human outputs and map appropriate responses. This effort will involve computer science, psychology, linguistics, artificial intelligence, and cognitive science—with scientists from all disciplines contributing to the necessary insights and inventions.

But hardest of all will be fostering a view of the computer interface as a comfortable place where people and machines truly "keep company." The world of computers outside the home—where the largest number of computer consumers still resides—contains two classes of users. There are those people

# MATH/PROTRAN<sup>TM</sup>

## IMSL's Natural Resource for Mathematical Problem Solving

who actually operate computers, and there are those who decide whether to adopt a particular kind of computer. The latter users are usually indifferent to the subjective experience of those who actually interact with machines.

Most organizations operate according to the ethos that those lower in the hierarchy should not be too comfortable, and that the tools provided for them should be "cost-effective." If one tool is a computer that is excruciating to use, so much the worse for the employee. Unless they are computer buffs, people higher up in an organization usually avoid computers and hence do not understand the feelings of frustration and lack of control they can engender.

The world of home computers is different. People who buy those machines are normally the users, and they put a high stake on their subjective experience at the interface. For them, the flavor of "keeping company" is the bottom line. Yet the loudly trumpeted home-computer revolution seems to have stalled. People perceive word-processing and spreadsheets as helpful and games as entertaining babysitters for the kids, and that's about all. Only when home computers provide a real help—and good company—will the true revolution begin. Then, inexorably, people in the workplace will come to expect and even demand the kind of conviviality they are accustomed to at home. That vision is seductive.

*RICHARD A. BOLT is a principal research scientist in M.I.T.'s Media Laboratory, where he is acting head of its Human-Machine Interface Group. He is the author of The Human Interface, (Lifetime Learning Publications, a subsidiary of Van Nostrand Reinhold, 1984).*

**M**athematical problem solving can be involved and time consuming, but it doesn't have to be. MATH/PROTRAN, one of IMSL's Natural Resources, is a powerful system for the professional who expects a straightforward approach to problem solving.

You don't need any programming knowledge to use this remarkable system. In a surprisingly short time, MATH/PROTRAN is at your command. Convenient "help" files provide on-line reference, and the system automatically checks your statements for errors.

MATH/PROTRAN lets you define problems naturally, in a few simple statements — and gives you effective solutions to problems involving interpolation and data smoothing; integration and differentiation; eigenvalues and eigenvectors; differential, linear and non-linear equations; as well as other mathematical procedures.

If you're currently solving problems using FORTRAN, you'll appreciate the ability to combine FORTRAN and PROTRAN statements for tailored problem solving. This added measure of flexibility sets MATH/PROTRAN apart from other systems of its kind.

MATH/PROTRAN is a member of the PROTRAN family of problem-solving systems for statistics, linear programming and mathematics. These systems use accurate, reliable numerical techniques to give you the consistently dependable results you have come to expect from IMSL, a world leader in affordable technical software.

MATH/PROTRAN is the natural resource for a wide variety of mathematical applications. And the low subscription rate makes this powerful system extremely affordable, even if only one person in your organization uses it.

To find out more about MATH/PROTRAN, return this coupon to: IMSL, NBC Building, 7500 Bellaire Boulevard, Houston, Texas 77036, USA. In the US call toll-free, 1-800-222-IMSL. Outside the US and in Texas, call (713) 772-1927. Telex: 791923 IMSL INC HOU.

Please send complete technical information about MATH/PROTRAN.

Name \_\_\_\_\_

Dept. \_\_\_\_\_ Title \_\_\_\_\_

Organization \_\_\_\_\_

Address \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

Area Code / Phone \_\_\_\_\_

Computer Type \_\_\_\_\_ TECH5

The IMSL PROTRAN problem-solving systems are compatible with most Control Data, Data General, Digital Equipment and IBM computer environments. Not yet available for microcomputers.

# IMSL

Problem-Solving Software Systems

Copyright © 1984 IMSL, Inc., Houston, Texas

102687593

# MIT Industrial Liaison Program

## Report

Paper relevant to the symposium:

"MEDIA TECHNOLOGIES"  
October 3, 1985

"Realistic Computer Animation"  
by  
Professor David Zeltzer

- 1) "Towards an Integrated View of 3-D Computer Animation," by  
David Zeltzer

This paper has been duplicated at the request of the speaker.



Distributed for Internal Use  
by Member Companies Only.  
May Not be Reproduced.

© MIT



# TOWARDS AN INTEGRATED VIEW OF 3-D COMPUTER ANIMATION

David Zeltzer

Computer Graphics and Animation Group  
The Media Laboratory  
Massachusetts Institute of Technology  
Cambridge, Massachusetts 02139

## ABSTRACT

To automate character animation and extend it to 3-D we need to create and manipulate three-dimensional models of articulated figures as well as the worlds they will "inhabit". *Abstraction* and *adaptive motion* are key mechanisms for dealing with the *degrees of freedom* problem, which refers to the sheer volume of control information necessary for coordinating the motion of an articulated figure when the number of links is large. A three level hierarchy of control modes for animation is proposed: *guiding*, *animator-level*, and *task-level* systems. Guiding is best suited for specifying fine details but unsuited for controlling complex motion. Animator-level programming is powerful but difficult. Task-level systems give us facile control over complex motions and tasks by trading off explicit control over the details of motion. The integration of the three control levels is discussed.

**KEYWORDS:** computer animation, simulation, human-machine interface.

### 1. Animation as Simulation

Currently there is much controversy about the nature of 3-D computer animation. Should such systems be based on simulation, keyframing, or an animation programming language? Should the interface be through graphical input devices, or through the keyboard? It is my purpose here to provide a conceptual framework for 3-D computer animation in general, and character animation in particular.

Automatic inbetweening has been the focus of attention in moving from conventional to computer-assisted 2-D animation[1,2]. From this point of view, it seems a natural extension to apply automatic inbetweening to 3-D animation, and in fact, a number of such systems have been developed[3,4,5]. Here I will argue that 3-D keyframing belongs to the first level of a three level hierarchy of control regimes for animating articulated figures.

In order to produce convincing character animation, conventional 2-D animators refer constantly to living models, or study motion pictures of living models, or draw directly from still frames of such motion pictures

(rotoscoping). That is, character animation is not a process of transforming lines and shapes on 2-D surface, nor is it simply the art of squashing, stretching, or otherwise exaggerating and caricaturing motion. Thomas and Johnston[6] make quite clear that the great success of the Disney animators was due in large part to the long hours they devoted to studying and observing the movements of humans and animals in preparing for a particular sequence. A character was successful precisely in proportion to how well the animator understood the kinematics of the figure, the structure and timing of a movement, and the effects of a movement on soft tissue and clothing. Once these elements were mastered, only then could the animator develop a character's personality by the judicious exaggeration or de-emphasis of particular attributes.

As long as animation requires the generation of many drawings by hand, simplicity and economy will be essential elements. 3-D computer animation, however, is an entirely different medium. The animator's energy is no longer invested in drawing and the tedium of

---

From *Proc. Graphics Interface 85*, Montreal, May 1985. Reprinted, with revisions, in *The Visual Computer*, Springer Verlag, to appear.

inbetweening. Instead, the focus is on creating an environment -- designing *microworlds* and populating them with interesting characters. Since frames are no longer generated by hand, rather than expecting simplified and stylized imagery, we assess computer animation in large part by noting how convincing is the simulation of three-dimensionality, lighting, background scenery, and surface texture. Animated images can take on added complexity and detail; they can be made to look more realistic, if that is desired, or more convincing in their other-worldliness. This applies equally to the *behavior* of figures and objects as well as their physical appearance.

3-D computer animation is thus a process of simulation in its most general sense: the specification of objects and transformations on objects. The notion of the computer as a simulation medium is not new; Turing showed after all, that a computer could simulate itself or any other -- a theme echoed often by Alan Kay [7, 8] The work of early graphics researchers was aimed at apprehending the *visual* complexity of the world by learning to simulate the effect of light, shade, texture and so on. Current synthetic imagery has achieved near photographic realism for certain classes of geometric- and stochastically tractable objects and environments. Now the task is to try and apprehend the *procedural* complexity of the world, which, as we shall see, requires computational models of many highly complex or only partly understood processes, including the dynamics and inverse kinematics of articulated motion, and commonsense planning and problem solving.

How can we specify and coordinate the behavior of the objects we wish to animate? There are three basic approaches.

- (1) We can explicitly describe the behaviors we are interested in. This is the *guiding* mode.
- (2) We can describe behaviors algorithmically, in some programming notation. This is the *animator level*.
- (3) Lastly, we can describe behavior implicitly, in terms of events and relationships. This is the *task level*.

For our purposes we can consider the domain of 3-D computer character animation to

be the control and coordination of the motion of articulated structures made up of rigid links. Differential scaling and other shape transformations are important but secondary to the motor control problem. In the next section I examine the fundamental problem we face in trying to coordinate the motion of articulated figures, and we will look at mechanisms for dealing with the complexities of figure animation. Later we will see that the graded implementation of these mechanisms gives the above three-tiered hierarchy of control modes.

## 2. The Degrees of Freedom Problem

The essential problem of coordinating the motion of an articulated figure is to generate appropriate values of the joint variables that control the position and orientation of each link. Joints may be modeled as *lower pairs*[9], such as rotary or sliding joints, or they may be more complex, as in a detailed model of the human knee. For a figure with  $n$  joints, we can think of an  $n$ -dimensional *pose space*, where we assign a coordinate axis to each of  $n$  degrees of freedom; and an  $n$ -component *pose vector*, which completely specifies a particular configuration. To animate the motion of a jointed figure, a pose vector must be specified for each frame of the sequence. To animate a minute of complex motion for a reasonably detailed figure, say, with 30 links, tens of thousands of values will need to be specified for the joint variables to display a new configuration each frame. Even if the sequence is keyframed, with a keyframe every two seconds, 30 pose vectors -- nearly a thousand values -- need to be specified. This is an example of the *degrees of freedom* (DOF) problem[10], which refers to the sheer volume of control information necessary for coordinating the motion of an articulated figure when the number of links is large, as in a human figure. It is the reason why animators find 3-D character animation so tedious.

Of course we are not interested in random motion; the movements of a figure must be "correct" in some sense to be of any use -- the robot programmer may wish to optimize energy expenditure, for example, and the animator will want the figure to move in some expressive manner. Viewed in this way, character animation is a problem of *search*. Not only do we have to generate pose vectors, we need to find a

*particular* set of variables out of an immense pose space -- if each of 30 joints has only 2 possible positions, there are over a billion potential configurations!

To complicate the problem, many figures of interest are kinematically *redundant*, possessing "extra" degrees of freedom that allow multiple solutions -- perhaps an infinity of pose vectors -- all of which satisfy a particular movement problem. The human arm, for example is redundant -- you can reach for an object with the elbow held high, low, or in between. That is, there are many arm configurations that will position the hand at some fixed location in space. This redundancy gives us the extra flexibility we need to reach around and over objects, and, in general, to maneuver in a cluttered environment. And it is why individuals can develop characteristic and expressive "styles" of movement.

For animation, of course, we are not interested in a single configuration, but a sequence of pose vectors -- a "hyper-path" through a many-dimensional pose space. So it is not surprising that even when the main features of a motion are known, it may take an animator many iterations to get the movement "just right".

In the next sections we look at two important techniques for dealing with the degrees of freedom problem.

### 3. Adaptive Motion

By adaptive motion I mean the ability of a figure controller to use information about the environment and the figure itself in the control process. That is, feedback can be used to guide the search through the huge space of potential configurations. To do this, at least the location and orientation of objects and their surfaces must be available to the animation software, and not just to the rendering programs, as is usually the case. Physical interactions between figures and objects are so ubiquitous in the real world -- touching, grasping, pushing, not to mention locomotion over a wide variety of surfaces -- that automatic collision detection, long of interest in CAD/CAM and robotics, should become an integral part of the animation environment.

Adaptive motion makes possible goal-directed and constrained behavior, since it allows the animator to describe movement in terms of relations among objects and figures. It lends generality to animation sequences, since animation software can adjust motion sequences for different scenes. This helps to hide unnecessary detail from the animator, since the burden of generating much of the control information can be left to the animation software.

Reynolds [11] has suggested that it would be desirable if an animator could establish "rules of behavior" for objects and characters in some imagined microworld. After establishing the initial conditions of this simulated universe, the animator would sit back and let the animation system generate the sequence. This amounts to a 3-D, computerized extension of the *straight ahead* style of 2-D animation[6]. Adaptive motion makes possible the extension of this technique to 3-D computer animation.

## 4. Abstraction

The importance of abstraction in dealing with the intellectual complexity of computer programming is well-known[12], and it is a basic tool for dealing with the kinematic and behavioral complexities of articulated motion as well.

There are five kinds of abstraction useful for controlling character animation: structural, procedural, functional, character and world modeling.

### 4.1. Structural Abstraction

A structural abstraction describes the kinematic properties of a figure, i.e., the transformation hierarchy, the nature of the allowable joint motions, and whether links are rigid or non-rigid (although we will deal only with rigid motion here). The notion of a transformation hierarchy is a generalization to 3-D of the familiar 2-D instancing systems described in graphics texts. Most 3-D animation systems provide some means of representing transformation hierarchies, e.g., Crow's *scn\_assembler*[13], Blinn's *artic*[14], and Reynolds' ASAS[11]. In these systems, joint transformations are represented as simple rotations and translations, sometimes including scaling, although more general representations for articulated motion, e.g., Denavit-

Hartenburg (D-H) notation[9, 15], have long been used in the field of mechanism design, and more recently, robotics[16]. *sdl*, the skeleton description language, is the tool for specifying structural abstractions for use in *sa*, an articulated motion system described in[17]. A similar tool, *mat*, is in use at the New York Institute of Technology[18].

At the Media Laboratory we are developing a set of graphical tools, implemented on a Symbolics 3600 Lisp machine, for designing and editing standardized kinematic descriptions of jointed figures. These descriptions have been expanded to include D-H representation for prismatic and rotary joints; the representation is extensible such that other joint types may easily be added. Such a standardized representation is essential in a distributed computing environment in which many programmers and animators, working on a variety of machines, wish to share graphical data and software. In addition, we can readily integrate robotics control algorithms into the animation software, since D-H notation provides a convenient and powerful representation for such motor control techniques.

#### 4.2. Procedural Abstraction

A procedural abstraction [19] is the representation of a movement algorithm *independent* of the structure of the figure it controls.

For example, the DOF problem is not so severe in the case of a robot manipulator with six or seven joints. Even so, humans are not good at calculating the necessary joint angles for controlling even a simple manipulator, and *resolved motion*[20, 21] is generally used. That is, the position and orientation of a target location are input, and the manipulator controller automatically computes the pose vector necessary to reach it.

Resolved motion control is an important example of the use of procedural abstraction in the solution of the DOF problem: a computation is specified that will transform the input parameters, i.e., the position and orientation of the target, into the output object, here the set of joint angles that will position and orient the end effector at the desired location in the workspace, if possible. Resolved motion control is independent of a particular kinematic structure, and can

be applied to figures of 6, 8 or more links, and for, say, a human figure, can be applied equally well to control either arms or legs[22, 23]. Other examples of procedural abstraction are the computation of trajectories for falling objects, the computation of the paths of colliding objects, or the use of spline curves for generating smooth motion. Such facilities are often provided for the animator ready-made, but may be constructed by the animator in animation systems embedded in high-level programming languages.

#### 4.3. Functional Abstraction

For the robot arm the number of links is small and the arm is treated as a single kinematic entity. But for a figure with many links, we want to be able to group together both the structural elements and the procedures that are necessary to effect a particular class of motions. Alternatively, we can impose constraints on the movements of a set of joints. We call such a grouping a *functional* abstraction. Functional abstractions are important because they allow the animator to *factor* the pose space into motor skills. If we already know the general "shape" of a motion, we need only consider a subregion of the total pose space. Say we want a figure's hand to grasp an object -- we already know which joints need to move, roughly how they should move, and moreover, we know this is a useful motion that we want to repeat often. We can cluster this group of joint movements around the task "to grasp", and attach one or more procedures to implement it (perhaps resolved motion). Once this motor skill has been defined, the details of its execution can be suppressed. That is, we need only supply the appropriate parameters, e.g., target location, fast or slow, hard or soft, to the motor program for the grasping skill. By specifying functional abstractions for grasping and other tasks, the animator is spared the burden of generating pose vectors and can instead think of the figure motion at a higher level -- in terms of the tasks and events that are to be performed.

Functional abstractions allow us to attach implicit *goals* to figure motion. By decomposing a figure's potential movements into a repertoire of skills we can associate the events and relationships the animator specifies with the skills (implemented as functional abstractions)

that the figure controller "knows" about. Moreover, if we allow functional abstractions to refer to other functional abstractions, it is possible to construct *behaviors* as compositions of simpler movements.

#### 4.4. Character Abstraction and World Modeling

In the physical world objects and figures interact in complex ways at many levels of detail. Adaptive motion requires at least efficient geometric representations for collision testing and path planning; goal-directed animation control requires in addition sophisticated mechanisms for knowledge representation.

Part of the problem involves structuring high-density graphical data bases to avoid exhaustive searches through long lists of surface elements to do, say, collision testing. Rather, we want to consider only those objects in the scene that are "near" to the figure. This means that the data base must be carefully organized spatially so that searches always proceed at the appropriate level of detail. Various hierarchical methods of structuring data to speed up occlusion testing have been reported, e.g., [24, 25, 26]. Franklin [27] describes a set of algorithms which are useful for intersection testing as well.

But the larger problem is to represent attributes, functionality, and relationships of objects in a scene so that we can simulate the mechanical behaviors and interactions of objects in general. We want this representation to be uniform such that there is no distinction between agents and objects. That is, while humans are in some sense active agents they also obey the laws of Newtonian mechanics; a person falls just like a rock when pushed off a cliff. On the other hand, an animator may want the chairs and tables to dance around the room when the villain leaves. It should be easy to ascribe such behaviors to otherwise inanimate objects.

In order to do this, we need to specify three things about any object: what it is, how it's put together, and how it behaves. The problem of representing physical objects is an active area of research in artificial intelligence [28]. Briefly, objects can be described in terms of a generalization hierarchy, such that instances of particular things appear as

specializations of more general classes of objects. 'Inheritance' is a key notion which means that object instances may make use of attributes and procedures associated with the class of which they are members. These serve as default values which may be overridden by specifying particular values for these 'slots' in the instances themselves. Often *multiple inheritance* is supported such that an object may inherit attributes from more than one superior class. For this reason, the generalization "hierarchy" is often a lattice, rather than a tree. A number of programming systems support this view of object representation, which is well-described in the literature [29, 30].

An object's location in such a generalization lattice tells us what it is, and what it looks like (i.e., how to render it). We can describe the structure of an object by associating with each node in the generalization lattice a transformation hierarchy as described above. The node for a human figure, for example, would have an associated tree structure describing the joints and links that make up the figure. Each of the links, in turn, would be an instance of a monolithic object located elsewhere in the generalization lattice. Instances of the class of human figures can inherit this structural description with local variables that specify the dimensions and movement constraints of a particular human being.

Like structural descriptions, each object must have an associated behavioral description. For simple objects, the behaviors would be correspondingly simple. The prototypical physical object, for example, might obey some subset of the laws of Newtonian mechanics. Articulated figures would have a repertory of skills, such as walking and grasping. But since behaviors too can be inherited, human figures, being instances of the class of physical objects, can inherit all those simple Newtonian behaviors we'd expect.

In addition, it is necessary to represent the mechanical interaction of objects in terms of a small, well-defined set of relationships. Since we want to represent a world changing over time, these relationships must be dynamic, and include links between objects that signify support, contact, containment, epsilon-proximity, and whether one object is a part-of, or a movable-part-of another. The part-of

relationship may well apply to objects that are themselves complicated assemblies, each with their own structural description, e.g., an engine is part-of a car.

Lastly, in order to do simple motor problem solving, we want to embed commonsense knowledge in object descriptions. That is, we want to be able to encode such knowledge as "One usually leaves a room by finding and opening the door."\* Such knowledge represents cultural information learned by individuals early on; it is "common knowledge" we all know about doors. It is appropriate to associate such information with the objects themselves. That is, in addition to modeling the physical and geometric properties of a microworld, the world knowledge base must contain cultural information attached to objects as well. We are currently investigating techniques for encoding such information in terms of uniform behavioral descriptions for all the objects contained in the generalization lattice.

#### 4.5. Text-Mediated and Device-Mediated Interaction

The power of an animation system derives ultimately from its available abstraction mechanisms and the implementation of adaptive movement, not simply by providing the animator with joysticks, knobs and dials. Much has been made of device-mediated interaction in computer graphics, especially in early work (e.g.[31, 32]), begun at a time when Fortran or assembly code may have been the only alternative means of human-machine communication. However, language will probably remain the medium of choice for describing algorithms and complicated spatial, temporal, and behavioral relationships. Much of the objection to text-mediated interaction really is an objection to typing. Progress in improving the ergonomics of the typewriter keyboard and ultimately, developments in speech recognition will go a long way towards ameliorating this aspect of the human-machine interface.

At the same time, there are many functions, e.g., picking, locating, and sketching, for which the graphical gesture clearly is the preferred mode of interaction. Perhaps the

\* This is not always the case in animation, of course! Therefore such representations must be easily modifiable.

ultimate example of graphical interaction is that of flying a simulated airplane, or steering the six-legged walk of a science fiction robot ant with a joystick. But these are large simulation programs built on a complex set of procedures. The user interacts with the top level of a hierarchy of abstractions, and it is this organization that allows small movements of the operator's hand on a joystick to be amplified into a complex of meaningful control signals with such a powerful result.

In the following sections, we will see how three levels of control result from the graded implementation of adaptive motion and abstraction mechanisms.

#### 5. A Three Level Hierarchy for Character Animation

We can classify animation systems as being either guiding, animator level, or task level systems. (For a similar classification of robot programming systems, see [33]).

##### 5.1. Guiding

Guiding systems are those with no mechanisms for user-defined abstraction or adaptive motion. There are a wide range of guiding systems, including motion recording[34, 35], shape interpolation[4], key-transformation systems[3, 4, 18], and notation-based systems[36, 37].

In *motion recording*, various devices are used to acquire kinematic data from a moving figure. The kinematic data is then used to control an animated figure. Such systems are usually limited to measurements of a restricted range of human movement in a laboratory setting, but offer a potentially rich source of data on human motion. *Shape-interpolation* (also known as "metamorphosis") is the 3-D analog of 2-D keyframing. Where there is a one-to-one correspondence between the points and faces of separate objects, inbetween frames can be computed by interpolating between the data points of the two objects. In *key-transformation* systems, whole objects are manipulated by affine transformations. Inbetween frames are generated by interpolating the transformation parameters and transforming the objects. Such systems usually allow the specification of transformation hierarchies, making articulated motion possible. In such *key pose* systems, e.g.,



BBOP, a p-curve facility[31] is provided so that the user can graphically specify velocities. *Notation-based* systems are an example of text-mediated guiding in which the user describes a movement in a choreographic notation or an alphanumeric equivalent (i.e. [34]).

### 5.1.1. Limitations of Guiding Systems

In guiding systems, the animator must specify in advance the details of motion. This is reasonable only in a relatively featureless environment. Suppose a human character is to walk over rough terrain. Walk cycles are not difficult to generate using keyframing or shape interpolation, but in this case, the walk cycle changes with each step, requiring a large number of intermediate configurations to ensure that the motion looks right. This is because the inbetween frames are computed without regard for other objects in the scene. If a foot goes through the floor, or the figure walks right through a wall, so be it. What is worse, if the character is to walk in another direction over different terrain, none of the earlier key configurations can be used.

In guiding, the animator has nearly complete control over the motion of a figure. Because of the nature of the DOF problem, this is both a blessing and a curse. The animator is free to design an expressive motion sequence in toto, but for complicated figures or intricate mechanisms this is a demanding or perhaps an impossible task, even with a well-designed device-mediated interface[38].

Most guiding systems include predefined procedural abstractions for smoothing motion based on one or several spline techniques[39]. Often these tools allow the animator to interactively adjust the spline parameters until some desired trajectory is achieved. Splining allows the animator to more closely simulate the dynamics of rigid bodies, e.g., acceleration and deceleration due to inertia, friction, or gravity, since motion that is linear and jerky doesn't look right and is often unpleasant to view. (In conventional animation acceleration and deceleration are referred to as *ease in* and *ease out* respectively, and must be calculated by hand and from tables). In general, splining provides convenient control over the velocity of many kinds of transformations, including changes in size, shape and color, in addition to changes in

position and orientation. The value of using parameterized curves to control animation was recognized early on[31], and the refinement of these techniques remains an active area of interest (see e.g.[40])., While the use of spline curves is a powerful simulation mechanism, spline techniques alone are not a general solution to the DOF problem, since the control of many transformations requires the generation and refinement of many splines.

To date, a number of interesting animation sequences have been produced using guiding systems at various commercial production houses and university laboratories. However, since powerful abstraction mechanisms are not provided, and because adaptive motion is not possible at all, guiding systems do not scale up well for use with complicated figures, and their utility for controlling animation in complicated environments is limited.

### 5.2. Animator-Level Systems

A number of animator level systems have been designed to allow the animator to specify motion algorithmically. A few of these systems, while not specifically designed as character animation systems, do provide some measure of one or both adaptive motion and abstraction.

#### 5.2.1. GRAMPS, ASAS, and MIRA

GRAMPS [41] has no facility for adaptive motion, but does allow the construction of motion macros based on functional abstraction. Joints can be grouped together and their input derived from dials, and the motion at the joints can be explicitly constrained to lie within some range of values. This is a good example of the interaction of a guiding mechanism (dials) and a functional abstraction (motion macros). While not designed as a character animation system, GRAMPS has been used to generate interesting animation of a human figure.

Craig Reynolds' ASAS [11] provides a set of low-level mechanisms for both abstraction and adaptive motion. The actor paradigm explicitly provides a general abstraction mechanism allowing the definition of transformation hierarchies (structural abstraction) and behaviors (procedural and functional abstractions). The message passing mechanism makes it possible to implement adaptive motion, since animated entities can report aspects of their physical

attributes or their internal states.

Another recently reported animation system, MIRA[5], is based on a programming paradigm closely related to actor-based systems, namely, the data abstraction[12]. MIRA provides a set of important abstraction facilities nearly identical to those of ASAS. While MIRA is not a message-passing system, the animator can set and examine the values of variables (of various data types), so that attributes of figures and objects can be used to influence the generation of movement.

### 5.2.2. TEMPUS

The group led by Norman Badler at the University of Pennsylvania has long been involved in research on representing and portraying human movement. They have developed TEMPUS[42, 43], a system for analyzing and displaying the movements of realistic human figures in a workspace. While not a general-purpose animation system, TEMPUS has sophisticated features for defining and modifying human figures, and for resolved motion control.

Because the domain of TEMPUS is restricted, unlike MIRA and ASAS, to positioning and orienting human figures, TEMPUS can be largely device-mediated. Users pick actions from a graphically displayed menu, and control motions using displays of simulated potentiometers. Available movements are rotation and translation of the whole figure, rotations at selected joints, and resolved motion of the limbs.

TEMPUS has no facilities for adaptive motion, and abstraction mechanisms available to the animator are limited to a parameterless macro facility which allows the user to group movement commands. The implementation of a flexible resolved motion algorithm for positioning the limbs of a human figure is an important step towards task-level animation.

### 5.2.3. Discussion

Because it is possible to implement adaptive motion, and to define structural, functional and procedural abstractions, animator level systems provide significant improvements over guiding in terms of the DOF problem. But as usual, there is a trade-off. Guiding systems are

relatively easy to learn and use, but lack the power to control complicated animation. Animator level systems, on the other hand, provide the computational power of a general programming language but at the same time saddle the user with all the problems so closely associated with software development. Thalman et al note that "it took 14 months to produce [a] 13-minute film," certainly highlighting the problem[5]. That is, while it is possible to develop complex motion in either ASAS or MIRA, it is not necessarily easy, since neither language provides explicit, high level support for developing functional abstractions or adaptive motion. Interestingly, Thalman et al. note that they found it necessary to integrate a guiding system, MUTAN[44], into their production scheme. I will have more to say about integrating control modes later on.

### 5.3. Task Level Animation

At the task level, the animation system must schedule the execution of motor programs to control characters, and the motor programs themselves must generate the necessary pose vectors. To do this, as we have seen, a knowledge base of objects and figures in the environment is necessary, containing information about their position, physical attributes, and functionality.

In[45] I outline one approach to task level animation in which motor behavior is generated by traversing a hierarchy of skills (represented as frames[46] or actors [47] in an object-oriented system) selected by rules which map the current action and context onto the next desired action. Albus of the Bureau of Standards has designed a robot control system based on a hierarchy of table-driven computing elements[48]. Powers has outlined a behavioral control hierarchy based entirely on servomechanism theory[49]. Both of these latter approaches seem to work well at the lower levels of motor control and what we might call instinct-driven behavior, but seem rather vague when it comes to behavior requiring symbolic interaction with the environment.

Task level motor control is a difficult problem under study by cognitive scientists, roboticists, and of course those interested in high level animation systems. In the near term we can expect the development of prototype

systems capable of generating rather simple behaviors. How well such systems scale up depends on our understanding of the motor control problem itself.

With task level control, the animator can only specify the broad outlines of a particular movement and the animation system fills in the details. Whether this approach is appropriate depends on the particular application. A non-expert user may be satisfied with the 'default' movements and figures the system provides if he or she can produce, in a reasonable amount of time and at a reasonable cost, an animation that gets the point across. A 'high-end' user, say, in the entertainment industry may want nearly total control over every nuance of a character's movement to make a sequence as expressive as possible. However, control over the expressive qualities of movement does not mean that the animator needs or wants a pure guiding system to generate pose vectors. The animator does need access to different levels of the control hierarchy in order to generate new motor skills and to 'tweak' the existing skills.

## 6. Integration of Control Modes

Guiding is the prevalent mode in most current interactive animation systems. The necessity for integrating all three modes of control stems from the inability of any one mode to provide complete yet economical control. Guiding is best suited for specifying fine details but unsuited for controlling complex motion. Animator level programming is powerful but difficult. Task level systems give us facile control over complex motions by trading off explicit control over the details of motion.

Part of the solution lies in applying guiding techniques at appropriate points in the motion control hierarchy. The key is the ability to decompose the movement repertoire into a manageable set of hierarchically organized skills. The notion of *browsers*, as implemented in Smalltalk[30] or Loops[29] suggests a powerful method for attaching guiding controls to motor skills. Suppose I have on my RGB monitor a shaded display of a human character. On my terminal screen is a representation of the structure of the character and its skills. Now suppose I trace a curve on the graphics tablet. If I specify that that curve represents a particular joint rotation, -- i.e., I point to the node for the little finger on my terminal, I should

immediately see on the display the little finger of my character wiggling. Suppose now I point to the node for "grasping with the left hand" -- I should see the figure's left hand open and close with the velocity I have specified. Lastly, if I pick the node labeled "walk", the figure should begin to walk across the screen, and this time, the curve I have drawn could determine, say, the speed of the gait.

This modular, hierarchical organization allows the user to identify the motion qualities that need to be adjusted, and at the same time it helps to localize the effect of such changes. This calls for a uniform representation of motor skills that incorporates, for each skill, a specification of the kinds of adjustments that are possible, and, in addition, a uniform set of mechanisms, e.g., p-curves, for interacting with skills.

## 7. Conclusion

I have presented a conceptual analysis of the domain of three dimensional computer animation, which is viewed as the process of simulating objects and their behaviors in a microworld specified by the animator. The degrees of freedom problem is the central issue in the coordination of articulated figures. Computer animation systems must be based on the appropriate set of domain concepts, namely adaptive motion and the five abstraction mechanisms, to enable the animator to define and manipulate interesting characters and environments in an expressive way.

The discussion of the three control modes suggest criteria for good guiding and animator level systems. Guiding systems have received the greatest attention to date -- the notion of an interactive, device-mediated interface has come to be viewed almost as a standard way of communicating with computers, as evidenced by the popularity of the "mouse-and-window" style of computing. In general, however, guiding should be seen as a mechanism for developing and controlling the behavior of complex systems, rather than just picking points, drawing lines, or generating scalar values for various transformation parameters. As suggested in above, we want to be able to attach the output of a physical input device at arbitrary levels of a behavioral hierarchy. While the meaning of a gesture depends, of course, on the process that is viewing it, the hard question is to find an

appropriate set of parameters for controlling a complex process, for example, facial expressions[50,51]. Once a natural control set has been determined, it is not hard to use input devices to generate parameter values interactively. There are two complementary design themes: How can we "plug in" guiding mechanisms to drive a given complex behavior? How can input device modules serve as standard "gesture amplifiers" that can be easily redirected to various functions of the figure control hierarchy?

An animator level language should incorporate the design features and principles we expect in a powerful programming language. Concealing the programming task from the user or sugar-coating the syntax is not nearly as important as providing the expressive power needed for animation. This is not the place for a discussion of the future of automatic programming, nor of the merits of the latest programming paradigm. The point is that the algorithmic description of behavior -- "Do this, then do that" -- is an essential and fundamental way to communicate about movement. More often than not the so-called "naive user" will quickly learn the syntax of an animation language only to become frustrated because the language is not powerful enough. Animation level languages and systems should therefore combine what we know about software technology with the mechanisms appropriate to motion control -- e.g., functional abstraction and adaptive motion.

Finally, adaptive motion in the form of collision testing, and resolved motion should be implemented, at least in part, as basic elements of any 3-D computer animation system.

The art and science of 3-D computer animation continues to evolve towards the simulation of hypothetical worlds complete with physical laws and figures possessing behavioral repertoires. It is by learning to construct and control these simulations that we give computer animation its expressive power.

#### References

- [1] N. Burtnyk and M. Wein, "Interactive Skeleton Techniques for Enhancing Motion Dynamics in Key Frame Animation," *Communications of the ACM*, vol. 19, no. 10, pp. 564-569, October 1976.
- [2] E. Catmull, "The Problems of Computer-Assisted Animation," *Computer Graphics*, vol. 12, no. 3, pp. 348-353, Proc. ACM SIGGRAPH 78, July 1978.
- [3] R. Chuang and G. Entis, "3-D Shaded Computer Animation -- Step-by-Step," *IEEE Computer Graphics and Applications*, vol. 3, no. 9, pp. 18-25, Dec 1983.
- [4] J. E. Gomez, "Twixt: A 3-D Animation System," *Proc. Eurographics '84*, North-Holland, September 1984.
- [5] N. Magnenat-Thalman and D. Thalman, "The Use of High-Level 3-D Graphical Types in the Mira Animation System," *IEEE Computer Graphics and Applications*, vol. 3, no. 9, pp. 9-16, Dec 1983.
- [6] F. Thomas and O. Johnston, *Disney Animation: The Illusion of Life*, New York: Abbeville Press, 1981.
- [7] A. Kay and A. Goldberg, "Personal Dynamic Media," *Computer*, pp. 31-41, March 1977.
- [8] A. Kay, "Computer Software," *Scientific American*, vol. 251, no. 3, pp. 52-59, September 1985.
- [9] J. Denavit and R. B. Hartenberg, "A Kinematic Notation for Lower-Pair Mechanisms Based on Matrices," *Journal of Applied Mechanics*, vol. 23, pp. 215-221, June 1955.
- [10] M.T. Turvey, H.L. Fitch, and B. Tuller, "The Problems of Degrees of Freedom and Context-Conditioned Variability" in *Human Motor Behavior*, ed. J.A.S. Kelso, Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1982, pp. 239-252.
- [11] C. W. Reynolds, "Computer Animation with Scripts and Actors," *Computer Graphics*, vol. 16, no. 3, pp. 289-296, Proc. ACM SIGGRAPH 81, July 1982.
- [12] M. Shaw, "The Impact of Abstraction Concerns on Modern Programming Languages," *Proc. of the IEEE*, vol. 68, no. 9, pp. 1119-1130, September 1980.
- [13] F. C. Crow, "A More Flexible Image Generation Environment," *Computer Graphics*, vol. 16, no. 3, pp. 9-18, Proc. ACM SIGGRAPH 82, July 1982.

- [14] J. F. Blinn, "Systems Aspects of Computer Image Synthesis," *Course Notes, Seminar on Three Dimensional Computer Animation*, ACM SIGGRAPH 82, July 1982.
- [15] C.S.G. Lee, "Robot Arm Kinematics, Dynamics, and Control," *Computer*, vol. 15, no. 12, pp. 62-80, December 1982.
- [16] R. Paul, *Robot Manipulators: Mathematics, Programming, and Control*, MIT Press, 1981.
- [17] D. Zeltzer, "Representation and Control of Three Dimensional Computer Animated Figures," Ph.D. Thesis, The Ohio State University, August 1984.
- [18] L. Williams, "BBOP," *Course Notes, Seminar on Three-Dimensional Computer Animation*, ACM SIGGRAPH 82, July 27, 1982.
- [19] R. D. Tennent, *Principles of Programming Languages*, Englewood Cliffs, NJ: Prentice-Hall, 1981.
- [20] D.E. Whitney, "The Mathematics of Coordinated Control of Prosthetic Arms and Manipulators," *Transactions of the ASME, Journal of Dynamic Systems, Measurement, and Control*, vol. 122, pp. 303-309, December 1972.
- [21] C. Klein and C. Huang, "Review of Pseudoinverse Control for Use with Kinematically Redundant Manipulators," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-13, no. 3; pp. 245-250, March 1983.
- [22] M. Girard and A.A. Maciejewski, *Computational Modeling for the Computer Animation of Legged Figures*, Proc. ACM SIGGRAPH 85, July 1985.
- [23] E. A. Ribble, "Synthesis of Human Skeletal Motion and the Design of a Special-Purpose Processor for Real-Time Animation of Human and Animal Figure Motion," M.S. Thesis, The Ohio State University, June 1982.
- [24] J. H. Clark, "Hierarchical Geometric Models for Visible Surface Algorithms," *Communications of the ACM*, vol. 19, no. 10, pp. 547-554, October 1976.
- [25] H. Fuchs, Z. Kedem, and B. Naylor, "On Visible Surface Generation by A Priori Tree Structures," *Computer Graphics*, vol. 14, no. 3, pp. 124-133, Proc. ACM SIGGRAPH 80, July 1980.
- [26] S. Rubin and T. Whitted, "A 3-Dimensional Representation for Fast Rendering of Complex Scenes," *Computer Graphics*, vol. 14, no. 3, pp. 110-116, Proc. ACM SIGGRAPH 80, July 1980.
- [27] W. R. Franklin, "3-D Geometric Databases Using Hierarchies of Inscribing Boxes," *Proc. Conf. Canadian Society for Man-Machine Interaction*, pp. 173-180, June 1981.
- [28] K. Wasserman, "Physical Object Representation and Generalization," *AI Magazine*, vol. 5, no. 4, pp. 28-42, Winter 1985.
- [29] M. Stefik, D. Bobrow, S. Mittal, and L. Conway, "Knowledge Programming in Loops: Report on an Experimental Course," *AI Magazine*, vol. 4, no. 3, pp. 3-13, Fall 1983.
- [30] L. Tesler, "The Smalltalk Environment," *Byte*, vol. 8, no. 8, pp. 90-147, August 1981.
- [31] R. M. Baecker, "Picture-driven Animation," *Proc. AFIPS Spring Joint Computer Conf.*, vol. 34, pp. 273-288, Spring 1969.
- [32] I. E. Sutherland, "Sketchpad: A Man-Machine Graphical Communication System," *Proc. AFIPS Spring Joint Computer Conf.*, vol. 23, pp. 329-346, Spring 1963.
- [33] T. Lozano-Perez, "Robot Programming," AI Memo 698, MIT, Cambridge, MA, December 1982.
- [34] T.W. Calvert, J. Chapman, and A. Patla, "The Integration of Subjective and Objective Data in the Animation of Human Movement," *Computer Graphics*, vol. 14, no. 3, pp. 198-203, Proc. ACM SIGGRAPH 80, July 1980.
- [35] C. Ginsberg and D. Maxwell, "Graphical Marionette," *Proc. ACM SIGGRAPH/SIGART Workshop on Motion*, pp. 172-179, April 1983.

- [36] T. W. Calvert, J. Chapman, and A. Patla, "Aspects of The Kinematic Simulation of Human Movement," *IEEE Computer Graphics and Applications*, vol. 2, no. 9, pp. 41-50, November 1982.
- [37] L. Weber, S. W. Smoliar, and N. I. Badler, "An Architecture for the Simulation of Human Movement," *Proc. ACM Ann. Conf.*, pp. 737-745, 1978.
- [38] D. Lundin, "3-D Modeling, A Personal Orthodoxy," *Course Notes, Seminar on Three-Dimensional Computer Animation*, ACM SIGGRAPH 82, July 27, 1982.
- [39] D. F. Rogers and J. A. Adams, *Mathematical Elements for Computer Graphics*, New York: McGraw-Hill, 1976.
- [40] D. H. U. Kochanek and R. H. Bartels, "Interpolating Splines with Local Tension, Continuity, and Bias Control," *Computer Graphics*, vol. 18, no. 3, pp. 33-41, Proc. ACM SIGGRAPH 84, July 1984.
- [41] T.J. O'Donnel and A.J. Olson, "GRAMPS -- A Graphics Language Interpreter for Real-Time, Interactive, Three-Dimensional Picture Editing and Animation," *Computer Graphics*, vol. 15, no. 3, pp. 133-142, Proc. ACM SIGGRAPH 81, August 1981.
- [42] J. Korein, J. Korein, G. Radack, and N. Badler, "TEMPUS User Manual," Unpublished, Dept. of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, September 1983.
- [43] Norman I. Badler, "Design of a Human Movement Representation Incorporating Dynamics," *Course Notes, Seminar on Three-Dimensional Computer Animation*, ACM SIGGRAPH 82, July 27, 1982.
- [44] Denis Fortin, Jean-Francois Lamy, and Daniel Thalman, "A Multiple Track Animator System for Motion Synchronization," *Proc. ACM SIGGRAPH/SIGART Workshop on Motion*, pp. 180-186, April 1983.
- [45] D. Zeltzer, "Knowledge-based Animation," *Proc. ACM SIGGRAPH/SIGART Workshop on Motion*, pp. 187-192, April 1983.
- [46] M. Minsky, "A Framework for Representing Knowledge" in *The Psychology of Computer Vision*, ed. P. Winston, New York: McGraw-Hill, 1975.
- [47] C. Hewitt, "Control Structure as Patterns of Message Passing" in *Artificial Intelligence: an MIT Perspective*, ed. R. H. Brown, Cambridge, MA: MIT Press, 1979, pp. 433-465.
- [48] J. S. Albus, *Brains, Behavior and Robotics*, Peterborough, NH: Byte Books, 1981.
- [49] W. T. Powers, *Behavior: The Control of Perception*, Chicago: Aldine Publishing Co., 1973.
- [50] F. I. Parke, "Parameterized Models for Facial Animation," *IEEE Computer Graphics and Applications*, vol. 2, no. 9, pp. 61-68, November 1982.
- [51] S. M. Platt and N. I. Badler, "Animating Facial Expressions," *Computer Graphics*, vol. 15, no. 3, pp. 245-252, Proc. ACM SIGGRAPH 81, August 1981.



# MIT Industrial Liaison Program

## Report

Papers relevant to the symposium:

"MEDIA TECHNOLOGIES"

October 3, 1985

"Synthetic Performers"

by

Professor Barry Vercoe

- 1) "The Synthetic Performer in the Context of Live Performance," by Barry Vercoe
- 2) "Synthetic Rehearsal: Training the Synthetic Performer," by Barry Vercoe and Miller Puckette

These papers have been duplicated at the request of the speaker.



Distributed for Internal Use  
by Member Companies Only.  
May Not be Reproduced.

© MIT

THE SYNTHETIC PERFORMER  
IN THE CONTEXT OF  
LIVE PERFORMANCE

Barry Vercoe  
Experimental Music Studio  
Media Lab.  
M.I.T.

Objectives

The purpose of this paper is to describe a new area of computer music research — one which is certain to become a major territory for future work. The research objective can be clearly stated:

to understand the dynamics of live ensemble performance well enough to replace any member of the group by a synthetic performer (i.e. a computer model) so that the remaining live members cannot tell the difference.

The import of this advance is to move computer music clearly into the arena of live music performance. I mean here to break clear of the "music-minus-one" syndrome that characterizes tape and instrument pieces of the present; I also mean to recognize the computer's potential not as a simple amplifier of low-level switching or acoustic information (keyboards and live audio distortion), but as an intelligent and musically informed collaborator in live performance as human enquiry.

The circumstance prompting this research has been an IRCAM commission to compose a work for flutist Larry Beauregard, along with access to the 4X real-time audio processor. My immediate instinct was to return to concepts and real-time software principles I had developed in 1972, while working on the design of a large digital synthesizer with real-time performer control. Although the present implementation is written in C and runs under RT-11 on the PDP-11/55 controlling the 4X, the principles remain the same as then, and could similarly be applied to any other target system.

The control structure for modelling a synthetic performer has three main parts:

I LISTEN:

1. Catch and parse incoming live events.
2. Extract tempo; score position, loudness, etc.

II PERFORM:

1. Set new performance tempo, loudness, phrasing.
2. Organize the computer performance.

III LEARN:

1. Observe unexpected Listen behavior or Perform difficulty.
2. Keep as adjunct to the full score for future preparedness.

The three areas will now be described in more detail.

2. The Synthetic Listener

Catching and parsing a sequence of events played by a professional flutist implies pitch detection at a speed almost impossible for audio methods alone. The strategy here was to employ two additional sources of data: fingering information, and a musical score. Through a series of optical sensors installed on the keys by flutist Larry Beauregard, the list of possible sounding pitches was reduced to three. By piping the audio signal into three appropriate filters, the 4X could then resolve the ambiguity in about 35 milliseconds.

Both pitch and time information were next jointly mapped onto elements of the score, in such a way as to permit reasonable variance by the live performer. Music recognition turns out to require a significant degree of rhythmic elasticity, combined with smaller amounts of rhythmic and pitch fault tolerance. Recorded errors of two or more notes should automatically induce more rigorous pattern matching, sometimes extensive relocation. In the event of complete distress, the best strategy is to hold the current course until something recognizable occurs.

Extracting a sense of tempo from such a matched sequence requires further absorption of effects like agogic time shift. Time shifts observed within a beat can be weighted by position: modifications in the early part of a beat generally have less tempo significance than those occurring towards the end. Once a new Listen tempo has been determined and found reasonable, it is then posted for Synthetic Performer consideration.

### 3. The Nature of Performance

Determining the correct Synthetic Performer tempo involves two levels of action. First, at 12 milliseconds intervals the posted Listen tempo (in the form of a beatsize) is sampled and accumulated in beat-bins. This serves to integrate tempo over time with fairly high resolution. Then, about once every 200 milliseconds the beat-bins are used to determine the apparent Live Performer score position. This is compared with the Synthetic Performer score position, and an appropriately graceful catch-up action determined. Just five or so such determinations per second seems to represent adequately the manner in which performers do this kind of thing.

Modelling the physiology of performance is a shade more tricky. My view is that the events of an intended performance remain in strict metrical terms until just moments before action, when they are suddenly converted to physiological control objects that are impossible to retract. The period during which the human is involved in the physiological gesture of performance will depend on the person and on the device, but is somewhere about one-tenth of a second. In my performance model, once a scored event has crossed this threshold it virtually explodes into a cluster of active object modules, each representing some aspect of the event (rise time, transient effects, frequency and loudness curves) and each needing CPU service. The processor honors these requests on a priority queue basis, sending control data to time-tagged buffers conceptually residing in the synthesizer itself.

### Learning to Improve

The network of Control Processes representing the above can be regarded as modelling a neural organism that is strictly instinctive, without learned response. The most demanding test of a Synthetic Performer is how well it behaves in the absence of previously gathered information — by sight-reading, as it were, on the concert stage. Although one cannot avoid imbedding some stylistic bias in real-time programs during the course of their development, there has been a painstaking effort here to limit its effect. For example, despite the test pieces being primarily from the late Baroque (Handel and W.F. Bach flute sonatas), the system was able to respond equally well to contemporary literature (Boulez Sonatine).

I initially included a learning strategy in the Synthetic Performer, but that code generally remains disabled because it complicates the development and testing of the instinctive model. Once those aspects manage to exhibit a suitably high level of robustness, I plan to further develop the adaptive and learned response mechanisms. I expect that the addition of learned responses to particular scores and to recognized live input will vastly improve the measurable musicianship of the Synthetic Performer. I hope to report on such developments in the future.

Meanwhile, this paper concludes with a demonstration of grossly different performances of a Handel flute sonata — the flute part played live by Larry Beauregard, and the harpsichord accompaniment offered in response by a synthetic performer.

### 4. Acknowledgements

I would like to thank the Guggenheim Foundation for support of this work. Indebtedness also to Larry Beauregard for his tireless patience and encouragement, and to Miller Puckette for hours of discussion.

## Synthetic Rehearsal: Training the Synthetic Performer

Barry Vercoe & Miller Puckette  
Experimental Music Studio  
Media Lab, M.I.T.

### 1. ABSTRACT

Computer tracking of live instruments aims to understand the dynamics of live performance well enough to replace any member of an ensemble by a synthetic performer (computer model) so that the others cannot tell the difference. Modelling the performer experience has 3 major parts: *listen* (extract temporal and other cues from the other players), *perform* (organize a synchronized and suitably matched performance), and *learn* (remember enough of each experience to benefit future encounters).

While encouraging progress has been made on parts 1 and 2 (Vercoe, 1984), methods to date have relied exclusively on highly-sensitive live tracking. There has been no learning or training capacity to allow the benefits of previous experience. The synthetic performer is essentially "sight reading every time" on the concert stage.

This paper outlines a method of integrating past and present experience into new strategies of rehearsal and improved performance. During a performance run, past or learned information is used to sensitize the perceptual and cognitive components, giving them a bias towards certain expected live input behavior. Eventual pitch contour matching and durational best fit are thus made easier and more robust. As before, rhythmic elasticity, and both pitch and rhythmic fault tolerance, are a necessary part of a practical performance system.

### 1. BACKGROUND AND OVERVIEW

During 1983-84, research was conducted jointly at the *Institut de Recherche et Coordination Acoustique/Musique* in Paris, France, by Barry Vercoe (MIT) and Lawrence Beauregard (flutist of the IRCAM Ensemble Intercontemporaine). The objective was to understand the dynamics of live ensemble performance well enough to replace any member of a group by a **synthetic performer** (i.e. a computer model) so that the remaining live members could not tell the difference. This aimed to break clear of the "music-minus-one" syndrome that has characterized tape and instrument pieces of the past, and to recognize the machine's potential not as an amplifier of low-level switches and keys, but as an intelligent and musically informed collaborator in live performance.

The research was carried out using standard repertoire (flute sonatas by Handel and W.F. Bach) in which the flute part was played live and the responding harpsichord part was "performed" by a strictly synthetic accompanist. The motivation for the work was an IRCAM commission for Vercoe to compose a work, *Synapse*, for Larry Beauregard and

the 4X real-time audio processor. Proof of how well the computer can behave as chamber music player in both traditional and contemporary contexts was then given in public demonstration at the 1984 International Computer Music Conference (Vercoe, 1984).

The live demonstration routinely showed a facile ability to track and remain in sync with wildly changing tempos. It also showed an ability to deal with heavy doses of live performer errors, sloppy rhythms and general distortion that would likely throw off a live accompanist. However, in these early implementations there was no data retention from one performance run to another: there was no performance "memory", and no facility for the synthetic performer to learn from past experience. The synthetic performer was essentially *sight reading* on the concert stage every time.

The aim of subsequent research has been to conceive and model a working representation of music rehearsal and music learning. This has meant rethinking the music abstractions so as to more adequately represent strongly structured scores, and to more easily model those elements of cognition and learning that depend heavily on structure. However, since many contemporary scores are only weakly structured (e.g. unmetered, or multi-branching with free decisions), it has also meant development of score following and learning methods that are not necessarily dependent on structure. This has led to a flexible score-following method in which dependence on temporal structure and on previous rehearsal information can be parametrically controlled.

### 2. THE ANATOMY OF PERFORMANCE

#### 2.1 Organizing one's own Contribution

Live music performance comprises three major parts:  
Perception and cognition of external controls.  
Organizing one's own performance.  
Learning from the experience.

Although computer synthesis has traditionally specialized in organizing a performance, it provides a very poor model. Live music performance is decidedly more procedural, and many of its aspects remain unevaluated until the last possible moment in time. The chronological value of a 2-beat duration, for instance, is determinable only near the end of its performance. This means that the processes which control time-sensitive parameters (e.g. crescendo over several beats) must remain active throughout, and are susceptible to outside influences (changing tempo, etc.) at any time. In order to construct a synthetic performer, we must first understand the anatomical principles that drive real performers.

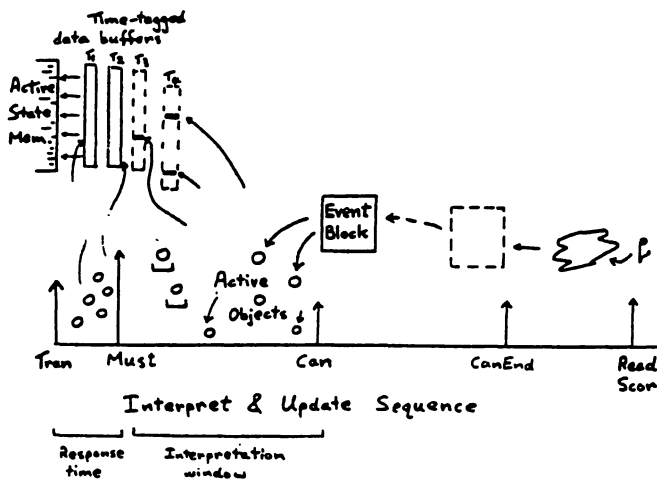


Figure 1.

The anatomical functions of a performance are depicted in Fig. 1. A musical note is initially understood as some future event, its distance away measurable only in beats. As the moment of performance draws near and the beat-defined event is drawn into active short-term memory, it passes a point where it can safely be given chronological definition ("can"). At this point the performer sets in motion the motor reflexes that commit him to the note's performance. The decision to perform may be delayed slightly, but not past a certain critical point ("must") which is the last chance to initiate the physical gesture that will deliver the note on time. The values of "can" and "must" in relation to the time of actual sound will depend respectively on the responsive sensitivity of the intended performance and on the physical inertia of the instrument and player concerned. Values of two-tenths and one-tenth of a second from actual sound are reasonable for a typical music situation.

The above scheme has been strictly represented in our synthetic performer system. The target instrument used in our studies has been either a digital model residing on a powerful audio processor (the IRCAM 4X), or a Yamaha DX7 synthesizer under MIDI control. In both cases, data is prepared in advance by a host processor and left in time-tagged buffers for automatic transfer to active state memory. The performer is modelled as a set of control processes running on the host processor. These sense external conditions and generate instrument update data values. Whenever an event crosses the "can" boundary, it bursts into life by spawning a host of some 20 active objects representing the multiple facets of a single sensitively performed note. These active objects compete for computing resources of a single CPU. Since they must do so in the "can-must" time span, the system is sensitive to the notion of performer overload.

## 2.2 Listen-Perform Synchronization

Computer performance of music can easily demonstrate that strictly synchronous behavior lacks much of the information we routinely seek from live performance. It is as if

the musical score acts as a carrier signal for other things we prefer to process. Much of this information derives from discrepancies between individual players. The degree of synchronization will vary as the live performer shifts the focus of his attention about. However, it is meaningful to ask: 'what are the tolerances involved, and what is the peak rate at which a skilled performer adjusts to incoming data?' The adjustment has an automatic throttle: upon sensing discrepancy, a performer will seek an aesthetic way of adapting, so as to preserve the integrity of his own line. How independent he can remain depends on what the score warrants and supports. We have here, in effect, a loosely-coupled system of performance and control, whose parameters depend on the topology of the score involved.

Listen-perform synchronization is based on two perceptions: currently observed tempo, and currently observed score position. Given an accurate observed tempo, and a comparison of the relative score positions, it is possible to determine by how much the synthetic performer should speed up or slow down. However, rhythmic time-shifts (either stylistic or in error) can distort both of these observations, and our initial synthetic performer implementation devoted considerable CPU attention to dealing with these aberrations. Many aberrant conditions can be described in the form of if-then productions, and much of our earlier success was due to this strategy. However, other quirks of interpretation can turn out to be quite localized and unsuited to such representation, even though they would likely occur at the same place in every performance of a given piece. These special moments catch a memoryless synthetic performer by surprise every time. Further progress on how skilled performers relate to one another is almost impossible without better models of score following, rehearsal and learning.

## 3. SCORE FOLLOWING

The problem of following a live player is complicated by the fact that neither the performer nor the sensors on his instrument are free from errors. As we move to the ultimately desirable situation in which we use only acoustic information from the performer, the frequency and magnitude of errors in the score tracker's input will increase. The success of the synthetic performer will be limited by our ability to reject errors while remaining sensitive to the useful information available.

Errors, both from the performer and from sensors, can be divided into qualitative errors (missing, extra, or wrong notes) and quantitative ones (early and late notes.) To reject the former it is sufficient to identify them; then we can ignore notes which do not correspond to the score, and use the timing information which is present even in wrong notes. Our best protection against quantitative error is to reduce it through averaging.

Our main desire is to extract reliable estimates for tempo and current beat, updating them as new events arrive. This amounts to finding our location in the score as a function  $\phi$  of time, which best fits the observed data. To do so we need to define a measure of how closely a theory about the current tempo and score position fits the incoming data.

Suppose that a part of the score is given by  $\{(s_1, x_1), \dots, (s_n, x_n)\}$ , where  $(s, x)$  denotes the event, "object  $x$  arrived at time  $s$ ." Suppose that we detect a performer as playing the sequence,  $\{(t_1, y_1), \dots, (t_m, y_m)\}$ . We want a theory of the form,  $\{(a_i, b_i)\}$ ,  $a_1 < a_2 < \dots$ ,  $b_1 < b_2 < \dots$ , which is taken to mean, "The event  $(t_i, y_i)$  played by the performer corresponds to the event  $(s_{a_i}, x_{a_i})$  in the score."

We assign to such a theory a cost, which measures its deviation from what we imagine to be a good performance. First we charge for combinatorial errors:  $p_m$  for each note in the score which is missing from the performance,  $p_e$  for each extra note in the performance, and  $p_w$  for each wrong note (i.e. whenever  $x_{a_i} \neq y_{b_i}$  .) To this we add the cost associated to metrical errors. We choose nonnegative weights  $\{u_i\}$  and take the least-squares fit of a line through the points  $(s_i, t_i)$ , weighted by  $u_i$ . This line,  $t=ps+q$  say, represents an assumption of a locally constant tempo of  $p$  real seconds per score second. The metrical cost of the theory is a constant  $p_t$  times the deviation of the  $(s_i, t_i)$  from the line  $t=ps+q$ ; we measure this deviation as

$$\sqrt{\frac{\lambda_1}{\lambda_2}}$$

where  $\lambda_1, \lambda_2$  are the minor and major moments of

$\{s_i, t_i, u_i\}$ . We find that we will want to drop some points out of this linear fit, which means that we are ignoring the times at which the performer plays some notes. To do this to a note which is right we charge another constant  $p_r$  and if it is wrong (in which case we are more willing to ignore its timing) we charge  $p_{rw}$ . The cost of the theory is the sum of all the above charges.

The problem of tracking the performance is that of finding the least costly theory. We then use the values of  $p$  and  $q$  associated to that theory either to predict what the performer will do next or to schedule the performance of an accompaniment.

Since tempo is not a global constant but is only nearly constant over short stretches of time, we actually wish to form a new theory associated to each note the performer plays. We choose the weights  $u_i$  to emphasize recent events over ones further in the past, thus keeping our tempo measurements local. We thus constantly update our estimates for tempo and current beat on the basis of fresh data.

Finding the absolute optimum theory for a note in a score of nontrivial size is probably intractable, but we have found an algorithm which works well in practice. For each pair  $(i, j)$  we find the best theory we can fitting the first  $i$  events in the performance to the first  $j$  events in the score. We choose the cheapest of four theories. First, the pair  $(i, j)$  might be a point of true correspondence between the score and the performance. The cost of this theory is the combinatorial cost of the best  $(i-1, j-1)$  theory plus  $p_w$  if the current note is wrong plus the cost of a linear fit of  $(i, j)$  with the other correspondences in the  $(i-1, j-1)$  theory. Second,  $(i, j)$  might be a correspondence whose time we wish to ignore; in this case we leave it out of the linear fit but add a charge of  $p_r$  or  $p_{rw}$  depending on whether the note is right or wrong. Third, the  $i$ th note of the performance might be an extra note; hence we add  $p_e$  to the cost of our best  $(i-1, j)$  theory; and finally the  $j$ th note of the score could be omitted, costing  $p_m$  plus the cost of the best  $(i, j-1)$  theory.

This algorithm is a fusion of the purely combinatorial one of Dannenberg (1984) with our own methods for following both pitches and rhythms (Vercoe, 1984). As such, it is able to combine hints from the two regimes to keep it in the closest possible synchrony with the performer. Via the mechanism of the parameters  $p_m, p_e, p_w, p_r$  and  $p_t$ , we may tune the algorithm to the specific tendencies of a performer/sensor pair. For example, if it is known that we almost never get extra notes, we will improve performance by increasing  $p_e$ , for the algorithm will rarely lie that a note is missing then. Our current setting is  $p_e = p_m = p_w = 1$ ,  $p_r = .4$ ,  $p_{rw} = .15$ , and  $p_t = 10$ .

In this way we can follow a performer. That is to say, if the performer plays his score without too many mistakes and without rapidly changing his tempo, we can approximate his instantaneous tempo and score position at any time. But we will still be sensitive to two elements of the performance which do not convey information about tempo and which we should be able to reject. First, a human performer realizes music in part by giving it a rhythmic microstructure; we should not try to follow this as a changing tempo from note to note but should learn to expect it and account for it when we track the performer. Second, there is the truly nonrepeatable microstructure. Even though we believe that this is musically essential we do not wish it to fool our following. The natural way to do this is to use memory of past performances constructively.

## 4. REHEARSAL, MEMORY AND LEARNING

### 4.1 Storing the Experience

Without information from previous performances, the Synthetic Performer's view of another performer's part will remain a prescribed contextual record. The score will be only minimally structured, conveying only the barest of syntactic features and nothing of performance or semantic interpretation. Every live performance heard, then, is at odds with this score, and requires the full services of facile though transient interpreters to maintain a reliable sense of score position. There is no learning from experience. The condition is typified by consecutive performance runs that are identically surprised by the same idiosyncratic input.

The first step towards admitting rehearsal as a determinant of subsequent expectations is to save away as much rehearsal data as might be useful. In the temporal domain, the information is of two major kinds: rhythmic aberration, and tempo warping. The two are not distinct, and one can easily be mistaken for the other in small localized contexts. The perceptual trick is to constantly gauge the effective tempo and current score position, then to regard the remaining distortions as rhythmic. As each incoming note is recognized in the score, the beat fraction by which the event is earlier or later than expected is written into its performance record. This amount is strictly a difference index with respect to the currently believed score position. There is no attempt to amend it on the fly, even in the face of changing opinion.

### 4.2 Post-Performance Memory Massaging

The effect of rehearsal on musical memory is to permit construction of new semantic concepts that will help delineate the score in its next performance. We can use the rehearsal data to improve future tempo detection, for example, by gathering statistics about how the performance of each note in the score tends to relate to the extracted tempo. After each performance, we take the event list of performer action times and incorporate this experience into the score by doing an unrehearsed, non-real-time tempo analysis of the entire event list, thus giving a score position as a function  $\phi$  of real time. Since we are not calculating this in real time, we can look into the future to get a more accurate estimate of  $\phi$ . For each event  $(s_i, x_i)$  in the score we collect the sample mean  $m_i$  and standard deviation  $\sigma_i$  of the measured discrepancies in score position between the observed note and the expected one, or  $s_i - \phi(t_j)$ , where  $(i, j)$  is a correspondence between the real score and the observed one.



The first correction we now make is for  $m_i$ . We do tempo matching not to the events  $(s_i, x_i)$  but to the mean-corrected events  $(s_i - m_i, x_i)$ . In this way we prevent that part of the microstructure of performance which repeats from rehearsal to rehearsal from skewing the tempo observations.

We use the sample standard deviations  $\sigma_i$  to guess the relative strength of the correlation between the time the note occurs and the actual beat it lands on. When we use the event  $(s_i, x_i)$  as part of a tempo determination we weight it by  $\sigma_i^{-1}$ , since that yields weighted averages of minimum variance.

In this way, we have used rehearsal memory to aid the score follower in precisely the language it can best understand. Since the follower sees its problem as one of finding a best fit between theory and observation, the most useful information we can give it is an estimate of how much on the average the data fails to meet the theory and at what points the data is more or less important to match. Similar memory massaging strategies can be used for other information tracks.

## 5. CONCLUSION

We have described a method of tracking live performers in such a way that they may be accompanied by synthetic performers that learn from rehearsals. Continual semantic reinterpretation of a piece under rehearsal is a complex, data intensive task, and a fully representative interpretation requires numerous rehearsals before the gathered statistics reliably anticipate live performer behavior. When that point is reached, however, the synthetic performer can be described as a well-rehearsed musical colleague, capable of robust yet sensitive collaborative performance.

Eventually it should be possible to develop a synthetic performer that would not require priming with a written score of what initially to listen for. Chamber music players typically perform from single part-books, building their sense of the full score strictly from the experience of rehearsal. In that this appears to be a prime route by which those players inform their overall performance, we would eventually like to understand a little of how that works.

## 6. REFERENCES

- Dannenberg, Roger (1984). Tracking a Live Performer. *ICMC Proceedings*, 1984.  
Vercoe, Barry (1984). The Synthetic Performer in the Context of Live Performance. *ICMC Proceedings*, 1984.