A Visual Software Agent connected with the WWW/Mosaic

Hiroshi Dohi Mitsuru Ishizuka

Dept. of Information and Communication Engineering,
Faculty of Engineering, University of Tokyo
Tokyo 113, JAPAN

Abstract

We have so far proposed the Visual Software Agent (VSA), an anthropomorphous agent with realistic face, eye, ear and mouth functions. It intends to realize a prototype for advanced human interfaces with natural communication capability beyond current iconic interfaces; that is, it allows face–to–face human–computer interaction with speech dialog. This paper describes why and how we combine the VSA with the World–Wide Web (WWW), a world–wide vast information database. One of these advantages is that we can describe information for the VSA by Hyper–Text Markup Language (HTML), a widely used language for hypermedia in the WWW. Another important advantage is that the VSA makes use of the Mosaic as the part of the system, a hyper-media viewer, and offers users another communication channel of human–computer interaction on the Mosaic in addition to a current mouse interface. A user can navigate on the Internet by speech dialog with the VSA. Because current Mosaic and HTML are designed for a hypertext–based system, we indicate that there exist mismatches to some extent between the current system and our VSA. The VSA–Mosaic interface developed here is especially useful for persons unfamiliar with a computer and physically handicapped persons.

1 INTRODUCTION

As computer systems come into our daily life and the opportunity for people to access these systems are increasing, advanced human–computer interaction becomes one of the most important research topics. One of its ideal form may be the human interface which can carry out many complex tasks by a few simple operations without any exercise. Although a today's pervasive interface style is direct manipulation (e.g. mouse, tablet, etc) on Graphical User Interface (GUI), it is not necessarily helpful for users; another type of interface with high naturality and intelligence is desired in coming new information–rich environments. Some research groups have designed and developed *software agents* for helping users on their daily or intelligent tasks[14, 6]. The word "*software agents*" generally means computer programs that perform intelligent tasks autonomously like human agents. In particular, "interface software agents" take care of interface tasks between a human and a computer system or vast information space[12]. In most cases, these software agents have no figure and no face. For example, Maes's Maxims system[14] displays a simple caricature; however, this caricature itself is not an *agent*, but simple indicator which conveys the state of the agent to the user.

When an agent has a speech dialog function with a user in addition to autonomous behaviors, its figure becomes to play an important role. It may be reluctant psychologically for a user to mutter to a simple physical box of computer. We feel uneasy whenever we speak to without a communication partner. An anthropomorphic agent with a realistic face embodies an environment close to "face–to–face" communication between a human and a computer.

TOSBURG-II[19] is an interface agent system which simulates speech dialog in a hamburger shop. It uses a 2-D animation of a clerk to facilitate interaction with users.

Talkman[20, 16] is also an anthropomorphous agent system with speech dialog. It has a realistic 3-D facial image, and changes its facial expression according to the internal state of emotion. It has proved quantitatively that interfaces with facial displays reduce the mental barrier between users and computers. It is also one of *believable agent* systems[1].

We have developed an anthropomorphous agent called "Visual Software Agent" or simply "VSA"[11, 7, 4]. The VSA has a realistic 3-D face as well as a speech dialog function. Its face has a texture image obtained from a real human and is always naturally rocking; these features give friendliness and sometimes fascination to users. Recent computer graphics technology makes possible to generate a vivid realistic facial image and change its facial expression in real-time. It is no doubtful that the realistic face close to a real human face is better than a simple 2-D animation.

A user can request an assistance of his/her work or ask questions with speech input to the VSA appearing on a display. Acceptable speech inputs are restricted to some topics because of described language and conversation knowledge in the VSA system. Also, the speech input can not be always recognized correctly because of the performance of a speech recognizer and the variation of the user's utterance. Thus the VSA

sometimes generate questions to assure the user's intension. These questions arisen from the miss recognition of the user's speech are usually annoying to users in conventional speech dialog systems. However, because of the friendly face, the user's unpleasantness is mitigated in the VSA system. (Such a speech dialog management system developed for the VSA is described in [9].) After the VSA captures the user's intension, then it can access an associated hyper-media database and generate answers. It behaves like an electronic visual secretary. The visual anthropomorphous agent VSA is intended to realize an advanced human interface prototype beyond current iconic interfaces.

These anthropomorphous agent systems with speech dialog function, including our VSA system, have worked considerably on restricted problem areas. The databases for the anthropomorphous agents have been particular ones. Thus they lack a vast information database. Here we combine our visual anthropomorphous agent VSA with the World-Wide Web(WWW)[2], a world-wide vast information database.

The advantages of this connection are 1) the Hyper-Text Markup Language (HTML) description identical to the WWW pages can be used for providing information through the VSA, and 2) the VSA can use vast information stored in the WWW and information services built in the WWW. As a result, hyper-media data in a standardized format can be used for the VSA as well as for the WWW. Moreover, the VSA can play a navigator or guide with a friendly face and a speech dialog function for the WWW in the Internet.

This paper describes a prototype of the VSA system connected with the WWW through the Mosaic[15]. A user can navigate on the Internet by speech dialog with the VSA. If we replace a general VSA module with a personalized one with individual knowledge, the VSA will be a personalized competent agent.

2 Visual Software Agent

Desirable characteristics of the VSA are as follows.

- a realistic human-like face as a communication partner,
- natural human face–to–face communication ability with speech dialog,
- vast information database, and
- intelligence.

Toward this end, many research works including real-time visual processing, speech processing, intelligent behavior, etc., are required.

2.1 Realistic Face

For generating the realistic image of the VSA, we use a texture-mapping technique[8] and a deformable three-dimensional wire-frame model of a human face. Our wire-frame model consists of approximately 500 surface patches (triangle polygons). We use a full face photograph for texture-mapping. First, the face area of the photograph is decomposed into small triangle polygons, and then their texture data are mapped and transcribed onto the wire-frame facial model. It can re-compose any realistic facial image which can be seen from wide angles within a certain limit of direction, because of the use of the three-dimensional wire-frame model. We can change its expression by making deformation of the model, i.e., controlling positions of vertices. The VSA winks his/her eyes, opens his/her mouth like speaking synchronized with a speech synthesizer.

2.2 Rocking Face

There have been many researches so far on facial image generation. A synthesized face may change their facial expression, i.e., smile, anger and sorrow, but the base face itself has a fixed posture. Whenever turning a synthesized head, its movement is usually too smooth; it traces strictly on a scheduled path. The real face alive is always rocking and never freeze, unlikely to a robot. Our VSA rocks his/her face naturally all the time. The rocking face gives intimacy that can not be obtained from static images.

2.3 Parallel Image Generation

In the case of the computer graphics art, the main concern is generating high quality images. It is not rare to take a few hours even for only single image generation. However, in human interface field, image synthesis speed becomes an important factor rather than its quality. We don't necessarily need superfluous high-quality images. For our application, we require "realistic" facial images, but it is meaningless to generate a small part in detail since the VSA's face is always rocking.

Figure 1: The Visual Software Agent (VSA)

Both texture-mapping and the deformation of a three-dimensional wire-frame model in real-time take high computational costs. Only a few high-end graphic systems can support real-time texture-mapping operation by hardware, but they are too expensive. A compact less-expensive hardware with real-

time texture-mapping is desirable for our purpose. To perform a real-time generation of texture-mapped animated images, we designed and implemented a simple texture-mapping system with general-purpose parallel micro-processors, Transputer[10]. It is called "*ViSA (Visual Software Agent)*" system. A prototype of the ViSA system has four transputer modules on a single board, and attains good performance for realistic facial image synthesis. It executes both texture-mapping and deformation more than 10 frames/sec.

Figure 1 shows an example of the VSA. The left photograph on the left console is the original picture of a human face for texture-mapping. The right photograph on the console is a magnified one of the left that is overlayed with a wire-frame model. A synthesized facial image is displayed on the right monitor. The background is the ViSA system. A user puts a headset on and talks with her. She (the VSA) looks at the user (turns her head), winks, and speaks.

2.4 System Configuration

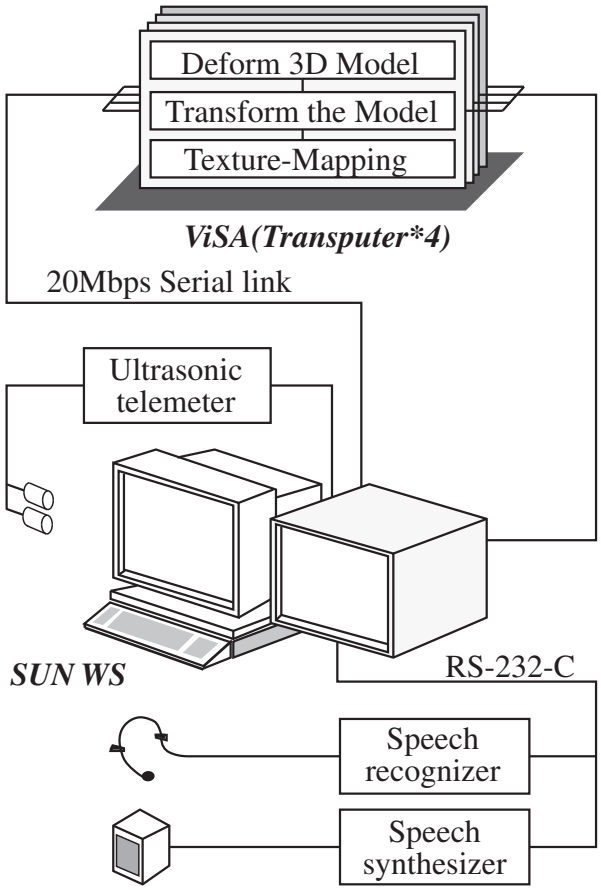

Figure 2: System configuration

Figure 2 shows the configuration of our prototype system.

The ViSA hardware for facial image synthesis is connected with a SUN workstation. Four transputers work in parallel. The realistic facial image is displayed on the workstation's console through a live video input or on an external monitor.

A speech synthesizer and a speech recognizer are attached to the workstation through RS–232–C for speech communication. Both are commercial products. The speech recognizer is a speaker-independent system, and can recognize continuous speech.

An ultrasonic telemeter is also attached to ensure the position of a user. Using this measurement, the VSA can direct her face to the user; as a result, a pseudo eye–to–eye contact is achieved, which is very important to improve friendliness in communication. Instead of the ultrasonic telemeter, we can use a vision system which we have developed.

3 VSA and WWW/Mosaic

3.1 WWW as a vast information database

Required characteristics of a desirable information database with which our VSA is expected to work are as follows;

- wide coverage area,
- large amount of information, including multimedia data (image, sounds etc.)

As a database grows, it becomes difficult for one or a few persons to update and maintain the database. That is, the database may require a standard format for many users' collaboration.

The WWW is a sort of the world-wide distributed information database. It is one of the databases which satisfy the above conditions. It is growing radically although it includes much junk data, and nobody knows the real capacity of the database.

The benefits of connecting the VSA with the WWW are that we can describe necessary information for the VSA information service by the HTML, and use accumulated vast information in the WWW as well for the VSA service.

3.2 VSA–Mosaic Communication

The WWW is based on a server–client model. When retrieving information from the WWW, the VSA accesses through the Mosaic, and does not communicate directly with the WWW. The Mosaic is a networked information retrieval tool which is developed at the National Center for Supercomputing Applications (NCSA). It is widely used as an interface to the Internet. The VSA makes use of the Mosaic as the part of the system, i.e., a hyper-media viewer.

On the Mosaic display, clicking a mouse on the highlighted string (called "anchor") invokes the retrieval of a linked object. Regardless of the type of a server, its interface is the same. However all of users can not manipulate a mouse satisfactory. In other words, it is unusable for people unfamiliar with a computer and physically handicapped persons.

Speech is an important media for communication or interface. The VSA offers users another communication channel of human–computer interaction on the Mosaic in addition to a mouse interface. It also offers

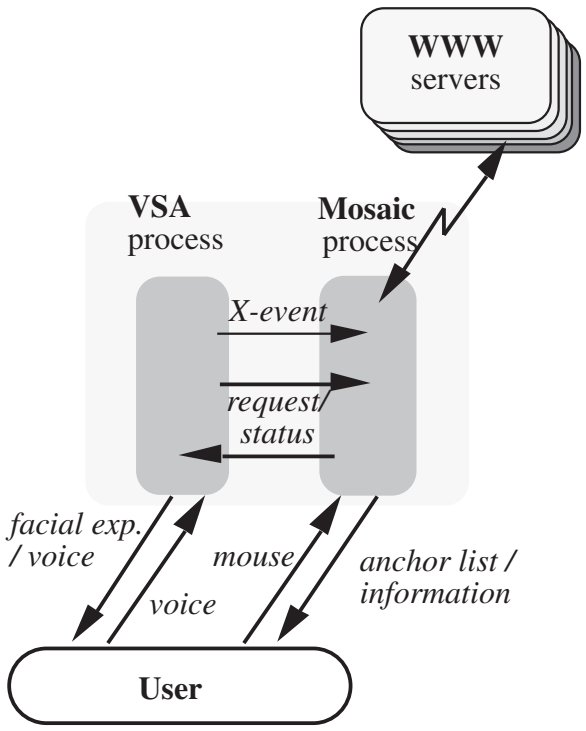

Figure 3: The design of VSA–Mosaic interface

a natural face–to–face communication with speech dialog.

Figure 3 illustrates the design of a VSA–Mosaic interface.

A VSA interface is designed and implemented as an independent process from the Mosaic one. Because of the design, it does not affect the original functionalities of the Mosaic at all. All mouse controls are available any time independent of the use of speech control. Also we can easily replace the VSA module with a personalized one. The VSA process and the Mosaic process may run on different workstations. Even if the VSA process is not invoked or fails to contact with the Mosaic process, the Mosaic works normally.

We have modified the source code of the Mosaic–2.4, and added an X–window event handler routine for the VSA–Mosaic communication. The VSA process creates an X–window event structure when the speech recognizer detects one of registered key strings, and sends it to the Mosaic process. It informs the Mosaic process the occurrence of a request from a user. Then some requests/statuses are exchanged between the VSA and the Mosaic through an X–Cut–Buffer.

3.3 Speech Commands

The VSA communicates with a user using speech dialog, and delivers a user's request to the Mosaic. The current VSA recognizes three types of speech commands for Mosaic operations.

- select–by–key–string
  When a user speaks one of registered key strings, an associated action is invoked. Each entry of a key string table includes pairs of the key string (not only single word) and the anchor string / URL (Uniform Resource Locator). Suppose the table has an entry that the key string is "*foo*" and its paired anchor string is "*bar*". When a user says "*foo*, please" or "please show me *foo*", the VSA sends its anchor string "*bar*" to the Mosaic process. It is equivalent to click with a mouse on the anchor string "*bar*." In current implementation, to increase the flexibility of the speech communication, we provide large varieties of key strings. It uses a maximum-length sub-string matching search. Thus, the string "*foo bar*" has priority over the word "*foo*". In case the URL instead of an anchor string is given, a specified page is open directly. It is equivalent to jump to the page using a hotlist function.

- select–by–number
  Our speech recognizer needs candidate words together with a set of grammar rules, though it can recognize user–independent, continuous speech. It is apparently impossible to register all anchor strings on a WWW server. Besides, some anchors like image anchors have no anchor strings. Thus the VSA system extracts anchors from a HTML text and presents an anchor list to the user on another window whenever a new page is opened. Each entry of the list is a pair of an index number and the anchor string. The speech input of the index number instead of the anchor string is also available for selecting an arbitrary anchor.

- page control
  These speech commands are used for page control. It is equivalent to click a certain functional button or a scroll bar on the Mosaic. These are reserved words such as;

  – *"Home page"*
    return to the home page

  – *"forward"/"back"*
    forward/back the page link

  – *"page up"/"page down"*
    page up/down on the page.

3.4 Communication Examples

Figure 4 shows an example display of the VSA–Mosaic interface. There are three windows on a screen. The right window shows a Mosaic page. Its appearance is the same as the original Mosaic system. The upper left window is the VSA with a rocking realistic face. The lower left window displays the anchor list of each Mosaic page.

Example 1 below shows a simple dialog example. A user speaks to the VSA, and the VSA replies with voice as well as changing the Mosaic page, if necessary.

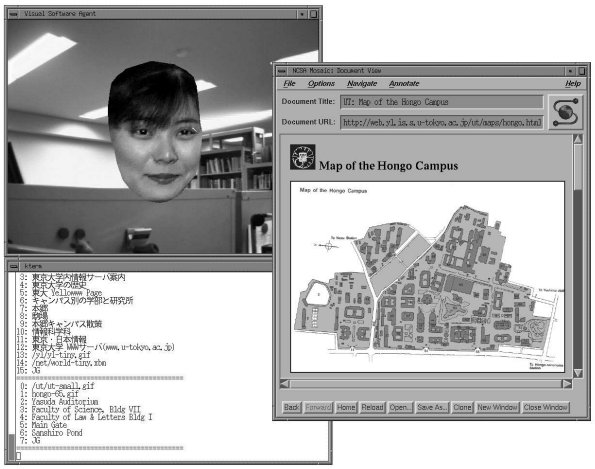

Figure 4: VSA–Mosaic interface

Although the current VSA doesn't have so many sentence patterns, one may feel that is relatively natural.

The speech recognizer used in our system returns both the sentence recognized and its score for decoding of the complete utterance. The score value is between 0 to 999 (best). If the score value is under a threshold, the VSA system discards the returned sentence and inquires again.

3.5 Problems

The WWW is designed for a hypertext–based system, and defines a common basic language, the HTML, for interchanging hypertext data. It is powerful enough to include multimedia data like images and sounds. However, there is no hints or guidelines to make a page manipulated well by speech dialog. Accordingly, there exist some mismatches and some problems listed below when the VSA speech interface manipulates arbitrary-written WWW pages.

- Location-dependent anchor
  In some cases, the location of an anchor may have important meaning rather than its anchor string. A simple but good example is the following one;

  "*XXXX* is here."

  The word with an underline ("here") is an anchor. We frequently find this form anywhere on Mosaic pages. It may be effective on a hypertext system. In our VSA–Mosaic system, when we say "here, please", it will work fine but it sounds strange.

- More than two identical anchor string
  It is no problem that more than two anchors have an identical string on one page. These anchors may have different link pointers or the same one. For example, they may appear as;

  "*foo* is here."
  "*bar* is here."

User: Please contact with "*the University of Tokyo*" server.

VSA: Yes. Just a moment please.
(connect with the server, then present an anchor list)
O.K. There are 5 anchors.

User: Number 0, please.
(0: Japanese version)

VSA: (open "*Japanese*" page, and present new anchor list)
O.K. There are 5 anchors.

User: Please show me c????s ...
(fail speech recognition)

VSA: Pardon?

User: Please show me "*campus map*"

VSA: (open "*campus map*" page, and present new anchor list)
O.K. There are 3 anchors.
...

Example 1. A simple dialog example.

  By using a mouse, it is easy to point a location out on a display. However, it becomes problem in the VSA system, which can not identify an intended anchor. In this case, the VSA system has to generate a question to discriminate anchors.

- Non-string anchor
  The use of non-string anchors like image anchors is one of the special features of the HTML. The image anchor has own file name, but no anchor string. An user has to use its index number on the anchor list, instead of the file name, in the VSA system. When the file has a meaningless name, it is not helpful to distinguish some image anchors.

- Too many anchors
  There may be more than a hundred anchors on one page. The fore part of the anchor list in the VSA system may scroll off when there are too many anchors.

To avoid these problems, a guideline of making WWW pages, which are accessed through the VSA–Mosaic system with the speech function, is required.

4 Conclusion

In this paper, we have described a prototype of the Visual Software Agent (VSA) system connected with the WWW, a vast information database. When a software agent has a speech dialog function with a user, its figure, particularly its human-like realistic face, becomes to play an important role to achieve friendly interface.

The WWW is designed for a hypertext–based system. It is powerful enough to include multimedia data like images and sounds. In hypertext format, the location information of anchors is often implicitly utilized for selecting next WWW page. This feature is good for the mouse interface. However, all of users can't manipulate a mouse satisfactory.

The VSA offers a user another communication channel, a face–to–face communication with speech dialog. It delivers a user's speech request to the WWW/Mosaic. A user can walk around the Internet by speech dialog with the visual anthropomorphous agent, VSA.

We have found some mismatches and some problems on the VSA–Mosaic communication through our experimental work. Speech dialog can't point location out well directly because of the characteristics of speech, though it is also an important media for human communication. Thus it can't fully replace a pointing device to manipulate arbitrary WWW pages. At present, a guideline of making WWW pages is required when they are manipulated well by speech dialog. It is important to know how to use various interfaces properly.

The VSA interface is independent from a vast database, and we can replace it with a various personalized one easily. The realistic anthropomorphous agent such as the VSA of this paper with speech dialog function and a vast information database is expected to play an important role for an advanced human–computer interaction as a competent agent in coming new information era. We are also working to add intelligence to the agent, so that the interaction between the agent and a user becomes more smooth and natural and the agent can navigate the user in the vast Internet information space.

References

[1] J. Bates: "The Role of Emotion in Believable Agents", *Commun. ACM*, Vol.37, No.7, pp.122–125, 1994

[2] T. Berners-Lee, R. Cailliau, A. Luotonen, H. F. Nielsen, and A.Secret: "The World-Wide Web", *Commun. ACM*, Vol.37, No.8, pp.76–82, 1994

[3] H. Dohi and M. Ishizuka: "Realtime Synthesis of Super–realistic Moving Goldfish using Small–scale Parallel Transputers", *Communicating with Virtual Worlds (N.M.Thalmann and D.Thalmann eds.)*, Springer–Verlag, pp.462–472, 1993.

[4] H. Dohi and M. Ishizuka: "Realtime Synthesis of a Realistic Anthropomorphous Agent Toward Advanced Human–Computer Interaction", *Software and Hardware Interfaces (G.Salvendy and M.J.Smith eds.)*, Elsevier, pp.152–157, 1993.

[5] O. Etzioni and D.S. Weld: "A Softbot–Based Interface to the Internet", *Commun. ACM*, Vol.37, No.7, pp.72–76, 1994

[6] O. Etzioni and D.S. Weld: "Intelligent Agents on the Internet: Fact, Fiction, and Forecast", *IEEE Expert*, Vol.10, No.4, pp.44–49, 1995

[7] O. Hasegawa, C-W. Lee, W. Wongwarawipat, and M. Ishizuka: "Realtime Synthesis of Human-like Agent in Response to User's Moving Image", *11th ICPR*, pp.39–42, 1992

[8] P.S. Heckbert: "Survey of Texture Mapping", *IEEE CG&A*, Vol.6, No.11, pp.56–67, 1986

[9] Y.Hiramoto, H.Dohi and M.Ishizuka: "A Speech Dialogue Management System for Human Interface employing Visual Anthropomorphous Agent", *Proc. RO-MAN '94*, pp.277–282, 1994

[10] INMOS Limited.: *"The Transputer Databook"*, second edition, 1989

[11] M. Ishizuka, O. Hasegawa, W. Wongwarawipat, C-W. Lee, and H. Dohi: "Visual Software Agent (VSA) built on Transputer Network with Visual Interface (TN–VIT)", *Proc. Computer World '91*, pp.36–46, 1991

[12] B.Laurel: "Interface Agents: Metaphors with Character", The Art of Human–Computer Interface Design (Ed. by B.Laurel), Addison–Wesley Publishing Company, Inc., pp. 355–365, 1990

[13] C-W. Lee, O. Hasegawa, W. Wongwarawipat, H. Dohi, and M. Ishizuka: "Realistic Image Synthesis of a Deformable Living Thing based on Motion Understanding", *Journal of Visual Commun. and Image Representation*, Vol.2, No.4, pp.345–354, 1991

[14] P. Maes, "Agents that Reduce Work and Information Overload", *Commun. ACM*, Vol.37, No.7, pp.31–40, 1994

[15] "NCSA Mosaic Home Page", URL: http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/help-about.html.

[16] K.Nagao and A.Takeuchi: "Social Interaction: Multimodal Conversation with Social Agents", *AAAI-94*, pp.22–28, 1994

[17] T.Oren, G.Salomon, K.Kreitman and A.Don: "Guides: Characterizing the Interface", *The Art of Human–Computer Interface Design, (B.Laurel eds.)*, Addison–Wesley, pp.367–381, 1990

[18] F.I.Parke: "Control Parameterization for Facial Animation", *Computer Animation '91, (N.M.Thalmann and D.Thalmann eds.)*, Springer–Verlag, 1991

[19] Y. Takebayashi: "Spontaneous Speech Dialogue System TOSBURG II –Towards the User-Centered Multimodal Interface–", *IEICE Trans. Information and Systems*, Vol. J77–D–II, pp.1417–1428, 1994

[20] A.Takeuchi and K.Nagao: "Communicative Facial Displays as a New Conversational Modality", *ACM/IFIP INTERCHI '93*, pp.187–193, 1993