Face-to-face 型擬人化エージェント・インタフェースの構築

土肥 浩† 石塚 満†

本論文では，自然感の高い顔と音声対話能力を備えたface-to-face型の統一的な擬人化エージェントインタフェースを実現する；ビジュアルソフトウェアエージェント（VSA: Visual Software Agent）の新しいプロトタイプの構成について述べる．新しいプロトタイプでは，顔画像，音声合成，音声認識，WWWブラウザ制御などをそれぞれ機能ごとにモジュール化した．これにより，TCP/IPネットワークで結合された複数のWSで構成される分散環境下で，7つの機能モジュールが協調的に動作する擬人化エージェントインタフェースを実現した．それぞれの機能モジュールは，複数のWS上で並列あるいは並行に動作する．新しい機能モジュールを動的に追加・削除したり，更新したりすることが容易に可能である．ユーザはテレビ電話を使うようにエージェントとの簡単な音声対話により，電子メイルの到着や時刻を尋ねたり，WWWブラウザの操作を依頼したりすることができる．またVSAシステムはWWW上で常時更新されていく多種多様の情報を人手を介さずに自動的に収集し，音声で答えられるように変換する．これによりユーザは同じ音声対話インタフェースを用いて天気予報やニュースなど，知りたいときに知りたい情報を得ることができる．

An Implementation of the Visual Software Agent Interface System Connected with WWW/Netscape

HIROSHI DOHI† and MITSURU ISHIZUKA†

This paper describes the design and the implementation of the Visual Software Agent (VSA) interface system connected with a voice-controlled Netscape. VSA is a test bed for an Internet-based interface agent with an unified face-to-face style interface. Besides the texture-mapped rocking realistic face, it equips a camera, a speech recognizer, a speech synthesizer, and some sensors; corresponding with eyes, ears, a mouth, and some sense organs, respectively. A user asks the realistic anthropomorphic agent some questions and orders some tasks, as if he/she usually asks his/her colleagues with a visual phone. It embodies a human-like face-to-face style communication environment, and therefore it will reduce the mental barrier for communication lying between a human and a computer. The system is also connected with World Wide Web (WWW)/the Netscape navigator.

1. はじめに

情報端末機器の小型化，高性能化により，コンピュータの使われる場面は飛躍的に拡大し，多様化している．現在のコンピュータインタフェースは，マウスに代表されるポインティングデバイスを使用したグラフィカルユーザインタフェース（GUI）が主流になっている．マウスは，難解な操作コマンド名を覚えたりキーボードに触れたりすることなく，簡単なボタン操作だけでコンピュータを扱うことができる点で，ユーザインタフェースに大きな役割を果たしている．その一方で操作が直接的（direct manipulation）であることから，比較的単純なタスクしか選択できない．また，すべての操作をマウスだけで統一的に行えるように設計されたシステムでは，たとえばダブルクリックのような基本操作がうまくできない人にとってはほかに選択肢がなく，逆に使いにくいものになるなど，必ずしも万能のインタフェースであるとはいえない．

良いインタフェースとは，ユーザが操作方法を覚えたり練習したりすることなく，単純かつ自然な操作方法で複雑なタスクを実行できるものであろう．我々は，従来のマウスやキーボードによる操作に加えて，簡単な音声対話インタフェースなど，複数のインタフェースチャネルを備えたビジュアルソフトウェアエージェントVSAの研究を進めている．マウス/キーボードによる操作と音声による操作は同じ優先度を持つため，ユーザは状況や環境にあわせて，いつでも自然で最適なインタフェースチャネルを自由に選択できる．

VSAの実行例を，**図1**に示す．

† 東京大学工学系研究科電子情報工学専攻
Department of Information and Communication Engineering, School of Engineering, University of Tokyo

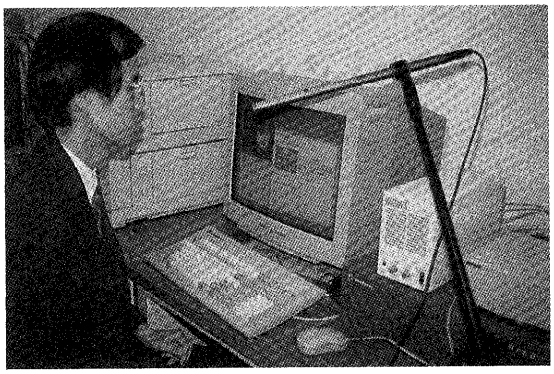


図1 VSA インタフェース
Fig. 1 An overview of the VSA interface.

VSA システムでは音声対話インタフェースに，一般の人にもよく知られているテレビ電話のような face-to-face 型（対面型）コミュニケーションスタイルを採用した．人と話をしているときに，用もないのに別の方向を向いたまま返事をされたりすると，不快になる．逆に，相手の顔や姿が見えると安心できるというのは，だれもが経験することである．そこでコンピュータという箱に向かって話をするという心理的抵抗感を和らげるために，実際の顔写真を基にした自然感の高い顔画像を CG 合成している．この顔は自然に揺らぎ，瞬きしたり，おしゃべりしたり，ユーザの姿を追いかけたりする．

VSA は World Wide Web（WWW）/Netscape とも接続されている．ユーザは，VSA との簡単な音声対話を通じて Netscape を操作し，インターネット情報空間から必要な情報を得ることができる．ユーザの音声指示とは独立に自発的に WWW サーバにアクセスして，天気予報・降水確率やプロ野球/J リーグ/大相撲の結果などのデータを収集し，音声応答用ローカルデータベースの内容を人手を介さず自動的に更新するサブエージェントを実装した．テレビやラジオはあらかじめ決められた時間にニュースや天気予報を放送している．これに対して VSA は限定的ではあるが，ユーザが知りたい最新の情報を，知りたいときに教えてくれる．

2. WWW ブラウザと結合した VSA インタフェース

2.1 擬人化エージェント

アニメーションキャラクタによる Presentation Agent や Life-like Agent と呼ばれるインタフェースがいろいろと開発されてきている．たとえば，Microsoft Agent[5] は，Windows 用に開発されたエージェントインタフェースである．多数の小さな画像ファイルを連続的に再生しているのであるが，愉快なキャラクタが音に合わせて動き，視覚的な効果は抜群である．

VSA システムでは，テレビ電話のような face-to-face 型擬人化エージェントインタフェースの実現を目指している．このためアニメーションキャラクタではなく，3 次元頭部ワイヤフレームモデルに実際の顔写真をテクスチャマッピングし，それをリアルタイムで動かすことで，より自然感の高い顔画像を持つ擬人化エージェントを実現している[1]．

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．3 次元頭部ワイヤフレームモデルは約 500 の頂点で構成されており，男性/女性に関係なく，スケールを変えることで顔写真に合わせる．頭部ワイヤフレームモデルと顔写真とのマッチングを行うために，VSA editor を開発した．画面に表示された顔写真の上で右目，左目，口の 3 点をクリックするだけで，顔写真からモデルに対応するテクスチャが切り出され，即座に自然に揺らぐ顔画像が生成できる．

2.2 Netscape との接続

適切なデータベースを用意することができれば，音声による簡単な質問を受けて，それに応答するインタラクションシステムを作ること自体はそれほど難しくない．しかし，コンテンツの少ないシステムは魅力に乏しい．そのデータベースの内容，すなわち多様なコンテンツをどのようにして集め，更新していくかが問題となる．ユーザがそれぞれのシステム独自のデータフォーマットでコンテンツを記述していくことは一般に困難である．たとえば，パーソナル（個人向け）アシスタントシステムにおいて，天気予報などのようにデータが時々刻々変化していく場合，ユーザがローカルデータベースの内容を自分で更新することは考えられない．

インターネット上の WWW では多くのユーザにより多種多様の情報が提供され，その内容は常時更新されている．したがって，WWW を VSA のための分散型大規模情報データベースとして利用することができれば，その応用範囲が飛躍的に広がる．ユーザはハイパーテキスト記述言語 HTML をはじめ，WWW で標準的に使用されているフォーマットでデータを記述できるだけでなく，これまでに WWW 上で蓄積されてきた膨大なマルチメディアデータをそのままの形で利用することが可能になるからである．

我々は，まず VSA をインターネット上の WWW ブラウザ NCSA Mosaic と結合した[2),8)]．その後，基本的な外観を受け継ぎ，Netscape と結合した VSA を

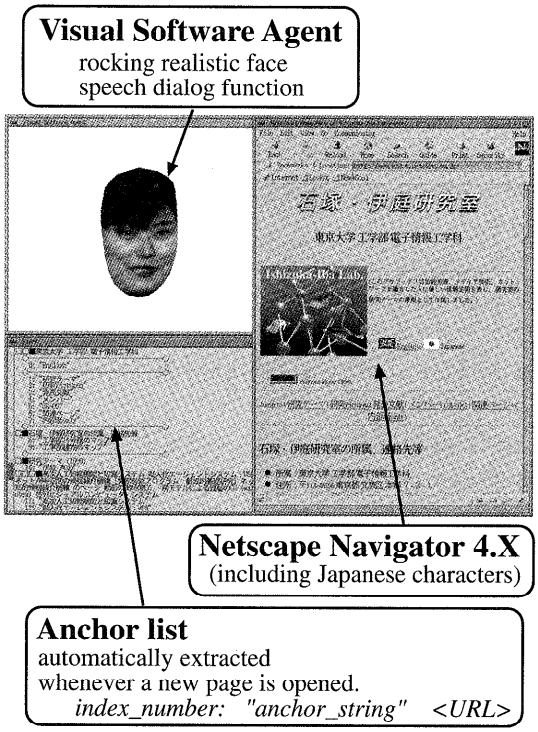


図2 VSA-Netscape インタフェース
Fig. 2 VSA-Netscape interface.

実現した．マルチメディアデータは，Netscape の画面を介してユーザに提示される．

Netscape と接続した VSA インタフェースの画面例を図2に示す．右画面が Netscape，左上画面が擬人化エージェント VSA，左下画面は各 Web ページから抽出したアンカーリストである．アンカーリストは，その時点で右側の Netscape に表示されているページに含まれるアンカーの一覧表で，新しいページがオープンされるごとに自動的に抽出，更新される．ユーザは従来のマウス操作に加えて，擬人化エージェント VSA との簡単な音声対話（たとえば，アンカーをクリックする代わりに，その文字列を発話する）によって Netscape や 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．

3. 実　　装

3.1 システム構成

システム構成を図3に示す．

Netscape 接続版 VSA は，基本的に7つのプロセスで構成されている．

- Proxy プロセス
  Netscape で新しいページがオープンされるごとにアンカーリストを作成し，音声–URL 変換プロセスに送る．
- Netscape
  ハイパーメディアデータを表示する．普通の Netscape である．
- 音声認識プロセス
  不特定話者連続音声認識装置を制御し，発話内容に最も近いと思われるテキストを音声–URL 変換プロセスに送る．
- 音声–URL 変換プロセス
  発話されたテキストとアンカーリストを比較して Netscape を制御したり，質問に対する応答テキストを生成したりする．
- 音声合成プロセス
  応答テキストから口形状の時系列データを生成し，合成音声と同期するように顔画像生成を制御する．
- VSA 画像生成プロセス
  ランダムに瞬きをしたり，揺らぎのある顔画像を生成する．合成音声に合わせて口形状を変化させたり，各種センサ情報に基づいて視線一致のためにユーザの方を向かせたりする．
- センサプロセス
  電子メイルの到着を確認したり，室温などの環境情報を収集して，ユーザの質問に答えるためのデータを提供する．また焦電型赤外線センサにより人体の検知を行い，超音波距離センサにより VSA とユーザ間の距離を計測する．たとえば，ユーザが近づいてくると「こんにちは」と挨拶し，立ち去る際には「さようなら」と挨拶させることがで

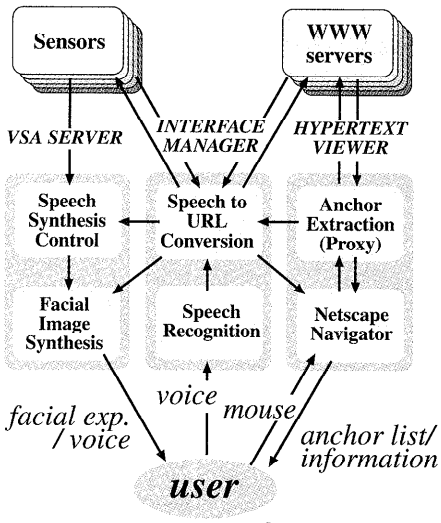


図3 システム構成
Fig. 3 System configuration.

きる．

3.2 分散協調処理

Netscape接続版VSAでは，それぞれの機能単位を1つのプロセスとして実装した．TCP/IPネットワークにより接続された複数のWS上でプロセス間通信（一部，共有メモリ）により結合され，協調的に動作する．ハードウェアの制約はあるが，それぞれのプロセスは同じ計算機上で実行してもよいし，負荷を分散させるなどのためにネットワークで接続された異なる計算機上で実行してもよい．人間にとって，何かを目で追いかけながら，しゃべり，同時に手を動かすことは容易であるように，それぞれの機能モジュールは並列，あるいは並行に動作する（現在，我々のシステムは，音声認識処理用としてSun WS 1台，顔画像合成，その他の処理用として，構成によりSGI WS 1〜3台を同時に使って実行している）．

各プロセスは，複数のプロセスからのコネクション要求を同時に処理することができるように実装されている．これにより，必要に応じて新しいセンサを動的に付加してその情報を受け取ったりすることが容易にできる．このモジュール化により，ハードウェア構成や動作環境に応じて最適なモジュールを選択し，交換することが可能になる．

3.3 合成音声と口形状の同期

"DECface"[7]では，英語音声合成ソフトウェアDECtalk（音声合成装置DECtalkの合成アルゴリズムをWSに移植したもの）を用いて，音声合成と顔画像生成を1つのWS上で同時に実行している．これにより，テクスチャマッピングした顔の唇を，合成音声に合わせて高い精度で動かすことができる．

VSAでは市販の規則音声合成装置をRS-232-Cで外部接続しているため，原理的にDECfaceのような精度の高い同期は実現できない．代わりに，まず音声合成装置に発話テキストを送り，発話開始までの時間と発話速度を考慮して，テキストから抽出した口形状列/発音記号列に従ってテキスチャマップした顔の口形状を変化させる．規則音声合成装置として，日本語用に「しゃべりん坊」(NTTデータ)，英語用にDECtalk（DEC）を接続している．

日本語の発話テキストは漢字の読み誤りのないように，また口形状を自動抽出するために，平仮名あるいはカタカナで記述する．さらに，アクセント位置と単語間のポーズ長を指定する．日本語は高低アクセントであり，アクセント位置とは語の途中で音が低くなる位置のことをいう．VSAは，日本語テキスト用として「あ」「い」「う」「え」「お」「ん」の口形状を表現できる．音声合成プロセスは発話テキストを音声合成装置に渡す前に先読みして，テキストに対応する口形状列データを生成する．

英語の発話テキストは，普通の英文テキストである．英語の場合には単語境界がはっきりしているので，音素辞書を検索して約40種類の発音記号列データに変換する．

UNIXオペレーティングシステムではリアルタイム性が保証されていないため，顔画像生成が音声合成に対して大きく遅れる場合には，共有メモリを用いて途中の画像をドロップさせて同期を確保している．

3.4 音声認識

音声認識には，市販の不特定話者連続音声認識装置DS-200（オージス総研）を用いている．日本語/英語（切替え）に対応している．この装置は，最も可能性の高いテキストとともに採点スコアを返す．スコアがしきい値に達しない場合は，エージェントがユーザに再入力を促す．ある得点以上のスコアを持つ場合には，音声認識が正しく行われたものと見なす．音声認識部分はモジュール化しているため，音声認識などの結果としてテキストを返せるものであれば何でもよい．

3.5 アンカーの抽出

Mosaicは約6万行のソースコードが公開されていたため，Mosaic内部に手を加えることによりVSAと結合し，アンカーリストを抽出していた．この機能は，論争支援マルチモーダル実験システムMr.Bengoでも使われている[9]．これに対してNetscape ver. 4.X以前のソースコードは公開されていないため，Proxyプロセスを WWWサーバとNetscapeの間に挟み込み，そこを通過するデータを監視することで，同等の機能を実現した．

Proxyを利用してアンカー抽出を行う場合，WWWサーバから送られてくるhtmlファイルを2度，解釈することになる．すなわち1度はProxyで，そしてもう1度はNetscape内部で解釈される．内部を解析したわけではないが，Netscapeは判断可能なものについては文法的に軽微な誤りがあってもそれらしく表示する場合がある．そのため，文法上の誤りにより，VSAのProxyでは正しいアンカーが抽出できないにもかかわらず，Netscape画面上では正しく表示される場合がある．逆に，Netscapeでは日本語の文字コードセットの自動判別機能が完全ではないため，Proxyでは日本語のアンカーを正しく抽出できるのに，Netscape画面上では文字化けする場合がある．

3.6 アンカーリスト

アンカーリストは，新しいWebページがオープン

されるごとに自動的に抽出，更新される．各エントリは，次の3つの項目からなる．インデックス番号は，ページごとの一連番号である．

＜インデックス番号，アンカー文字列，URL＞

画像アンカーの場合には，「*nn*:【画像】(*_URL_*)」と表示される．

アンカーリストは，2種類の表示方法を選択できる．1つは，フラットな一覧表示である．もう1つは，アンカーとともに横線（
）およびヘッダ文字列（<H*n*>行）が表示される．またヘッダのレベル n に応じて，アンカーおよびヘッダ文字列がインデントされる．画像アンカーの場合でも，ヘッダとの位置関係および画像のURLから，アンカーリストとページ上のアンカーの対応関係を得ることはそれほど難しくない．

4. 音声によるWWWブラウザの制御

4.1 音声コマンド

VSAは，日付や時刻，電子メイルの到着などの音声での問合せのほかに，インターネット情報空間の散策を支援する4種類の音声コマンドを認識する（[ ]は，省略可能であることを示す．また，これ以外の文字列に置き換えることもできる）．

_を見る:={ を見せて下さい | が見たい | ...}

- アンカー文字列によるセレクション
  【例】『○○ [_を見る]』
  アンカー文字列を発話すると，そのアンカーが選択される．マウスでアンカーをクリックするのに相当する．
- インデックス番号によるセレクション
  【例】『○番 [_を見る]』
  たとえば画像のようなアンカーはアンカー文字列を持たない．文字列の代わりにアンカーリストのインデックス番号を発話することによっても，任意のアンカーを選択できる．
- 登録したURLによるセレクション
  【例】『○○○サーバ [_を見る]』
  あらかじめキーワードとURLを登録しておくことで，発話により，そのURLにジャンプする．ブックマーク機能に相当する．
- 予約語によるページ制御
  【例】『ホームページ [に戻る | _を見る]』
  『次 [のページ] [に進む | _を見る]』
  『前 [のページ] [に戻る | _を見る]』
  （ページ間の移動）
  『上 | 下 [_を見る]』
  （ページ内の移動）など
  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．

マウスによる操作と音声による操作は，完全に同じレベルで扱われる．騒音下など音声入出力インタフェースが適さない場合でも，従来どおりマウスやキーボードを使ってNetscapeを普通に実行できる．マウス操作と音声操作を自由に混在させても構わない．したがって，ユーザは状況に応じていつでも，より適切なインタフェースを選択することができる．このインタフェースは，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．

4.2 拡張（アンカー）リスト

拡張（アンカー）リストは，アンカーになっていないキーワードを特定のURLに関連づけたり，テキストを別の文字列に置き換えたりすることにより，仮想的にアンカーを拡張するものである．特定のWebページに対してのみ有効なもの（ローカル）と，すべてのWebページに対して有効なもの（グローバル）の2種類がある．たとえば，テキストの「日本語」を「Japanese」に置き換えるように登録しておけば，「Japanese」というアンカーを含む英語のWebページで「日本語を見せて下さい」とエージェントに話しかけると，日本語のWebページに切り替わる．ただし拡張リストによる別の文字列への置き換えは，1度しか起こらない．

4.3 アンカー名の部分発話

近藤ら[10]が開発したJAM（Japanese Aware Multimedia）は，不特定話者連続音声認識を用いて，音声でWWWブラウザを制御するシステムである．JAMでは，Webページから抽出したアンカー名をまず音素列に変換し，次に音声認識用の文法を生成することにより，音声認識システムの語彙を動的に切り替えている．長いアンカー名は，先頭から3文節（デフォルト）読めばよい．

VSAシステムでは，アンカー名を「先頭から」読むという制限をなくし，アンカー名の任意の一部分を発話すればよいようにした．一部分の発話からではアンカーを一意に決定できない場合が生じるが，VSAは規則音声合成装置を備えており，ユーザに対して複数のアンカーが該当することを音声で知らせ，再入力を促す．このとき，アンカーリストも更新され，可能性のあるアンカーだけに絞り込まれる．

なお文献10)には，『(VSAシステムでは,) このシステム用に特にキーワードやインデックス-アンカー名対応表を埋め込んだWebページが必要である』と紹介されているが，これは誤解である．JAMと同様に，Webページからアンカー名とURLの対応表を動

的に自動抽出しており，アンカー名と発話キーワードの文字列マッチングをとっているので特別のページを用意する必要はない．もちろん必要があれば，拡張リストにより特別のキーワードを付加することは可能である．

4.4 アンカー文字列とのマッチング

音声認識プロセスから渡されるテキストと，**表1**の4つのリストとのマッチングをとることにより，次の動作が決定される．アンカー文字列と予約語が重なる場合があるので，マッチングの順序はテキストの終わり方により変化する．

- *Case A.* 特定の語句で終わる場合

テキストが，たとえば「を見せて下さい」「が見たい」などという語で終わっていれば，それより前の部分を検索キーワードとする．この場合，キーワードはまずアンカーとの部分一致が試みられる．マッチするものが1つだけの場合は，そのアンカーが選択され，対応するURLにジャンプする．マッチするものが複数ある場合には，候補となるアンカーのみで新しいアンカーリストを構成し，ユーザに対して再入力を促す．マッチするものがない場合には，拡張リスト（ローカル），同（グローバル）の順に完全一致が試みられる．拡張リストによりテキストが別の文字列に置き換えられた場合には，再度アンカーとの部分一致が試みられる．URLに関連づけられた場合には，そのURLにジャンプする．拡張リストともマッチしなかった場合，予約語リストとの完全一致が試みられる．いずれにもマッチしなかった場合，擬人化エージェントがユーザに対して，「その項目はありません」と音声で伝える．

- *Case B.* それ以外の場合

テキストが，たとえば「を見せて下さい」「が見たい」などという語で終わっていなければ，テキスト全体を検索キーワードとする．この場合，キーワードは，まず予約語リストとの完全一致が試みられる．予約語リストにマッチしなかった場合は，アンカーとの部分一致，続いて拡張リスト（ローカル），同（グローバル）の順に完全一致が試みられる．

この結果，たとえばあるWebページに「上」というアンカーが含まれる場合，「上を見せて下さい」と発話するとそのアンカー「上」に対応したURLにジャンプし，単に「上」と発話すれば，Netscape画面がスクロールアップする．もしWebページに「上」というアンカーが含まれていなければ，「上」「上を見せて下さい」のいずれも画面がスクロールアップする．

表1 マッチングの方法と順序
Table 1 Keyword matching.

リスト	検索方法	*Case A*	*Case B*
アンカーリスト	部分一致	1	2
拡張リスト（L）	完全一致	2	3
拡張リスト（G）	完全一致	3	4
予約語リスト	完全一致	4	1

5. サブエージェント

音声対話システムでは，ユーザの質問に対して遅滞なく反応（応答）できることが重要である．ユーザから質問を受けるごとにWWWサーバをアクセスしていたのでは，円滑に対話を進めることができない．ユーザが本当に必要としている情報はWebページの中のごく限られた一部分で，言葉にすれば実は一言で伝えられるという場合も少なくない．たとえば，「明日の天気が知りたい」という場合，それは全国各地の天気予報一覧が知りたいわけではなく，一般に自分のいる場所の天気予報や降水確率を知りたがっていると考えるのが自然である．WWWブラウザでは，文字列やイメージなどのアンカーをクリックすることにより，それにリンクされた情報を得ることができるように設計されている．ところがこれは，アンカーをクリックすればユーザが即座に必要な情報を入手できることを保証するものではない．アンカーのクリックは，WWWブラウザに対して，リンクされたデータをWWWサーバからダウンロードするように指示するだけである．ネットワークの状態によっては，そのデータが実際にユーザに提示されるまでには長い時間を必要とするかもしれない．細いネットワークで大量の画像データを転送し，長い時間をかけて全国各地の天気を美しく表示することがいつも有効であるとは限らない．

そこで，あらかじめ特定のWWWサーバにアクセスして，音声応答に必要なデータをそのWebページから切り出しておく．この作業をバックグラウンドで専門に行うプロセスを，サブエージェントと呼ぶ（**図4**）．サブエージェントは指定されたWebページからHTMLファイルをダウンロードし，特定のキーワードを含むパラグラフあるいは特定のデータを自動的に切り出すための一種の情報フィルタリングプログラムである．UNIXのcron（clock daemon）で定期的に起動されるほか，必要に応じてVSAからも随時起動される．データを定期的に更新しているWebページでは，それぞれ独自のページレイアウトや表記法を持っている．たとえば，プロ野球やJリーグの結果は，次のように表記されることが多い．

チームA　2－1　チームB

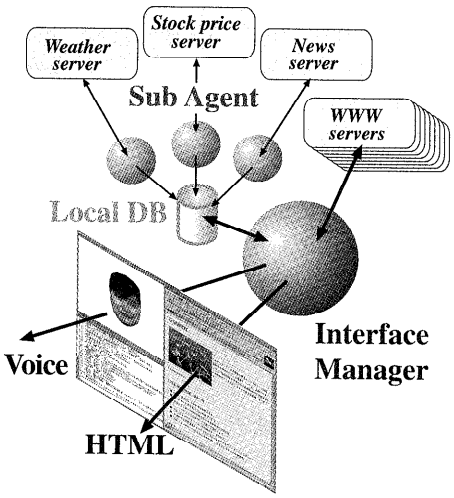


図4 サブエージェント
Fig. 4 Sub agent.

すべての内容を解析して分類することは大変であるが，このような限られたパターンのみを抽出することは非常に簡単である．天気予報などはインラインイメージを使ったテーブル形式で記述されていることが多いが，基本的な処理は同様である．切り出されたデータは，音声応答用ローカルデータベースに蓄積される．

6. 考　　察

6.1 エージェントの顔

インタフェースにおける顔の効果についてはいくつかの論文が報告されている．Walker ら[6]は，インタフェースに顔を用いると余計に時間がかかるだけで，顔を安易に用いるべきではないと指摘している．Takeuchi ら[4]は，インタフェースにおける顔はユーザの注意を強く引きつけるが，ユーザはその顔の表情から何かを理解しようとするので，本来の仕事への集中力を低下させるとしている．Koda ら[3]の研究でも，インタフェースにおける顔はユーザの注意を強く引きつけることが報告されている．Koda はまた，ユーザは擬人化することについて正反対の意見を持つ2つのグループに分かれることを示唆している．たとえば，一方のグループは人の顔よりも犬のキャラクタを好み，もう一方のグループは人の顔の方を好むといった具合である．

基本的に顔がユーザの注意を強く引きつける点では一致しているが，これらの評価が困難であるのは個人の能力，計算機操作の習熟度，作業環境などの要因が大きく影響するからである．また合成した顔画像，音声合成，不特定話者連続音声認識のいずれもがまだインタフェースに十分な性能に達していない状態で，顔の効果だけを評価することは難しい．

6.2 アニメーションの顔 vs. リアルな顔

アニメーションの顔は，画像生成の負荷は軽いが，自然感に欠ける．逆に，テクスチャマップによる顔は，自然な印象を与えるが，その生成には高い計算コストが必要である．従来，テクスチャマップによる顔をリアルタイムで動かすためには非常に高価なグラフィックス専用システムが必要であったが，計算機性能が劇的に向上した現在では，それほど特殊なシステムを必要としない．テクスチャマッピングをハードウェアで実行する専用回路を備えていれば，CPU 側にはほとんど負荷がかからない．このため，テクスチャマッピングの計算コストは問題にならなくなってきている．

「リアルな顔は，ユーザに過剰な期待を抱かせるので良くない（だからアニメーションの顔の方が良い）」という意見が大勢を占めているが，現在の音声認識/合成の能力では，実際に少し音声対話をしてみればユーザが過剰な期待をするとは思われず，わざわざ顔のクオリティだけを下げる意味はないかもしれない．

アニメーションでいろいろな顔を作るにはそれぞれデザインする必要があり，たとえば，それぞれの個人が自分の顔を作ることは難しい．これに対して VSA では1枚の写真からエージェントの顔を生成しており，エージェントの顔を変えるには基になる写真を交換するだけでよい．男性の顔も，女性の顔も，1つの同じワイヤフレームモデルのスケールを部分的に変えることによって実現されている．現在の VSA では顔だけを生成しているが，自然に揺れ動いたり瞬きをしたりする顔をしばらく見ていると，ユーザに「慣れ」が生じ，すぐに違和感が消えるようである．

アニメーションの顔とリアルな顔のどちらが優れているかについては，まさに Koda の指摘のように2つのグループに意見が分かれる．VSA システムでは人間同士の face-to-face 型コミュニケーションスタイルの実現を目標としており，WS の能力がさらに高くなれば将来的には擬人化エージェントシステムに，よりクオリティの高いリアルな顔画像が使われるようになると考えている．

6.3 音声による Netscape の操作

音声による操作は，使い方を覚えたり練習したりする必要がなく，だれでもが手軽に使える点で優れている．もし音声認識能力が完全であれば，キーボードよりも速い入力が可能であるといわれている．また独立したコミュニケーションチャネルとして，我々は日頃，キーボードで文章を入力しながらだれかに時刻を尋ね

るといったことも，無意識のうちに行っている．キーボードとマウスの関係のように，そのつど，手を動かして持ち替える煩わしさがない．

一方，音声の持つ性質として，マウスは容易に平面上の任意の位置を指定できるのに対し，音声では特徴的な名前を持つ目印がある場合を除いて位置を直接指定することが難しいなどの欠点がある．また技術的な課題として，現在ではまだ十分な音声認識性能が得られない．そのため，キーボードのように大量のテキストを安定して入力することには適していない．

たとえばNetscapeの場合，音声認識が完全に行えると仮定すると，基本操作の大部分は音声でも可能である．ステージ上で大型スクリーンを用いたプレゼンテーションを行う場合でも，マウスやキーボードを講演者の手元に置く必要がない．ただし実際には音声認識能力は十分ではなく，またWWW自体がもともと音声により操作されることを想定していないため，1つのページに同名のアンカー（たとえば，「ここ」をクリックなど）が複数存在したり，画像アンカーやURL（ファイル・パス），各種の記号を含んだテキストのように音声ではうまく指示できないといったことが発生する．日本語のテキストでは，漢字（同音異義語，異体字），平仮名（全角，半角），カタカナ（全角，半角），英字（全角，半角）が複雑に混在し，さらに送り仮名の使い方も一定していないなど，同じ音に対する表記法が多数存在するという問題もある．

VSAシステムではアンカー文字列の部分発話を可能にした．これにより，音声認識できる単語数に対するアンカーの選択の幅が大きく広がる（我々は市販の音声認識装置を使用しているので，音声認識できる単語数についての議論はしない）．アンカー中に各種の記号と単語が混在している場合でも，単語の部分だけを使って部分一致検索できる．VSAはユーザと簡単な音声対話をすることができるので，音声入力があいまいな場合や部分一致検索で，複数のアンカーにマッチして一意に選択できない場合には，候補を絞り込んだ新しいアンカーリストを提示したうえで，ユーザに対して再入力を促す．ユーザはアンカー文字列の代わりにインデックス番号を発話することにより，アンカーを選択することもできる．この方法では，画像アンカーや記号アンカーも含めて，アンカー文字列を持たない（発話できない）場合にも対応できる．

音声認識プロセスは独立しているため，より高性能な音声認識装置が登場した場合には容易に置き換えができる．我々が現在使用している音声認識装置は，あらかじめ発話される可能性のある構文と単語リストを定義しておく必要がある．しかし音声–URL変換プロセスには，音声認識プロセスに対する独立性を高めるために，音声認識可能な単語リストをいっさい与えていない．すなわち，音声認識装置が認識できるかぎりにおいて，どのような単語でもアンカーリストの部分一致検索の検索キーワードとなりうる．これは音声認識装置の性能向上に比例して，音声によるアンカー選択の幅が広がることを意味する．

VSAシステムでは音声入力は重要なコミュニケーションチャネルであるが，すべての操作を音声入力で完全に置き換えることを目指しているわけではない．ユーザが環境に応じて，その時点で最も適切なインタフェースを自由に選択できることが重要である．すなわち音声入力が適さない場合には，まったく同じプライオリティレベルで従来のマウスによる選択を自由に混在させることができる．

6.4 サブエージェント

VSAでは，現在の時刻や電子メイルの到着を尋ねるのと同じように，ニュースや天気予報を尋ねることができる．テレビ電話のような，統一的な音声対話インタフェースを実現している．

システムがデータの信頼性を保証できる場合，すなわちユーザがエージェントを十分信頼できるならば，エージェントはWebページをユーザに示さなくても，その内容の一部を音声で答えることができる．エージェントはユーザに対して必ずしもすべてのデータの出所を示す必要がない（もちろんユーザが要求すれば，エージェントはその根拠を答えなければならない）．そのデータがWWWから得られたものでも，CD-ROMから得られたものでも，あるいはそれ以外の方法で得られたものでも構わない．ところがサーチエンジンを利用する場合，エージェントは検索の仲介をすることはできるが，必ずしもその検索結果のデータの信頼性を保証できない．Webページの内容理解の研究も数多く行われているが，たとえばニュースや天気予報など，WWWから得たデータの信頼性をシステムが保証するためには，実際には特定のWebページしか利用できないことになる．また実用上，データを抽出するWebページを固定しても構わない場合が多い．

現在はトピックごとに特定のWebページをアクセスしており，それぞれのページレイアウトに依存したデータ抽出ルーチンが起動される．データを識別するためのタグを持つXMLフォーマットの普及が望まれる．

7. まとめ

本論文では，人間の日常的な face-to-face 型コミュニケーションスタイルを採用し，複数のインタフェースチャネルを備えたビジュアルソフトウェアエージェント VSA の構成について述べた．従来のマウス/キーボードに加え，日常的コミュニケーション手段である音声対話機能と，自然に揺らぎユーザと視線一致する顔画像を備えている．TCP/IP ネットワークで接続された複数の WS による分散環境下でも 7 つのプロセスが協調して，1 つの擬人化エージェントインタフェースとして動作することを確認した．

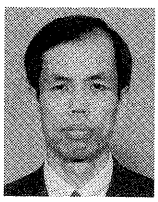
これによりユーザはテレビ電話を扱うような感覚で，エージェントに対して時刻や電子メイルの到着を尋ねたり，WWW ブラウザの操作を依頼したりできる．さらにユーザは同じ音声対話インタフェースを用いて，限定的ではあるが WWW 上で更新されるニュースや天気予報などの知りたい最新情報を，知りたいときに得ることができる．

謝辞 本研究の一部は，文部省科学研究費補助金基盤研究（B）展開（課題番号 10558048）および奨励研究（A)(課題番号 10780169）の支援を受けた．

参考文献

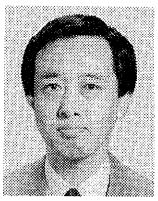
1) Dohi, H. and Ishizuka, M.: Realtime Synthesis of a Realistic Anthropomorphous Agent toward Advanced Human-Computer Interaction, *Human-Computer Interaction: Software and Hardware Interfaces*, Salvendy, G. and Smith, M. (Eds.), pp.152–157, Elsevier (1993).
2) Dohi, H. and Ishizuka, M.: A Visual Software Agent: An Internet-Based Interface Agent with Rocking Realistic Face and Speech Dialog Function, *AAAI technical report 'Internet-Based Information Systems'*, No.WS-96-06, pp.35–40 (1996).
3) Koda, T. and Maes, P.: Agents with Faces: The Effects of Personification, *Proc. 5th Int'l Workshop on Robot and Human Commun. (RO-MAN '96)*, pp.189–194 (1996).
4) Takeuchi, A. and Naito, T.: Situated Facial Displays: Towards Social Interaction, *Proc. CHI '95 Human Factors in Computing Systems*, pp.450–454 (1995).
5) Microsoft Corp.: Microsoft Agent Home Page, http://www.microsoft.com/workshop/imedia/agent/default.asp
6) Walker, J., Sproull, L. and Subramani, R.: Using a Human Face in an Interface, *Proc. CHI '94 Human Factors in Computing Systems*, pp.85–91 (1994).
7) Waters, K. and Levergood, T.: DECface: An Automatic Lip-Synchronization Algorithm for Synthetic Faces, Technical Report CRL 93/4, Cambridge Research Center, Digital Equipment Corporation (1993).
8) 土肥 浩，石塚 満：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，信学論，Vol.J79-D-II, No.4, pp.585–591 (1996).
9) 新田克巳，長谷川修，秋葉友良，神嶌敏弘，栗田多喜夫，速水 悟，伊藤克亘，石塚 満，土肥 浩，奥村 学：論争支援マルチモーダル実験システム MrBengo，信学論，Vol.J80-D-II, No.8, pp.2081–2087 (1997).
10) 近藤和弘，ヘンプヒル，C.：音声認識を用いた WWW ブラウザとその評価，信学論，Vol.J81-D-II, No.2, pp.257–267 (1998).

(平成 10 年 6 月 5 日受付)
(平成 10 年 12 月 7 日採録)



土肥 浩（正会員）

1985 年慶應義塾大学理工学部電気工学科卒業．1987 年同大学院修士課程修了．同年東京大学生産技術研究所勤務．1993 年より同大学工学部電子情報工学科助手．研究分野は知的擬人化エージェントインタフェース，画像メディア技術，ネットワーク化知的情報環境．ACM 会員．



石塚 満（正会員）

1971 年東京大学工学部電子工学科卒業．1976 年同大学院博士課程修了．工学博士．同年 NTT 入社，横須賀研究所．1978 年東京大学生産技術研究所助教授，同教授を経て，1992 年より工学部電子情報工学科教授．研究分野は人工知能，知識処理，画像メディア技術，擬人化エージェント，ネットワーク化知的情報環境．IEEE, AAAI, 情報処理学会，電子情報通信学会，映像情報メディア学会，画像電子学会各会員．