論　文

WWW/Mosaicと結合した自然感の高い擬人化エージェントインタフェース

土肥　浩[†]　　石塚　満[†]

A Visual Software Agent Connected to WWW/Mosaic

Hiroshi DOHI[†] and Mitsuru ISHIZUKA[†]

あらまし　自然な人間の顔をもち，インタラクティブにユーザと音声対話できる知的擬人化エージェントであるビジュアルソフトウェアエージェント(VSA)は，マルチメディア時代のヒューマンインタフェースの形態として重要である．これは人間の日常的なface-to-faceコミュニケーションを実現する．このVSAを，インタネット上のWWW (World-Wide Web)用ブラウザ Mosaicと結合した．WWWで標準的に使用されているHTML言語で記述されたデータをそのままの形で，擬人化エージェントインタフェースから利用することができる．ハイパテキストデータの表示には，Mosaic画面を利用している．ユーザは擬人化エージェントとの簡単な音声対話によってMosaicを制御し，WWWサーバから必要な情報を引き出すことができる．マウスの操作が困難な計算機に不慣れな人や身体的ハンデのある人にも有効である．また，音声によりWWWから情報を引き出す場合に生ずるデータ作成上の問題点を明らかにした．

キーワード　擬人化エージェント，ソフトウェアエージェント，ユーザインタフェース，WWW，Mosaic

1. まえがき

計算機が広く一般に利用されるようになるにつれて，より人間の日常的コミュニケーションに近い自然なインタフェースが重要になってきている[1]．我々はこれまで，ビジュアルで知的な擬人化エージェントを介するマルチメディア時代の先進的ヒューマンインタフェースの実現を目指して，ビジュアルソフトウェアエージェント VSA (Visual Software Agent)の研究を進めてきた[2]～[4]．VSAは自然な人間の顔をもち，インタラクティブにユーザと対話できる知的擬人化エージェントである．操作方法を覚えたり練習したりすることなく，単純な操作方法で複雑なタスクを実行するために音声対話機能を備え，人間の日常的なface-to-faceコミュニケーションスタイルを利用している．

何か問題が生じたときにそれを解決する一つの手段は，その問題に詳しくて身近にいる人に質問し，助言あるいは答えを求めることである．ところが実際には，その問題に詳しい人がいつも身近にいるとは限らない．また運良くそういう人に恵まれたとしても，頻繁に質問を繰り返していては相手の仕事を妨害してしまうかもしれない．VSAはテレビ電話のように画面の中におり，ユーザの音声等による質問に答え，またタスクを実行する．

擬人化エージェントが備えるべき属性として，次のものが挙げられる．

- コミュニケーションパートナーとしての自然感の高い顔
- 音声対話などの自然なコミュニケーション能力
- 大規模な情報データベース
- 大量の情報を処理する知性

これらを実現するためには，リアルタイムの画像処理や音声処理，データベースや人工知能技術など，さまざまなテクノロジを統合する必要がある．

我々はこれまでに並列プロセッサを使用して，コミュニケーションパートナーとしての自然感の高い顔画像の実時間生成を実現した．また，その顔画像

† 東京大学工学部電子情報工学科，東京都
Department of Information & Communication Engineering, Faculty of Engineering, The University of Tokyo, Tokyo, 113 Japan

をもつ擬人化エージェントに対して，日付や時間，電子メールの到着などを音声で質問し，その結果を音声で聞くようなプロトタイプシステムViSA (Visual Software Agent System)を実装した[4]．これらの研究から，適当な情報データベースを用意することができれば，簡単な対話による自然な擬人化エージェントインタフェースが実現可能であることを確認した．

World-Wide Web (WWW)[5] はCERN (European Center for Nuclear Research) を中心に研究が進められている分散型のハイパテキストデータベースである．テキスト以外に音声やイメージも扱うことができる．WWWサーバは比較的簡単に立ち上げることができ，ユーザ自身が容易にデータを公開することができることから，現在も爆発的に拡大している．もしハイパテキスト記述言語HTML(Hyper-Text Markup Language)で記述されたWWWデータを，何らかの変換操作を加えることなくそのままの形で擬人化エージェントのための情報データベースとして利用することが可能であれば，その応用範囲は限りなく広がる．

本研究では，擬人化エージェントVSAとWWWクライアントソフトウェアの一つであるMosaicを結合したシステムを実装した．擬人化エージェントインタフェースのための大規模情報データベースとして，インタネット上のWWWサーバをアクセスする．ハイパテキストデータの表示には，Mosaic画面を利用する．ユーザは，マウス操作に加えて音声対話でも必要な情報を引き出すことができる．マウスの操作が困難な計算機に不慣れな人や身体的ハンデのある人にも有効である．

2. ビジュアルソフトウェアエージェント (VSA)

2.1 自然感の高い顔

コミュニケーションパートナーとして自然な印象をユーザに与えるためには，自然感の高い顔を生成する必要がある．アニメーションによるエージェント[6]〜[8]は，画像生成の負荷は軽いが自然感に欠ける．逆にテクスチャマップ[9]によるエージェント[4],[10]はユーザに自然な印象を与えるが，その生成には高い計算コストを必要とする．

VSAでは，約400の頂点と約400の面からなるディフォーマブルな3次元頭部ワイヤフレームモデルに実際の人物の顔写真をテクスチャマッピングすることにより，自然感の高い顔画像を合成する．顔写真は，人物を真正面から写したものを1枚だけ使用する．頭部ワイヤフレームモデルは顔の前面部だけで後頭部はなく，後ろを向くことはできないが，ユーザインタフェースとして使用するには十分である．ワイヤフレームモデルの頂点の位置を動かしてその形を部分的に変形することにより，瞬きをしたり，規則音声合成装置と連動して喋るように口を開くといったさまざまな顔を1枚の写真から合成する．

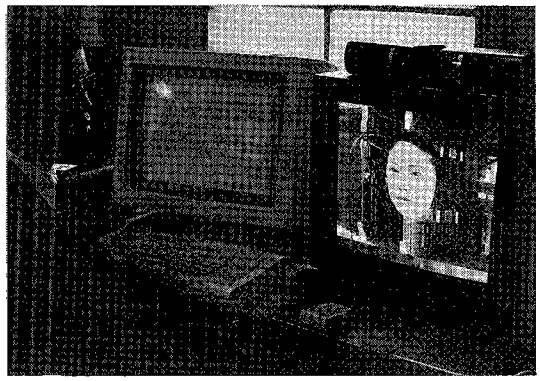

図1 ビジュアルソフトウェアエージェント
Fig. 1 Visual Software Agent.

VSAの実行例を，図1に示す．コンソール画面内の左側の画像は，テクスチャマップのもとになる正面顔画像である．同右側は，3次元ワイヤフレームモデルを顔画像にオーバラップし，その一部を拡大した画像である．合成された顔画像は右側のディスプレイに表示されている．背景は，ViSAシステムである．ユーザはヘッドセットを装着し，音声でVSAに対して質問したり，仕事を依頼したりすることができる．

2.2 揺らぐ顔

顔画像の合成については，これまでにも多くの研究がなされてきた．喜び，悲しみ，怒りなど，表情を変化させることのできる顔が合成されている．しかし，それらの多くは顔自体が静止していた．また顔の向きが変わる場合でも，あらかじめ決められた軌道の上を非常に滑らかに動くため，人工的であった．生きている人間の顔は微妙に動いており，ずっと静止したままであることはない．そこで我々は，顔自体に揺らぎを加えた．

顔の揺らぎは，次の3種類の組合せで実現している．

- 超音波距離計で測定したユーザの位置による揺らぎ

本システムではユーザの横位置に超音波距離計を設置

し，画面に並行にユーザの位置を検出している．これに連動してVSAは常にユーザの方向を向き，ユーザとの視線一致を実現している．距離計の分解能は約1cmである．ユーザが体を揺らしたり腕を動かしたりすると，それにつれて合成した顔が水平方向に揺らぐ．

- 口の開閉に伴う揺らぎ

発話に合わせて口を開閉するときに，顔の向きを少し水平・垂直方向に変化させる．

- ランダムな間隔・強さの揺らぎ

ユーザが全く動かず，口を閉じたままの状態でも，ランダムな間隔で顔の向きを0〜3度，水平・垂直方向に変化させる．

これにより，静止画では得られない自然感を与えることができる．

2.3 並列画像生成

我々の擬人化エージェントインタフェースでは，自然感の高い顔画像を実時間で生成することが必要である．しかし，その顔はいつも自然に揺らいでいるので，細かい部分を精密に合成することには意味がない．ワイヤフレームモデルの変形やテクスチャマッピングなどの実際の操作は，小規模並列ビジュアルコンピューティングシステムViSA上で実行される．ViSAは，要素プロセッサとして32ビットマイクロプロセッサ・トランスピュータT801（Inmos社製）を使用している．現在は4個のプロセッサを使用しており，並列に動作する．アセンブラレベルでの最適化を繰り返すことにより，毎秒10〜15フレームの画像生成速度を達成している．

3. VSAとMosaicの結合

3.1 システム構成

プロトタイプシステムの構成を，図2に示す．

VSAの顔画像を実時間で生成するViSAシステムが，Sunワークステーションに接続されている．顔の向きや口の形などは，Sunワークステーション上に実装されたViSAサーバが決定する．ViSAサーバはTCP/IPにより接続された他のワークステーションからの画像生成要求を受け付け，1フレーム当り40バイトの制御パケットをViSAシステムに転送する．合成された自然感の高い顔画像は，外部ディスプレイに表示される．ワークステーションのライブビデオ入力を通して，X-windowの一つのウィンドウ上に表示することもできる．

またユーザとの音声によるコミュニケーションを実現するために，規則音声合成装置，音声認識装置が接続されている．音声に関しては，市販の装置を利用した．音声認識装置は，不特定話者連続音声認識が可能である．あらかじめユーザの音声を登録する必要がなく，文章を認識することができる．

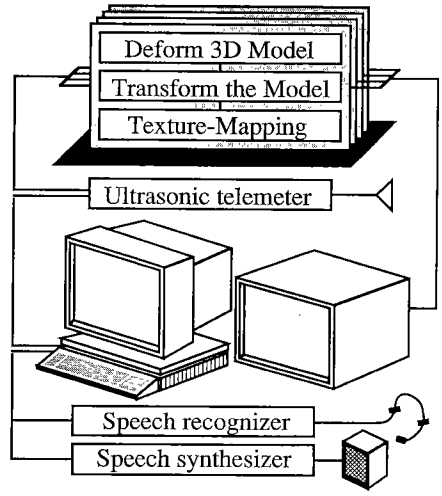

図2 システム構成
Fig. 2 System configuration.

3.2 VSA–Mosaic インタフェース

Mosaic [11] は，NCSA(National Center for Supercomputing Applications)が開発したWWW用のブラウザであり，大部分の操作を統一的にマウスで行うことができるように設計されている．Mosaicでは，ハイライト表示の文字列あるいはイメージ（これらをアンカーと呼ぶ）をマウスでクリックすることにより，そのアンカーにリンクされた情報を呼び出すことができる．

VSA–Mosaicインタフェースは，二つの視点から捉えることができる．

Mosaicを中心として見た場合には，VSAはWWW/Mosaicとユーザの中間に位置し，マウスの代わりに音声で必要な情報を引き出すためのインタフェースエージェントとして振る舞う．ユーザと自然なコミュニケーションをとり，ユーザの要求をMosaicに伝達する．マウスの操作が困難な計算機に不慣れな人や身体的ハンデのある人にも有効である．

またVSAを中心として見た場合，WWW/Mosaicは擬人化エージェントインタフェースのための分散型大規模情報データベースと考えられる．VSAは適当なURL (Uniform Resource Locator)を直接指定し

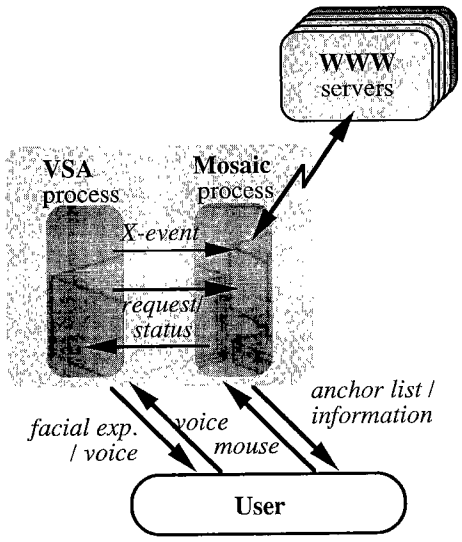

図3 VSA–Mosaicインタフェース
Fig. 3 VSA–Mosaic interface.

たり，リンクをたどることにより，必要なデータを引き出す．擬人化エージェントシステムVSAで提供する情報を特殊な形態で記述することなく，WWW/Mosaicで標準的に使用されているハイパテキスト記述言語HTML(Hyper-Text Markup Language)で記述できる点も，応用を広げる上で大きな利点となる．

VSA–Mosaicインタフェースは，二つの独立したプロセスから構成される．二つのプロセスの関係を，図3に示す．これは，できるだけもとのMosaicの機能に影響を与えないようにするためである．これにより，改造されたMosaicは，VSAプロセスとの接続に失敗した場合でも，通常のMosaicとして実行できる．また二つのプロセスを，異なるワークステーション上で実行することも可能である．

マウス操作と音声操作は，完全に同等に扱われる．従って両者を自由に混在させることもできる．音声認識装置がユーザの発話からキーとなる文字列の一つを検出すると，VSAプロセスはXイベントを生成し，それをMosaicプロセスに送る．すなわちXイベントは，Mosaicプロセスに対してユーザから何らかの処理要求があったことを通知する．その後，VSAプロセスとMosaicプロセスの間で，バッファを介して処理要求の種類や実行状態などの情報が交換される．

VSAが認識するコマンドは，大きく三つのタイプに分類できる．

- キー文字列によるセレクション

VSAインタフェースは，内部にキー文字列テーブルをもっている．キー文字列テーブルは，ユーザの発話から抽出されたキー文字列と，Mosaicのアンカー文字列あるいはURL (Uniform Resource Locator) との対応表である．ユーザの発話により，アンカー文字列あるいはURLがMosaicに送られる．キー文字列は単語である必要はなく，文章でもよい．この文字列変換には，最長部分一致検索が使われる．例えば“本郷のキャンパス”は，“本郷”や“キャンパス”に優先する．

キー文字列に対して同じ文字列をアンカー文字列として登録すると，アンカー文字列を発話することにより，そのアンカーが選択されることになる．

キー文字列に対してURLを登録しておくと，「○○○について知りたい」とか「○○○サーバにつないで下さい」と発話した場合に，指定されたサーバに接続される．

- インデックス番号によるセレクション

音声認識装置の性能による制限のために，音声対話によりマウスのもつ機能を完全に置き換えることはできない．我々の使用している音声認識装置は不特定話者連続音声認識が可能であるが，あらかじめ想定していない文章を認識できるわけではない．ユーザの自由な発話を，テキストに変換できるほどの能力はない．WWWサーバのリンクは常時変化していくので，ユーザが検索する可能性のあるすべてのアンカー文字列をあらかじめ登録しておくことは不可能である．更に，例えば画像のようなアンカーはアンカー文字列をもたない．そこでVSAは新しいページがオープンされるごとに，そのページに含まれるアンカーの一覧表（アンカーリスト）をユーザに提示する．アンカーリストは，Mosaic本体とは別のウィンドウに表示される．それぞれのアンカー文字列にはインデックス番号が付けられている．ユーザはアンカー文字列を発話する代わりに，インデックス番号を発話することにより任意のアンカーを選択することができる．

- 予約語によるページ制御

このコマンドはアンカーを選択するのではなく，ページ間やページ内の移動など，あらかじめ決められた方法でページを制御する．Mosaicでは，ページ画面下にあるボタンやスクロールバーの機能に相当する．これらは予約語として扱われる．[ ]は，省略可能であることを表す．

– ホームページ [に戻る]

- 次 [のページ] [{ に進む | が見たい }]
  (ページ間の移動)
- 前 [のページ] [{ に戻る | が見たい }]
  (ページ間の移動)
- { 上 | 下 } [が見たい]
  (ページ内の移動)
- その他

3.3 実 行 例

VSA–Mosaicインタフェースシステムの実行例を示す.

User: 「東京大学サーバにつないで下さい.」
VSA: 「はい. お待ち下さい.」
(指定されたサーバと接続し, アンカーリストを提示する.)
「はい. 五つのアンカーがあります.」
User: 「0番をお願いします.」
(0: Japanese version)
VSA: ("*Japanese*" ページをオープンし, 新しいアンカーリストを提示する.)
「はい. 五つのアンカーがあります.」
User: 「キ????プ を見せて下さい.」
(音声認識に失敗した.)
VSA: 「もう一度, お願いします.」
User: 「キャンパスマップを見せて下さい.」
VSA: ("*campus map*" ページをオープンし, 新しいアンカーリストを提示する.)
「はい. 三つのアンカーがあります.」
...

音声認識装置は, 発話内容に最も近いと思われるテキストとその採点スコアを返す. 採点スコアは0 (全く一致しない) から999 (完全に一致する) までの値をとる. 採点スコアが経験的に決められたしきい値に達しなかった場合には, VSAはそのテキストを放棄し, ユーザに対して再入力を促す.

図4は, VSA–Mosaicインタフェースの実行画面例である. X–window上に, 三つの画面が表示される. 右側のウィンドウは, VSAとのインタフェースをもつMosaicである. 外見的にはオリジナルのMosaicと変わりない. 左上のウィンドウには, 擬人化エージェントVSAが表示される. 左下のウィンドウにはアンカーリストが表示される. 2本の横線の間が, 1ページ分のアンカーである.

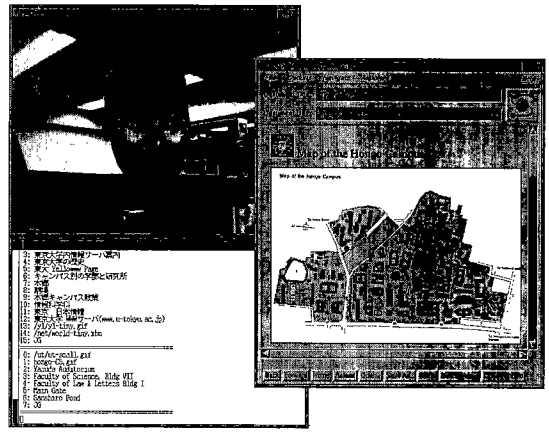

図4 VSA–Mosaicインタフェース実行例
Fig. 4 An example of VSA–Mosaic interface.

4. WWWとの接続における問題点

WWWでは, イメージや音声も含めたハイパテキストの交換のための記述言語HTML(Hyper-Text Markup Language)を定めている. しかし, それ自身が音声により操作されることを想定したデータ作成のガイドラインは示されていない. このため, 音声対話でWWWの情報を引き出す場合に, いくつかの問題が生じる.

(1) 位置に依存するアンカー
アンカーの文字列よりも, そのアンカーの存在する位置の方が重要な意味をもつ場合がある. 次に挙げるのは, 単純ではあるが非常に良い例である.

"XXXX は ここ"

下線で示した単語("ここ")がアンカーである. 我々はMosaic画面上でこのような書き方をしばしば見かける. ハイパーテキスト型のシステムではこの書き方は直観的であり, 有効であるかもしれない. これに対して, 音声により制御するシステムで上記の例を扱う場合を考える. 確かに「"ここ"をお願いします」と発話すれば意図したとおりの結果が得られるかもしれないが, 本人以外には全く意味が通じない.

(2) 同名のアンカー
HTMLでは, 一つのページの中で複数のアンカーが同じ名前をもつことを許している. しかも, それらが同じリンクをもっていても, 異なるリンクをもっていても構わない. これらは, その位置により区別が可能だからである.

マウスはポインティングデバイスであり，容易に任意の位置を指定できる．これに対して音声では，特徴的な名前をもつ目印がある場合を除いて，位置を指定することは難しい．

（3） 非文字列のアンカー

画像や音声のように非文字列のアンカーを扱えることはHTMLの優れた特徴の一つである．しかし，これらを音声で指定することは困難である．例えば画像の場合，アンカー文字列の代わりにファイル名をもつ．但し，このファイル名がその画像の内容を表しているとは限らない．

（4） 多過ぎるアンカー

HTMLでは，リンクにより必要なデータを容易に結び付けることができる．その反面，一つのページに100以上のアンカーが存在することも珍しくない．この場合，VSAが提示するアンカーリスト自体が画面内に収まらないことが考えられる．

これらの問題は，データの作成の仕方に大きく依存している．(1)や(4)は，データ作成者が注意を払うことで回避できる．(2)については，一つのページ内で同名のアンカーが常に同じリンクをもつことが保証されていれば，音声でも意図した通りのリンクをアクセスできる．(3)については，例えばHTMLのコメントタグを利用して，画像や音声データの内容について画面には表示されないコメントを付加し，それを検索する等の方法が考えられる．

プロトタイプシステムでは，上記のような問題点はあるものの，擬人化エージェントを介する音声対話によりMosaicを制御して必要な情報を引き出すことができることを確認した．

あまり計算機を使用していないユーザにとって，自然感の高い顔はユーザの注意を画面に強く引きつける効果がある．また計算機という「箱」に向かって話をするのではなく，話しかける対象が存在することで，ユーザの心理的抵抗感が軽減される．これに対して計算機を日常的に使用しているユーザは，自然感の高い顔や連続音声認識にあまり関心を示さない．これは現在の音声認識技術のレベルがまだ十分ではなく，キーボードやマウスの方が安定して入力できるためと考えられる．しかし音声は独立したコミュニケーション手段であり，例えばキーボードで文章を作成しながら，同時に音声で情報検索をすることもできる．また音声認識において十分な信頼性が得られるならば，複雑な処理をマウスを用いるよりも簡単に実行できる可能性がある．擬人化エージェントインタフェースを十分に活用するためには，音声認識装置の性能向上が必要である．

最近はインタネット上から自動的にアンカーとWWWページの対応を収拾してデータベース化するための Robot [12] と呼ばれるソフトウェアもいくつか開発されており，今後，WWWデータをうまく利用することがますます重要になると考えられる．

5. むすび

本論文では，ビジュアルソフトウェアエージェントVSAとWWWクライアントソフトウェアMosaicを結合したシステムを実装した．擬人化エージェントインタフェースのための大規模情報ベースとして，インタネット上のWWWサーバをアクセスする．通常のWWW用に作成されたデータをそのままの形で擬人化エージェントから利用することができる．但しデータの作成にあたっては，音声対話により操作されることを考慮する必要がある．

プロトタイプシステムでは，ビジュアルソフトウェアエージェントとユーザとの簡単な音声対話によってMosaicを制御し，WWWサーバから必要な情報を取り出すことができることを示した．これは，マウスの操作が困難な計算機に不慣れな人や身体的ハンデのある人にも有効である．音声は我々人間の日常的コミュニケーション手段の中心である．しかし，音声単独では空間的な位置を正確に指し示すことができない．また現在の音声認識装置の性能による制約から，マウスのようなポインティングデバイスを完全に置き換えることはできない．

現在はMosaicからアンカーリストのみを自動的に取り出し，それ以外の情報は，そのままユーザに提示している．今後，Mosaicからのすべての情報をトラップしてその内容を理解／学習し，ユーザの行動を積極的に支援できるような知的なインタフェースエージェントの実現が望まれる．

謝辞 本研究の一部は，文部省科学研究費補助金試験研究(B)(2)（課題番号06558045）ならびに奨励研究(A)（課題番号07780234）による．

文　献

[1] B.Laurel, "Interface agents: metaphors with character," in The Art of Human-Computer Interface Design, ed. B.Laurel, pp. 355-365, Addison-Wesley Publishing Company, Inc., 1990.

[2] M.Ishizuka, O.Hasegawa, W.Wiwat, C.W.Lee, and

H.Dohi, "Visual Software Agent (VSA) built on Transputer Network with Visual Interface (TN-VIT)," Int'l Symp. Computer World '91, pp. 36–46, 1991.

[3] O.Hasegawa, C-W.Lee, W.Wongwarawipat, and M. Ishizuka, "Realtime synthesis of human-like agent in response to user's moving image," 11th Int'l Conf. Pattern Recognition, IV, pp. 39–42, 1992.

[4] H.Dohi and M.Ishizuka, "Realtime synthesis of a realistic anthropomorphous agent toward advanced human–computer interaction," in Human–Computer Interaction: Software and Hardware Interfaces, eds. G. Salvendy and M. Smith, pp. 152–157, Elsevier, 1993.

[5] T.Berners-Lee, R.Cailliau, A.Luotonen, H.F.Nielsen, and A.Secret, "The World–Wide Web," CACM, vol.37, no.8, pp. 76–82, 1994.

[6] N.I.Badler, C.B.Phillips, and B.L.Webber, "Simulating Humans –Computer Graphics Animation and Control," Oxford University Press, 1993.

[7] 竹林洋一, "音声自由対話システム TOSBURG II —ユーザ中心のマルチモーダルインタフェースの実現に向けて," 信学論(D–II), vol. J77–D–II, no.8, pp. 1417–1428, 1994.

[8] E.Gasper, "Multimedia man-machine interface using anthropomorphic agents," Int'l Symp. Computer World '91, pp. 105–109, 1991.

[9] P. Heckbert, "Survey of texture mapping," IEEE CG&A, vol. 6, no.11, pp. 56–67, 1986.

[10] A.Takeuchi and K.Nagao, "Communicative facial displays as a new conversational modality," ACM/IFIP INTERCHI'93, pp. 187–193, 1993.

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

[12] "List of Robots,"
URL: http://web.nexor.co.uk/mak/doc/robots/active.html.

(平成7年8月20日受付，12月1日再受付)

土肥　浩

昭60慶大・理工・電気卒．昭62同大大学院修士課程了．同年東大生産技術研究所勤務．平5工学部電子情報工学科助手．擬人化エージェントによるヒューマンインタフェース，並列画像処理の研究に従事．ACM，情報処理学会各会員．

石塚　満　(正員)

昭46東大・工・電子卒．昭51同大大学院博士課程了．工博．同年NTT横須賀研究所勤務．昭53東大生産技術研究所助教授．平4工学部電子情報工学科教授．人工知能，知識システム，画像理解，擬人化エージェントによるヒューマンインタフェースの研究に従事．IEEE, AAAI, 情報処理学会，人工知能学会，画像電子学会各会員．