サイバースペース上の仮想人物による実時間対話システムの構築

四倉達夫† 藤井英史† 森島繁生†

コンピュータの処理能力の急速な発展により，複数のユーザがネットワークを介してサイバースペースを共有し，対話や協調作業を行うインタラクション環境が整ってきている．この仮想空間への没入感覚と臨場感を向上させるためには，実空間と同等の自然さで，人間同士のコミュニケーションを実現する必要がある．そこで本稿では，自分自身の姿を投影した顔を持つアバタ（Avatar）を仮想空間上に生成し，マイクから入力された自然音声に同期させて会話時の口形状の推定をリアルタイムに実施し，同時にキー入力された感情情報によってアバタの表情合成を行うシステムを提案する．このシステムによりサイバースペース上で多人数が参加可能なフェイストゥフェイスの対話環境が実現可能となった．

Realtime Face-to-face Communication System in Cyberspace Using Voice Driven Avatar with Texture Mapped Face

TATSUO YOTSUKURA,† EISHI FUJII† and SHIGEO MORISHIMA†

Recent advances in computer performance can generate an interaction environment in which multiple user can share cyberspace and communicate each other to make a cooperative work. An avatar in cyberspace can bring us a virtual face-to-face communication environment. In this paper, we realize an avatar which has a real face in cyberspace and construct a multi-user communication system by voice transmission through network. Voice from microphone is analyzed and transmitted, then mouth shape of avatar are synchronously estimated and synthesized on real time. And also facial expression is controlled on real-time by key-input.

1. はじめに

コンピュータの処理能力の急速な発展と，ネットワークの社会への浸透によって，人間同士のコミュニケーション形態も多様化してきている．コンピュータ上に構築された仮想的な世界，いわゆるサイバースペース[1]上においてテキストベースで対話を行うチャットは新しいコミュニケーション形態をもたらした．このチャットにおいてアバタは対話の相手の姿を表現する重要なシンボルであるが，最近ではより臨場感のある対話に近づけるように実画像をアバタに貼り付けたり，音声を同時に送るなどの工夫が行われている．しかしネットワークの帯域の限界から十分なクオリティの映像を伝送することは難しく，十分に自然な対話環境は実現できていない．

一方，通信の分野では従来のテレビ電話やテレビ会議システムのように通常の通信回線を通してカメラから取得した画像を信号として忠実に伝送するという発想から，ネットワークを通してコンテンツ情報を伝えるという目的にそのフォーカスが移りつつある．モデルベース符号化は，顔の表情を少数のシンボル情報として抽出・記述し，画像は受信側でコンピュータグラフィックスとしてローカルに合成するという方式である．現在では一般的な3次元オブジェクトや仮想空間情報まで拡張されて議論され，MPEG4の標準化活動につながっている．

このような背景から，本稿では仮想空間上に自分自身の姿を投影したアバタを置き，アバタの目を通して見える空間や相手の姿を画像として表現し，仮想会議システムや仮想空間での協調作業環境を実現可能な，いわゆるフェイストゥフェイスでの多人数でのコミュニケーション環境の実現を目的とする．

このような環境における臨場感や没入感を決定する重要なファクタは，画面を通して見る相手とのアイコンタクトや，顔の表情および感情の表現，さらに音声

† 成蹊大学工学部
Faculty of Engineering, SEIKEI University

の伝送と，それに同期した唇の動きであり，このようなファクタによって仮想空間においてもある程度のノンバーバルな対話が実現可能であると考えられる．またアバタの表現としては，個人情報をできるだけ反映できると同時に，コミュニケーションを円滑に行うための工夫，たとえば感情を顕著に表現することができたり，親しみのもてるCartoon的な形状や挙動をもたせることを考慮して，できる限りコミュニケーションギャップを取り除くことが重要である．

そこで本稿では，人間の顔形状になるべく忠実な3次元モデルをアバタの顔として導入した．この3次元モデルは任意の人物の顔画像に整合させて個人モデルを作成することができる．アバタの表情変形はユーザ自身がキーボード入力することによって意図的にコントロールできるように考慮した．これによって表情豊かでない日本人も感情を豊かに表現できるように考慮した．またマイクから入力された音声は，オンラインで伝送することとし，これに同期した口の動きのアニメーションは，この音声の分析によって実現している[2)]．

このシステムによって複数のネットワーククライアント同士の対話が実現され，サイバースペース上で相手の表情変化を見ながら会話することができる．サイバースペースの管理およびユーザ情報管理を行うサーバと，サイバースペースおよび自分から見える複数のアバタをローカルに合成するクライアントから構成されており，参加人数の追加は容易な拡張性が考慮されている．

なお本稿ではサイバースペースを「レンダリングされた3次元データや音声等のコンテンツによって表現される空間や環境」と定義している．

2. アバタのモデル化

表情の表出と音声に同期した唇の動きを実現可能な顔画像を持つアバタを生成するため，カメラから取得した対象人物の正面顔画像に，三角形ポリゴンで構成させる顔の標準ワイヤフレームモデルをマニュアル整合し個人モデルを作成する（**図1**）．このモデルは約850ポリゴンの三角形パッチにより構成されていて，格子点数は約480点から形成される．

ポリゴン数は形状の変化の際の演算量およびレンダリングの処理時間に直接関係する．ここでは実時間でのコミュニケーション実現のため，動きの変化の激しい部分にのみ細かいポリゴンを割り当て，全体的な演算量の削減を行っている．このモデルにテクスチャマッピングを施すことによって顔合成画像を作成する．また歯および口内部のモデルを追加した．

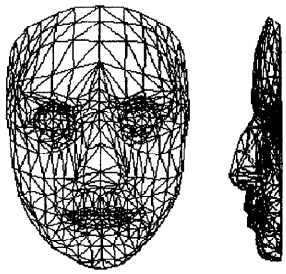


図1 三次元顔モデル
Fig. 1 3D face model.

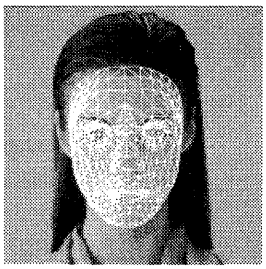
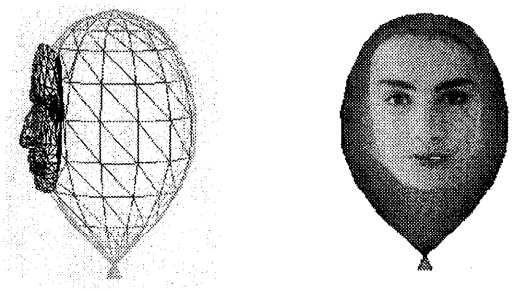


図2 対象人物に顔モデルを整合した状態
Fig. 2 Personal model fitted to 2D image.



ワイヤーフレームモデル　　テクスチャモデル

図3 アバタ本体
Fig. 3 Avatar model.

顔モデルを対象人物に整合した様子を**図2**に示す．この整合処理はマウス操作によって容易に実行することができるように，GUIが工夫されている．ユーザ登録時にこの整合処理を1度だけ行う必要があり，以後は個人モデルの選択のみ行う．アバタ本体には，風船形状の3次元モデルを用いた．このモデルによってアバタの各感情を表現する際に，風船自身の形状および色の変化によって対話相手に分かりやすく伝えることが可能となる．風船形状の合成画像はシェーディングによってレンダリングを行っている．顔画像の合成モデルと風船形状モデルを合体させ，アバタを**図3**のよ

うに表現した．

3. 表情および口形のパラメータ化

表情変化や会話時の口形状変化などを表現する顔画像を再現するために，3次元顔モデルの幾何学的変形のための基準となる特徴点の設定と，その移動量の記述，そして特徴点の周囲の格子点の移動規則などを定める必要がある．ここではモデル変形の基礎となる表情と口形の変形パラメータついて述べる．

3.1 表情パラメータ

表情パラメータとして心理学の分野で提案されているFACS（Facial Action Coding System）[3]と呼ばれる表情の記述法がある[4]．FACSは，顔面筋の位置および動きの方向を解剖学的に考慮して顔の表情をAU（Action Unit）と呼ばれる44個の基本動作に分類している．あらゆる表情はAUの組合せで表現できるとされ，FACSは表情記述単位として顔画像の分析，合成分野で広く用いられている．各AUはいくつかの顔面上の特徴点の3次元移動ベクトルとして定義されている．表情変化は3次元モデルの特徴点をAUの強さによって移動させ，特徴以外の格子点は，特徴点の移動に基づく補間によって制御される．表情の種類としてこのAUの組合せによって表現された，怒り，悲しみ，喜び，嫌悪，驚き，恐れの6基本感情を標準として用意した．もちろん，このAUの編集によってユーザ自身で表情をカスタマイズするためのAUエディタも用意されている．

3.1.1 表情の3次元計測

3次元モデルを変形するためのルールを決定するために，AU表出に熟練した人物に各AUの表出を行ってもらい，モーションキャプチャシステムによって顔の各部の動きの3次元計測を行った．顔の各部に48点のマーカを張り付ける（図4）．マーカとモデルの格子点は1対1に対応しているわけでなく，測定限界による制約からマーカ数を増やすことは不可能であるため，マーカの移動量からその周辺に位置するモデル上の格子点の移動量を決定する必要がある．

そこで，48点のマーカから任意に選び出された隣接した3点で三角形を作り，その範囲内のワイヤフレームモデル上の格子点を検索し，検索されたモデルの格子点の移動量をマーカの移動量によって内挿補間する．ただし，本来の正確な奥行値（Z値）を持たないモデルと実測データの間で3次元的に対応をとることは困難であるため，観測されたマーカ座標点とワイヤフレームの格子点を頭部を取り囲む円筒平面上に写像してこの処理を行う（図5）．図6のようにマーカの3点で作られた3角形内で検索されるワイヤフレームモデル上の点をPとする．仮に $\overrightarrow{PA}$ と $\overrightarrow{PA'}$ が逆方向であるなら範囲内，同方向ならば範囲外である．同様にしてBとCで行い，検索される点を特定することができる．そして3頂点に対する重み係数を算出し加重平均し，検索された点Pの移動量を決定する．ここで定義された格子点の移動量は標準ワイヤフレームに対して定義され，整合処理を経て得られた個人モデルの移動量は，顔のサイズに応じてノーマライズを行っている．

図4 マーカの配置
Fig. 4 Location of markers for expression.

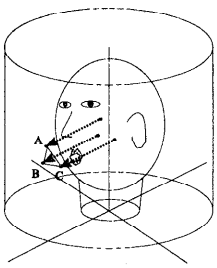


図5 円筒座標への射影
Fig. 5 Projection to cylindrical coordinate.

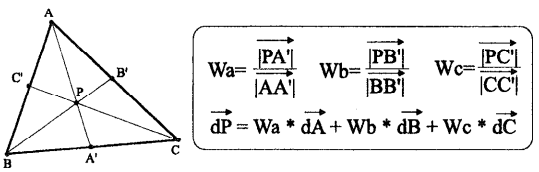


図6 加重平均
Fig. 6 Weighted average.

3.1.2 アバタ表情の合成

モーションキャプチャによって得られた各AU表出時のモデルの格子点の3次元移動量を基に，AUのコンビネーションによる移動量を計算して得られた基本6感情の合成画像を図7(a)～(f)に示す．これはあくまで標準として用意するもので，ユーザによるカスタ

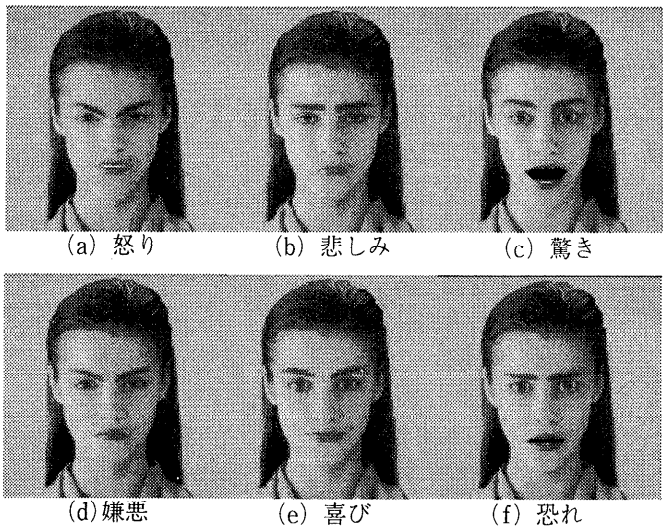


図7 顔モデルによる各表情合成画像
Fig. 7 Synthesized image of basic expressions by face model.

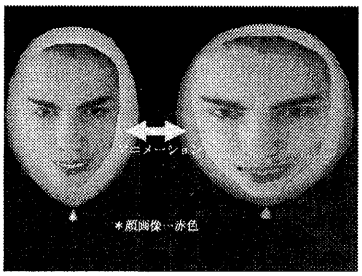


図8 表情"怒り"のアバタ合成画像
Fig. 8 Avatar's anger face.

図9 マーカの配置
Fig. 9 Marker location for mouth movement.

マイズは容易に実行できる．風船形状自体も，各表情に対応した動きを実現して，感情状態を顕著に表現できるように工夫した．たとえば怒りの場合には風船はしだいに膨張し，それにともなってアバタ全体が赤色に変化していく．また驚きの場合，風船形状は激しく上下に飛び跳ねるような変化を行う．幸福では前後左右に揺れて，全身で喜びを表現するように定義した．各々の形状変化によって相手へ伝えたい感情を明確に表現することが可能であると考える．図8に怒りのアバタ全体画像を示す．この感情状態に対応した風船形状の変化についても，ユーザによるカスタマイズが可能である．

3.2 口形パラメータ

発話時の口の形状を表現するために，先に述べたAUとは異なる，口領域の変形に限定したパラメータを用いる．日本語の発音の口形には，異なる発音でも同じような口形となる同口形異音が多く存在する．よってすべての音韻に対応する口形を独立に用意する必要はない．また音声分析性能の限界から，細かい子音に関しての識別は困難である．特に日本語では，大半が母音区間と考えられるので母音区間の口形再現が自然さに大きく寄与すると考えた．そこで5つの母音（/a/, /i/, /u/, /e/, /o/）と閉口の口形を基準とし，すべての口形はこれらの補間によって実現できると仮定している．

3.2.1 口形状の3次元計測

母音発話時の口形を表現するため，図9のように40点のマーカを口周辺に配置した．これを2台のカメラによって3次元計測し，実際に母音を発話している際のマーカの移動量を求め，表情パラメータ決定の際の格子点の移動量算出方法と同様に求め，母音発話時の標準顔モデルの格子点移動量を算出する．

3.2.2 口形パラメータ

口領域の動きを少数のパラメータで表現するために，

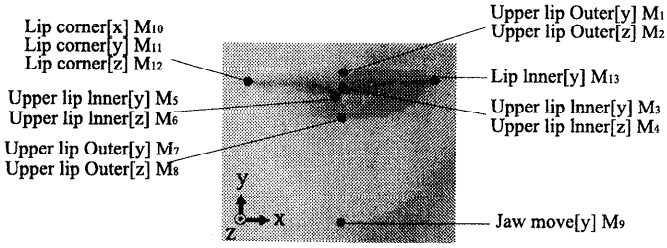


図10 各パラメータの位置
Fig. 10 Location of each parameter.

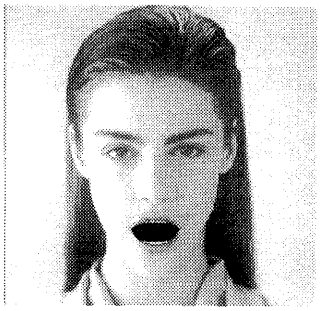


図11 口形/a/の合成画像
Fig. 11 Synthesized image of mouth shape /a/.

口領域の制御点として図10のような13個を定めた．3次元計測結果に基づいて，この制御点自体の移動量の算出，さらに制御点以外の格子点の移動量算出ルールを定めた．この13個の座標値によって，唇の形状を一意に決定することができる．図11にこの口形パラメータによって表現された口形/a/の合成画像を示す．

4. 音声情報から口形の抽出

アバタの口形状をリアルタイムに決定するため，ユーザから入力された音声をフレームごとに分析することによって，毎フレーム口形パラメータを推定する．特徴パラメータとして計算時間が比較的少なく，また発話者の声道特性と放射特性の特徴を表現していると考えられるLPCケプストラム係数とした．入力音声は16 [kHz]，16 [bit]とし，分析フレーム長および周期は32 [ms]で切り出す．LPCケプストラムから口形パラメータへの変換は図12のような3層フィードフォワード型ニューラルネットワークを用いている．入力層はLPCケプストラム次数と同じ20ユニット，出力層は13個の口形パラメータに相当する．さらに中間層は経験的に20ユニットとした．学習パターンは5母音のLPCケプストラムとそれぞれの発話時の口形パラメータ，および無発音時の周囲の環境雑音から求めたLPCケプストラム係数と閉口口形とした．収束までに100万回の学習を行った．このニューラルネットの重み係数は基本的に話者依存性が強く，基本的には話者ごとに事前に学習を行う必要がある．しかし，学習時間の関係から，後に述べる話者適応処理によってこの学習を省略することもできる．

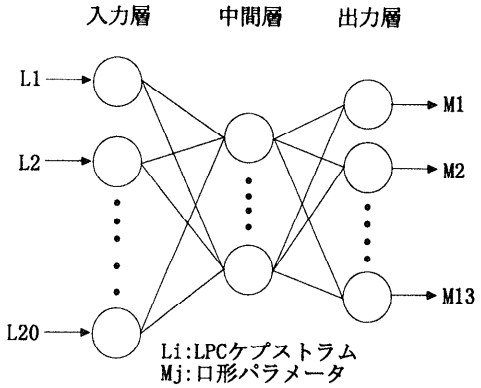


図12 口形パラメータへの変換に用いたニューラルネットワーク
Fig. 12 Neural network for conversion to mouth shape parameters.

5. 実時間対話システム[5),6)]

サイバースペース上でのリアリティのある多人数コミュニケーションを実現させるためには，ビデオレートと同程度の表示速度でリアルタイムに画面合成を行うことが要求される．アバタや仮想空間の生成には3次元モデルを構成するポリゴンを用い，テクスチャマッピングあるいはシェーディングを施して画面表出を行う．この際，そのポリゴンに含まれるピクセル数に比例した処理時間が必要となるため，被験者に不自然を与えないシステムを構築するためには必要最小限な3次元モデルのピクセル数を用い，かつリアルな描画が求められる．また描画が満足されたとしても，一連の処理過程において1つでも低速な処理が含まれていればシステム全体の処理速度を低下させる要因となる．

そこで本システムでは，主にサイバースペースの管

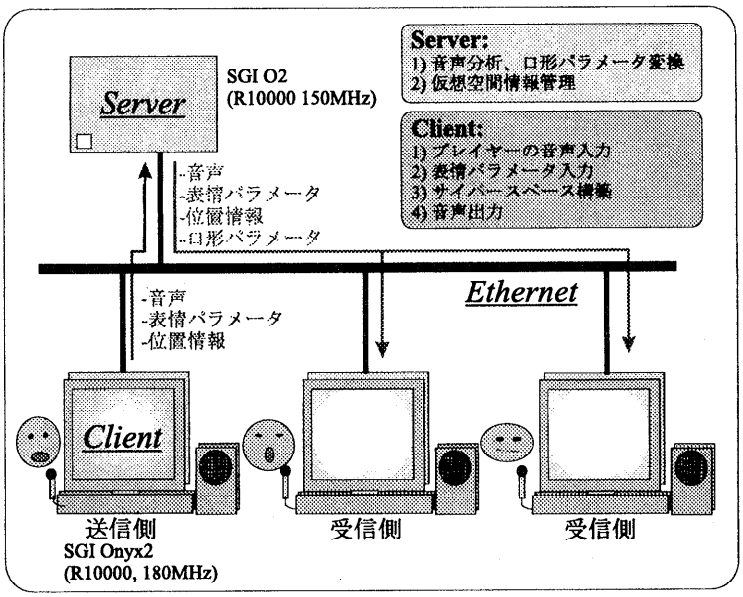


図 13 システム構成図（3 者間）
Fig. 13 System feature for 3 people communication.

理を行うサーバ部，そしてサイバースペースの表出を行うクライアント部の 2 ブロックに分け，各ブロックにおいて内部処理の効率化，高速化を図った．またブロックを 2 分化することにより参加人数が増えた場合，クライアントを参加人数分用意して，サーバ部は大幅なアルゴリズムを変更することなくシステム構築できるように配慮した．現時点では，音声の圧縮伝送を行っていないため，ネットワークの帯域の限界から，違和感なくコミュニケーションできるシステム最大参加人数は 3 人となっている．

5.1 システム構成

コミュニケーションシステムの構成は図 **13** のようなクライアント–サーバモデルとなっている．サーバ部では主に各ユーザの音声分析，仮想空間の管理を行い，各々のクライアントに分析結果を送信する．またクライアント部ではユーザの音声・表情入力，相手ユーザの音声出力，空間構築，アバタ配置，表情合成等を行う．各ワークステーションはイントラネット（10BASE-T）上に接続させており，各プロセスは，このネットワークを通じてプロセス間での通信を行いながら互いのデータの受渡しを行う．

5.1.1 クライアント部

1）音声入力

各ユーザの発話音声をマイク入力し，16 [kHz], 16 [bit] でサンプリングしバッファ内に順次貯えていく．音声の分析フレーム長である 32 [msec] 以上のサンプルがバッファに貯えられると，これが切り出されてネットワーク制御プロセスによって音声サンプルをパケット化し，サーバ部に送出される．

2）感情表現

ユーザが表現したい感情をキーボード入力によって指定する．相手の画面内に合成されたアバタがこの感情情報に反応し，表情変化が行えるようにする．基本表情はあらかじめファンクションキーにアロケートしておき，表情表出のタイミングに応じてキー入力を行う．決定された表情パラメータ（AU）はパケット化されサーバ部に送られる．本システムでは基本 6 感情を標準でインプリメントしているが，任意に感情を増やしたり，再定義したりするなどのカスタマイズが可能である．

3）位置情報入力

ユーザの化身であるアバタはネットワーク上に構築された 3 次元のサイバースペース内を自由自在に移動が可能である．移動方法としては，マウス，キーボードを用いる．マウスを動かすと同時にサイバースペース上のアバタも前進，後退，右旋回，左旋回も行う．キーボードでも同様の動作が可能である．ユーザ位置情報入力プロセスではマウス，キーボードで決定されたサイバースペース上での現在位置を獲得し，その位置情報をパケット化しサーバ部に送出する．なお，アバタから見た視線方向は，現時点では眼球の回転は考慮していないので，顔の正面方向と一致している．

4）アバタ・空間生成

サーバ部において音声分析を行い算出された口形パラメータ，そして各表情パラメータ，さらに音声サンプルを統合したパケットをクライアントで受け取る．クライアントでローカルに走るアバタ画像合成プロセスで各アバタの顔モデルの格子点を移動させ表情および口形変形する．さらに，変形されたモデルにテクスチャマッピングを施し，画像データとしてビデオメモリに格納する．アバタ本体もまた表情パラメータによって形状変化，動きの付加を行いシェーディングを施しメモリに格納する．次に仮想空間プロセスにて空間構築のため空，地面を生成し，サーバから送られてきた位置情報から空間内に各アバタを配置し，ビデオメモリに格納する．これら貯えたビデオメモリは2つの画像メモリからなるダブルバッファ構成となっており，描画のタイミングによる画面のちらつきを解消し，画面への出力高速化が可能となる．

5）音声出力

サーバ部から受け取った各ユーザの位置情報により，相手のユーザと自分との距離を把握できる．その情報から相手の自然音声の音量に強弱をつけ，音声に遠近感を持たせ，相手ユーザの音声をスピーカにミキシング出力する．パケットを受け取った後にスピーカから音声がプレイバックされる時間に比べて，グラフィックス生成時間は長時間を要する．したがって，音声出力する際に経験的に一定の遅延を付加して，音声と口形状との同期を図っている．この遅延は，グラフィックスの生成スピードのみによって決定されるので，クライアントの描画性能によって一意に決定できる．なお，現時点ではクライアントでの音声の圧縮・伸長処理は行っていない．

5.1.2 サーバ部

1）メディア変換

各クライアントから送られてきた 16[Bit]，16[KHz]，512 サンプルを1パケットとした自然音声のパケットを受け取り，4章で述べた音声分析方法によりLPCケプストラムを算出する．次にニューラルネットを用いて口形パラメータに変換する．音声分析プロセス・パラメータ変換を行うプロセスは各ユーザごとマルチスレッドによって並列に実行が進められていく．

2）仮想空間情報管理

各クライアントから送られてきたアバタ位置情報の管理を行う．各位置情報を基に仮想空間内でのクライアント同士の距離を算出し，送り先以外の自然音声，各ユーザの口形，表情パラメータとともにネットワーク制御プロセスによってパケット化を行い，それぞれのクライアントに送出する．

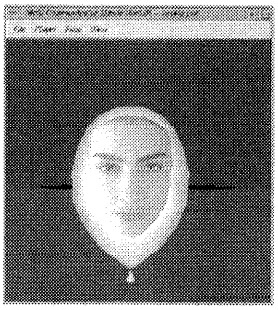


図14 Walk Mode
Fig. 14 Walk Mode.

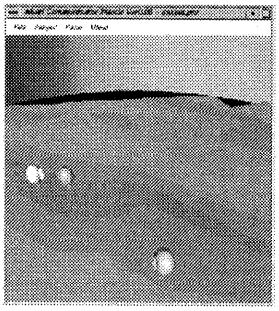


図15 Fly Mode
Fig. 15 Fly Mode.

5.2 ディスプレイ方法

画面表示として2つのモードを用意して，自らのアバタの現在位置の把握，自由な視界変換を可能にした．各視界はウインドウ内のメニューバーによって選択できる．まずWalk Modeは仮想空間内をアバタの目から見た映像を通して移動する際に用いるモードで，通常のフェイストゥフェイスの対話はこのモードで行われる（**図14**）．いわば主観モードである．

Fly Modeでは，自分の位置とは無関係に任意の方向から，仮想空間内を見渡すことができるモードである（**図15**）．空を飛ぶ鳥の視線からの景色を画面に表出できる．このモードを用いることで自分の現在位置，相手との距離を第3者的に確認することが可能となる，いわば客観モードである．

6. ユーザ適応

新たなユーザがコミュニケーションシステムに参加する際，必要なものは正面画像1枚，5母音の音声サンプルのみが基本である．

6.1 顔整合ツール[7)]

ユーザごとの顔モデルの整合を行うため**図16**のよ

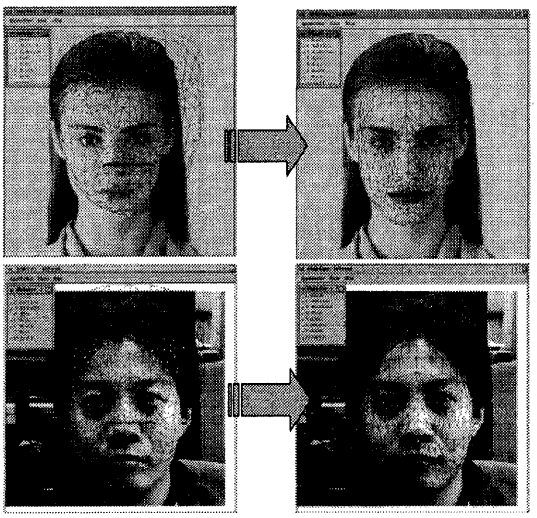


図16 顔整合ツールウインドウ
Fig. 16 Window for face fitting tool.

うなGUIツールを開発した．まずユーザの顔画像を読み込む．顔モデルのワイヤフレームモデルの格子点を動かし画像と顔モデルの整合を行う．点の移動は，はじめはマクロに制御して，しだいに細かく位置合わせできるように配慮されている．また実際に表情変形してみて，不自然な部分はインタラクティブに位置修正できるように配慮されている．特に目と唇の部分は表情変形に重要であるため綿密な整合が必要である．左図は整合前の編集画面であり右図は整合された後の画面を示している．このツールを用いて顔モデルを完成させる所用時間は約5分程度である．

6.2 話者適応

被験者が更新されるたびに，新しい学習データからニューラルネットを収束させるのは非能率的である．そこであらかじめ収録した75人分の学習データで重み係数のデータベースを構築した．この中から被験者に最適な重み係数を自動的に選択する．新しい被験者には，実験開始直前に5母音を発声してもらい，データベースの中から1つずつ選択された重み係数によって順次口形推定を行い，基準の5母音の口形に最も近いものを生成できる重み係数をその人物の最適な係数と判断して，話者適応を実施した．

6.3 音声からの口形推定評価

1995年8月にロスアンゼルスで開かれたACMのSIGGRAPH'95において，インタラクティブデモ展示を行った[8]．このデモでは，会場を訪れた人物の顔正面画像と5母音の音声をその場で取り込み，モデル整合と話者適応処理の後に，リアルタイムでマイクから入力された音声を分析して，口形を合成する処理を行い，合成された顔画像を通じて2者間で対話を行うというものであった．このデモにおいて，来場者160人の整合処理と話者適応を実施し，すべての人物において自然な口形と表情の合成が可能であることが明らかとなった．なお，この際の表情合成速度は毎秒10フレームであり，すべて外国人を対象として，対話は英語で行われた．整合処理は経験のある人物によって実施したが，平均1分程度の所要時間であった．

7. システム評価

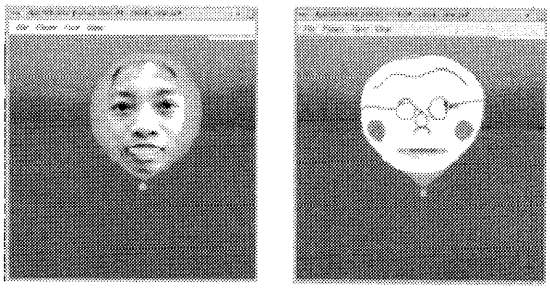
音声伝送によるサイバースペースでのコミュニケーションを実際に被験者に体験してもらい，本システムの操作性，実用性等を調査するために主観評価実験を行った．特に対話の自然さに重点をおいて評価するため，ユーザ数は2者に限定して評価を実施した．

7.1 実験方法

被験者2者間によるコミュニケーションを実施してもらい，被験者の顔画像をモデルにマッピングして合成したものと，Cartoonキャラクタの顔画像を用いてアバタを生成したものの2種類を使用した（**図17**）．実験は研究室内のイントラネット環境としたが，相手被験者の声が直接聞こえないように密閉型のヘッドフォンを用いた．システム構成はSilicon Graphic社製のワークステーションを用い，サーバ部はIRIS O2（MIPS RISC Processer R10000，150 MHz）を使用し，2者のクライアント部にはIRIS Indio2 Impact（MIPS RISC Processer R4400，250 MHz）およびIRIS ONYX2（MIPS RISC Processer R10000，180 MHz）を使用した．描画フレームレートはONYX2では約10フレーム，Impactでは約7.5フレームとなった．まず被験者に本システムを利用して20ないし30分ほど対話を試みてもらい，その後，以下の質問に回答してもらうという形式をとった．

質問1）システムの操作しやすさ．
質問2）相手プレイヤの化身であるアバタの口形状の動きと相手音声との同期はうまくとれているか．
質問3）アバタの口形状が正しく表出しているか．
質問4）被験者の顔画像を用いたものとCartoonを使用したものとどちらが親しみやすいか．
質問5）アバタの風船形状は違和感を感じず，受け入れられるか．

計5項目による主観評価実験を行った．各項目における評定尺度は5段階相対尺度とし，各尺度に［10，5，0，−5，−10］の各得点を与えた．なお，被験者は21歳~23歳の男性9人，女性1人，計10人の学生で，



被験者　　Cartoon キャラクタ

図17 顔画像合成後のアバタ（表情：驚き）
Fig. 17 Avatar's synthesized face with surprise.

10人中2人同士のチームを5チーム行い，1チームごとに被験者の顔画像を用いたアバタ，そしてCartoonを使用したアバタをそれぞれ1回ずつ，計2回行った．

7.2 実験結果

5項目における実験結果は以下のとおりである．

質問1）非常に操作しやすい方を10，非常に操作しにくい方を−10とした結果，平均は2となりほぼ操作しやすいと感じた被験者としにくいと感じた被験者の意見が分かれた．主としてあげられる原因としては，双方でウォークスルーさせていると，なかなかアイコンタクトがとりにくいこと．また自ら表出した表情をモニタリングできないために，感情表出しているようすが直接的に伝わってこない点があげられる．この点については，改良の余地は十分にある．

質問2）非常に同期していた方を10，まったく同期していなかった方を−10とした結果，平均は3となった．音声と口形との同期を実現する際，経験的手法によって遅延を決定したが，フレームレートが毎秒10フレーム以下であるために，十分な同期を感じとることができなかったと思われる．今後は処理の最適化によって最低毎秒20フレーム程度の合成速度が必要であると思われる．

質問3）口形状が正しく表出されている方を10，まったく表出されていない方を−10とした結果，平均は1.5となり音声分析による口形抽出による表現は意見が分かれた．特にCartoonキャラクタを被験者のアバタとした際，口形状が正しく表出されていないという意見が多かった．これはCartoonキャラクタの場合，唇は線画で表現されるため，口形状が十分に認識しにくいという原因によっていると思われる．

質問4）被験者の顔画像を使用した方が非常に親しみやすい方を10，Cartoonキャラクタ画像を使用した方が非常に親しみやすい方を−10とした．結果，平均は5.5となり顔画像を使用したアバタを利用した方が親しみやすいという意見が大多数を占めた．デフォルメされたCartoonキャラクタのような顔画像を使用するとモデルと顔画像との整合が難しく，感情がうまく表出されないと考えられる．

質問5）非常に違和感を感じずに受け入れられる方を10，まったく受け入られない方を−10とした．その結果，平均は1.0となり，受け入れられるという意見の方が若干強かったが，全体から見て意見が分かれた．感情の表現力があり，親しみやすいなどの受け入れられる側のコメントを受けた反面，顔画像がテクスチャマッピングによるものなので後頭部や体の部分も実際の人物に近い方がより受け入れやすいという意見もあった．

8. まとめ

自然音声から口形状を変形させるアバタを用いたサイバースペース上での実時間の多人数コミュニケーションシステムを構築した．また，滑らかなアバタのアニメーションを実現するために音声入力から画像出力までの一連の処理プロセスを分割し，各プロセス内の処理を最適化することによって高速化を図った．評価結果から，おおむね良好な結果を得たが，改良の余地が十分にあることも指摘された．

また，現状では表情パラメータの決定方法はキーボード入力である．現在は音声からの感情の自動推定を行う方向で検討を進めているが，このためには感情を表現するための最適な音声パラメータ発見の必要がある．さらに音声だけではなくカメラを用いた実時間の画像分析をも導入することによって，感情推定精度を向上させる検討も進めている．

さらにカメラからの利用者の視線追跡とアバタの眼球回転を考慮することによって，実際のアイコンタクトを実現したフェイストゥフェイスでのコミュニケーションが実現され，より臨場感のある対話が実現されると考えている．

本システムでは現在，音声の伝送の際に生じる膨大なパケット量のため，4者間以上でのコミュニケーションを行う際には，パケット落ちが生じ，再生音声が一部欠損してノイズが生じることがある．今後は音声の圧縮，システム全体のアルゴリズム最適化を行いパケット量の軽減を図り，より多くの人物が参加可能なコミュニケーションシステムの構築を進めていく．さらにアバタに人体と頭部を付加して，実際の人物に近

い形状表現を実現し，臨場感を向上させる予定である．

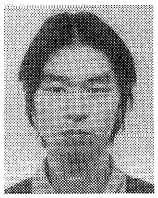
参考文献

1) Gibson, W.: *Neuromancer Ace Sience Fiction Books*, New York (1984).
2) 宮下，坂口，森島：ヒューマンコンピュータインタラクションのための音声から画像へのリアルタイムメディア変換，インタラクション'97，pp.53–54 (1997).
3) Ekman, P. and Friense, W.V.: *Facial Action Coding System*, Consulting Psychologist Press (1977).
4) 原島，武部：顔の3次元モデルに基づく表情の記述と合成，電子情報通信学会論文誌A，Vol.J73-1, No.7, pp.1270–1280 (1990).
5) 四倉，藤井，小林，森島：音声から画像へのメディア変換を用いたサイバースペース上での多人数コミュニケーションシステム，インタラクション'98論文集，pp.133–134 (1998).
6) 四倉，藤井，小林，森島：音声による実時間口形・表情制御可能なサイバースペース上での仮想人物の実現，電子情報通信学会技術研究告，MVE97-103, Vol.97, No.566, pp.75–82 (1998).
7) 森島，八木，金子，原島，谷内田，原：顔の認識・合成のための標準ソフトウェアの開発，電子情報通信学会技術研究告，PRMU97-282, Vol.97, No.596, pp.129–136 (1998).
8) Morishima, S.: Better Face Communication, *Visual Proc. SIGGRAPH'95*, p.117 (1995).

(平成10年 6 月 8 日受付)
(平成10年12月 7 日採録)

四倉 達夫（学生会員）

平成10年成蹊大学工学部電気電子工学科卒業．同年同大学大学院修士課程在学中．自然音声による実時間口形・表情制御可能な3次元多人数コミュニケーションシステムに従事．



藤井 英史（学生会員）

平成9年成蹊大学工学部電気電子工学科卒業．同年同大学大学院修士課程在学中．3次元モデルを用いた音声から顔動画像へのメディア変換・Action Unitによる表情の合成等の研究に従事．情報通信学会会員．

森島 繁生（正会員）

昭和57年東京大学工学部電子工学科卒業．昭和59年同大学大学院修士課程修了．昭和62年博士課程修了．工学博士．同年成蹊大学工学部電気電子工学科専任講師．昭和63年同大学助教授，現在に至る．平成6年より1年間トロント大学客員教授．コンピュータグラフィックス，コンピュータビジョン，マルチモーダルインタフェース等の研究に従事．平成4年電子情報通信学会業績賞受賞．IEEE，ACM，日本音響学会，テレビジョン学会等各会員．