音声言語情報処理 7－1
（1995．7．20）

臨場感通信会議におけるマルチモーダルインタラクション

宮里 勉　　岸野 文郎

ATR通信システム研究所、京都府

本稿では，ATR通信システム研究所で試作した，計算機により生成される仮想の三次元環境内で異なる3地点にいる参加者が協調作業を行なう3地点間の臨場感通信会議システムについて紹介し，仮想物体操作におけるマルチモーダルインタフェースについて述べる．臨場感通信システムの構成，特徴，システム実現のための要素技術について述べるとともに，94年9月19日から22日までの期間に行った国際電気通信連合（ITU）全権委員会議の電気通信展示会場の国立京都国際会館とATRを結んだデモにおける協調作業の例を紹介する．

Multi-modal Interaction in Virtual Space Teleconferencing

Tsutomu MIYASATO　　Fumio KISHINO

ATR Communication Systems Research Laboratories, Kyoto, 619-02, Japan

This paper introduces a virtual space teleconferencing which is expected as an example of a human-oriented telecommunication system. A virtual space teleconferencing at ATR is a new paradigm "Communication with realistic sensations". By using this system, participants can engage in the conference with the sensation of sharing the same space, and cooperative work can be performed among remotely located participants. We have built an experimental system which connects three different sites by 1.5 Mbps ISDN in commercial use. The system has two large screens in which real-time reproduction of 3-D human's whole body images are realized. Participants at three different sites can feel as if they are at one site. They can also cooperatively work through virtual common space using multi-modal interface, in this case hand gesture and natural language. Virtual manipulation is improved by using natural language combined with hand gesture. By using natural language, we can not only manipulate 3-D virtual objects but also generate and modify them. Our system was demonstrated at telecommunication exhibition in ITU-PP'94 held in Kyoto in Japan from Sept. 19-22, 1994. Promising results for real-time cooperative work using the experimental system are demonstrated.

1．まえがき

通信網の発達は我々の社会を高度情報社会へと変遷させつつある．これに伴い，人と人の意思の疎通を支援するヒューマンコミュニケーションの重要性が増してきている．電気通信は，距離を隔てた人間同士のコミュニケーションを支援するためのものであり，電話による音声主体の意志の疎通から，テレビ電話やテレビ会議など映像を伝え合う視覚による意志の疎通を図るものへと発展してきた．

しかし多量の情報を伝送できる画像通信システムを用いても，現在までのところ通常の面談会議のように，多くの人が一堂に会しているような感覚，臨場感を再現することは困難である．

これに対して，ATR通信システム研究所では，互いに異なる複数の場所にいる多くの人々が，あたかも一堂に会した面談の感覚で会議を行なうことができる臨場感通信会議を提案し，その実現に必要な要素技術の研究を進めている[1,2]．臨場感通信会議は，バーチャルリアリティ（VR）と電気通信とを複合した会議であり，単に相手側の映像を伝送するという考え方ではなく，同じ雰囲気が感じられる”コミュニケーション空間”をコンピュータで生成するという新しい考え方に立っている．臨場感通信会議では，仮想の空間内で，実空間での面談会議で経験するように，会話をしている二人の視線が一致するのを他の参加者が確認できたり，また，議長が発言のきっかけを掴みたい人の視線を感じたりできる．

本稿では，計算機により生成される仮想の三次元環境内で，異なる地点にいる参加者が協調作業を行なう3地点間の臨場感通信会議システムについて紹介し，仮想物体操作におけるマルチモーダルインタフェースについて述べる．

なお，平成6年9月に行った，国際電気通信連合（ITU）の全権委員会議[*]の会場の国立京都国際会館とATRを結んだデモにおける協調作業の例を交えて紹介する．

2．臨場感通信会議システム

2.1 臨場感通信会議

コンピュータグラフィクス（CG）により仮想的な3次元空間を生成し，会議参加者の人物像をこの仮想空間に合成表示することにより同一の会議室にいるかのような感覚を再現する．この方式の中で扱われる人物像は，テレビカメラより撮像した人物像を処理し，3次元モデルを生成したCG像であり，あらかじめ表面形状とテクスチャーはデータベースに記憶されており，通信時は変化情報のみ送られる．

臨場感通信会議のイメージを図1に示す．仮想の会議室が3次元CGで生成され，人物も3次元的に処理され表示される．その際，仮想会議室の照明条件，受け手側の視点の位置等がコンピュータでリアルタイムに計算される．このようにして実空間での面談会議と同様の環境を提供することができ，更に仮想空間を共有して協調作業環境を提供することもできる．図1のイメージ図では新しい車のデザインを3地点に離れた会議参加者がディスカッションしている様子を示している．また，様々な仮想空間を作り出すことができるため，会議の目的にふさわしい環境にすることによって[3]，従来の通信会議にはない付加価値が生まれるものと考えられる．

さらに，CG人物像を用いる他の利点として，表情を制御できるという点がある．日本人の表情は欧米人に比べて乏しいと言われるが，CG顔により表情をオーバーにしてめり張りを付けることができる．したがって，CGによる表情合成により，ノンバーバルな情報を相手に誇張して伝達することや会議相手の文化や国民性に合わせた表情の翻訳も可能となる．

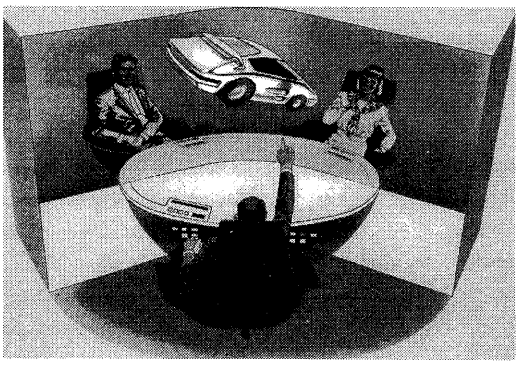

図1　臨場感通信会議のイメージ

[*]平成6年9月に京都で開催．同会議は，電気通信分野における国際連合の唯一の専門機関であり，最高意思決定機関である．

2.2　システムの特徴

臨場感通信会議を実現するためには，1）実空間と同じ感覚で観察可能な高臨場感表示技術，2）人物像を同じ会議室にいるかのように，かつ，面談会議と同じような感覚を再現する人物像処理技術，3）離れたところにいる人々があたかも同じ作業空間で協調作業を行なうような環境を提供する高度協調作業技術，などの要素技術の確立が必要である．本システムは次の特徴を持つ．

1）高臨場感表示

表示には，70インチの大型スクリーン2枚を横に密着して並べ，継ぎ目のない横長の立体画像を表示することにより，視野角120度の包み込まれるような視界を生み出している[4]．なお，表示画面は，縦×横が約2000×500画素であり，立体画像は液晶シャッタメガネで観察する．

また，本システムでは画像そのものを伝送するのではなく，センサーからの位置データだけを伝送し，表示側の計算機がCG画像をリアルタイムに生成する．

2）人物像処理技術

協調作業相手を実際に仮想空間内に実在するように表示できれば，協調作業時において作業相手の動向を画像からより直感的に知ることができる．人物像の表示は3次元モデリング，実時間動き検出，人物像生成の3つの要素から構成される[5]．表示されている人物画像は実際に撮影した画像でなく，3次元の人物モデルにテクスチャーを貼ったCG画像である．このCG画像がどれだけ人間らしく見えるかは，センサーであるカメラが参加者の表情をどれだけ正確に検出できるか，また，得られた検出データから人物モデルをどれだけ自由に変形できるかに依っている．

図2に人物像の処理の流れを示す．3次元モデリングでは，人物の表面形状を3角パッチで近似し，三角パッチの集合体であるワイヤーフレームモデルを作成する．また，同時に人物の表面の色彩情報も取得する．これらの処理には3次元デジタイザなどの機器を用いている．

人物の実時間動き検出では，顔の表情，頭，体，腕，指の動きを検出する．顔の表情の変化は，実際の人物の顔に特徴追跡用のマーカを貼り，これらのマーカを画像として追跡することにより検出する．人物はヘルメットをかぶり，このヘルメットにテレビカメラ

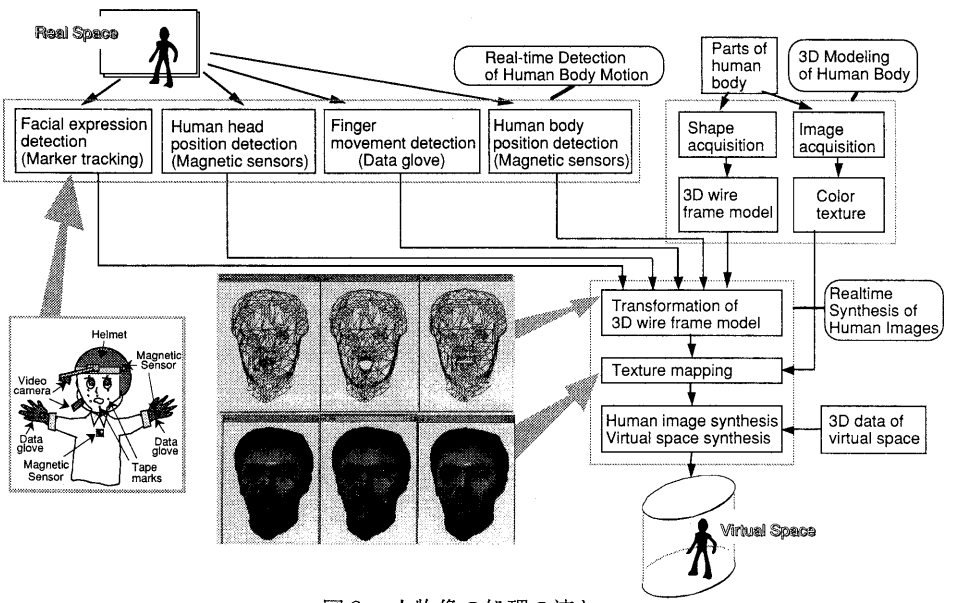

図2　人物像の処理の流れ

を固定して，顔に対して常に一定の位置から画像を得ている．人物の頭部と体，腕の３次元空間における移動と回転は，それぞれの部分に装着した３次元位置センサにより検出する．指の曲げは特殊な手袋を用いて検出する．以上のようにして検出された動き情報は，人物像のワイヤフレームモデルの変形に使用される．あらかじめ獲得されているワイヤフレームを変形し色彩情報をマッピングすることで人物像の３次元像が得られる．

３）高度協調作業

神輿の各パーツを組み合わせて，３地点間で話し合いながら完成させるという協調作業において，手振りによるジェスチャに加えて音声認識と衝突検出を導入した．ジェスチャ認識は各関節の曲げ角と環境の状態（対象物と手のひらの中心の距離，対象物の大きさなど）を基に決定される．また，特定話者認識方式による音声認識を用いて，仮想物体の配置，結合，移動，拡大・縮小，形状の変形，取り消し，を行っている[6]（図３）．音声コマンドには同義語を用意し，指示の自由度を持たせている．さらに，ジェスチャと音声コマンドとを組み合わせることにより，手が届かない距離にある部品に対して，手をかざしながら音声で「屋根を持ってくる」と言えば，かざした手の位置に屋根が飛んでくる．

また，仮想空間内の作業支援として高速衝突検出機能[7]を加え，組み立て作業中に仮想物体同士が衝突した際には衝突面の色が変化すると共に衝突音が発生する．さらに，結合を目的とした物体の近くまで物体を運ぶと，結合音と共に所定の位置と傾きで自動的に結合する．

３．３地点臨場感通信会議の実施

３．１　システム構成

図４に３地点臨場感通信会議システムの構成を示す．デモでは，京都国際会館の展示会場と会場から約40 Km離れたATR内の２つの部屋の計３ヵ所を実際に

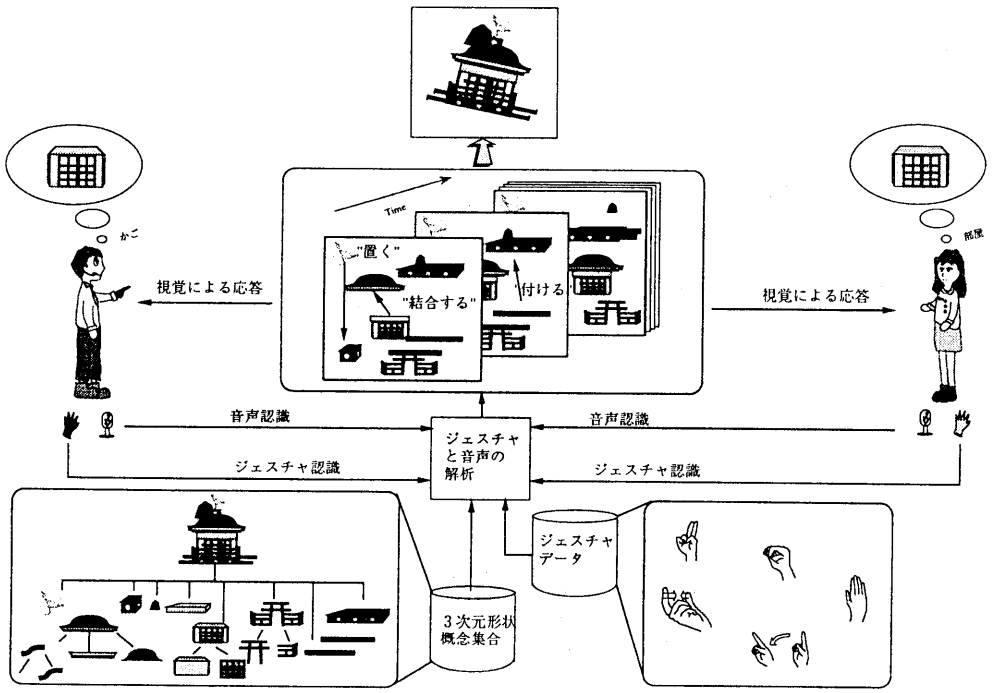

図３　音声と手振りによる対話的な物体の生成と操作

結んだ．展示会場とATR間は1.5 Mbps 1回線で結び，ATR内の2部屋はイーサネットで結んでいる．デモの主会場である京都国際会館イベントホールには，広視野の70インチディスプレイを2面用い，左画面にATR-1会場の参加者を，右画面にATR-2会場の参加者をそれぞれ表示した．主会場の参加者は液晶メガネをつけて画面を見ることにより，表示されているATR-1，-2の両参加者，および画面に表示されている物体を立体画像として見ることができる．また，主会場の参加者はセンサー付きの手袋をはめることにより，表示されている物体を手で掴んで操作できる．一方，ATR-1，-2の参加者はセンサー付きの手袋の他に，視線および表情検出用の小型カメラとヘッドセットを取り付けたヘルメットを被っている．

図5に，臨場感通信会議システムにより仮想の神輿を製作している様子を示す．

3.2 デモ内容

実際には遠隔地にいる3人があたかも一堂に会しているかのように，仮想の会議室内で共同作業を行って神輿を製作する過程を示した．デモでは，離れた場所にいるもの同士が協力して神輿を作ることが可能なことを示し，見学者にも体験させた．

デモの流れは，神輿のパーツ集めから始まり，神輿の組み立て，変形，神輿の完成となる．神輿用パーツの呼び出しは音声認識で行い，各パーツの組み合わせによる神輿の原形の組み立てには，手操作と音声認識を用いた．また神輿の一部の拡大・縮小や変形などの操作は音声認識により行った．音声認識は，特定話者認識であり，作業者の足元のフットペダルの上下により音声認識機能の開始・終了を制御した．

4日間におけるデモを通して，3地点を結んだ仮想空間内での高度な協調作業の有効性を確認したとともに，操作を数名の操作者が行うことで，以下の知見が得られた．音声認識の使用により，手が届かない距離にある部品の操作や，変形のための抽象的な指示が可能になり，協調作業の操作性を向上させた．また，音声認識により指定された物体を明確にするための視覚フィードバックは，認識機能が動作しているか否かの判断が早まり，操作者の不安解消に効果があった．また，現在のシステムには，力および触角フィードバックは備わっていない．したがって，視覚だけでは対象物をつかんだというフィードバックが弱いため操作に戸惑いが生じる．今回用いた仮想物体同士の衝突時の衝突音の発生や結合音の発生は，直接の力および触角フィードバックには及ばないが，次善の手段として有効であり操作性を向上させた．ただし，操作者が物体を掴んだ際の音の発生は操作者を混乱させることになった．これは，使用した音の選択の違和感および音像を制御していないことによるものと思われる．

4. 課題

以下に，試作した臨場通信会議システムの実演から得られた課題を述べる．また，本研究会の趣旨に沿って，特に音声の入力に関連した私見を述べる．

4.1 融通性のあるシステムへ

3地点の参加者による神輿の製作では，予め部品同士の結合面が決まっていた．したがって，音声による曖昧な指示でも部品同士は所定の位置に収まった．しかし，任意の物体同士の結合の場合には，結合面は一意に定まらない．したがって，音声単独ではなく，ポ

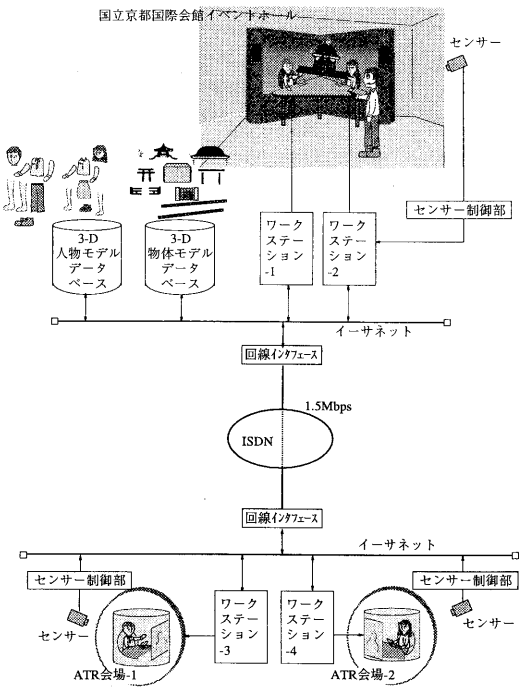

図4 臨場感通信会議システムの構成

インティングジェスチャと組み合わせた指示が必要である．さらに，結合が可能な面同士の自動判定・推定のための機能も必要となろう．

また，特殊手袋の着用は誰にでも歓迎されるものではないので，今後，特殊な手袋に代わる手形状入力手法として，画像処理を用いた非接触による形状認識手法[8]を組み込む必要がある．同様に，眼鏡無しの大画面立体表示装置[9]の組み込みにより，実空間と同様な表示が可能となり臨場感の増大が期待される．

ところで，前述の物体の結合における自動判定・推定機能には反するが，部品のでたらめな組み合わせから奇抜な物が出来上がる場合がある．そこで，神輿をできるだけ実物に忠実に再現できることを目指した今回のデモに対し，型にはまらない従来とはまったく異なった神輿の製作を支援するような，製作者の創造性を活性化する環境[10]の積極的な提供も考えられる．

ここで，CG合成顔の品質について述べたい．CGによる静止した状態の合成顔を見せると，一般的な反応として，ロウ人形のようだ，生気が感じられない，実物と差がある，などの意見が出る．この点に関して，現在改良を進めているところである．

しかし，言い訳に聞こえるかもしれないが，写真であっても，実物とは違和感はある．CG顔（人物）の品質評価において，静的な状態で評価してもあまり重要ではなく，動的な評価において実物と同じにすることが重要である，と思っている．

他人の第一印象を決定する主要因は顔や表情ではあるが，それは会話を交す以前においてである．つまり，最初、悪い第一印象でしかなかった人が、話している内に印象が良い方向にガラリと変わった，ということは日常よく耳にすることである．

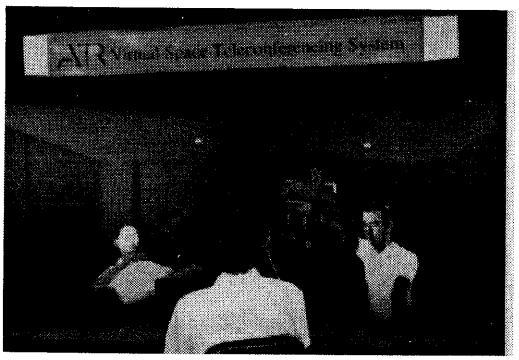

図５　３地点の参加者による神輿の組立て

つまり，人と人とが，単に一瞥するだけではなく，共同・協調作業をする際には会話の存在が必然的である．CG合成顔を静的に評価する方法では，その顔の第一印象しか評価していないことになる．したがって，静的な評価よりも、会話行動を通しての相手の印象に関わる動的な評価が重要なのである．また，会話中の重要な情報として相手の視線の動きの知覚がある．この点は，現在のCG画像でも実物と同じに知覚できる[11]．

実際に，CG人物相手に会話をしていて会話が弾んでくると，相手がCG像であることを忘れることがある．相手の肉声からの情報に補間効果があるのだろうか，頭の中では普段見慣れている相手のイメージ（パーソナリティ）と会話しているようである．

したがって，相手のパーソナリティを再現するのに高品質のCG像は必要であるかもしれないが，必ずしも十分ではなく，逆に，極論すれば，パーソナリティが再現できれば低品質のアニメ像でも構わないのかもしれない．

4.2　音声認識は客寄せパンダか？

前述したように，臨場感通信会議システムでは音声による言語表現の入力ならびに手振りによる表現を用いて３D仮想物体の生成や加工を行う．

いつもながら，音声認識を用いたデモの一般の人への受けは高いものである．しかし，通常よく見られる例として，予め予定された語彙を用いるようにしたデモでは良好に動作するが，実際に使用する段になると認識がうまく行かない，ということがよく起こる．たとえば，「竿」と言っても認識されず，「棒」と言い直して認識される．そこで，「竿」という語彙をデータベースに追加して，ひとまず「竿」は認識できるようになる．しかし，問題はそんなことではないと思う．

認識できる語彙の蓄積数を増やした結果，メモリの消費，高速検索のためのアルゴリズムの採用によるシステムの負担増の割には，やはり音声認識機能が使われない，ということになりかねない．

もちろん，筆者は，音声認識によるユーザインタフェースを否定するつもりはない．「ライトをつけて」と言えば部屋の明りが点灯する，あるいは，仮想

空間中の未知の物体を指さして尋ねると，その物体に関する詳しい情報を答えてくれる，などのような場合には音声認識入力は非常に有効である．さらに，他人との話し合いにより自分の考えがまとまることを思えば，前述したように計算機との間の言葉によるブレインストーミングも発想支援にとって有効である．

しかし，「音声が人間にとって自然であるから」という理由で音声入力を使うことには疑問を持っている．CGにより合成された仮想物体の操作に関する意図を計算機に伝える手段として自然言語が適切だとは考えられない．（ここは研究室内でも意見が分かれるところである．）確かに，現在のシステムでは粘土をいじるように自由に仮想物体を作れないため，音声操作に頼らざる得ない．しかし，触角・力フィードバック装置[11]が装備された時点でも物体操作のために音声が積極的に使われるであろうか？人が実際に物を製作する際に，はたして，システムに音声で指示するであろうか？製作者の性格にも依るであろうが，一般に，創造的な製作活動においては，人は無口になるものである．頭の中の物体の既存のイメージ，あるいは試行錯誤で物体のイメージを作る際に，人は一々音声化して表現しないと思われる．人間国宝と呼ばれる職人の仕事振りが良い例であろう．つまり，口を出すより手の方が早いのである．「かつぎ棒を神輿に取り付ける．前後が同一の長さになるようにして，神輿の縁と平行にする」などと音声で指示するより，直接自分の手で取り付けた方が早い．作図ツールを音声入力で使おうなどとは思わないのと同じである．

以前，音声認識機能付のマウスの評価を行ったことがある[12]．音声認識の利用は，文書編集作業中にキーボードとマウスの間で手を往復させるのが煩わしいということから出たものである．評価結果は，音声認識を使うと編集作業（カーソルの移動）が短縮化できるということであった．しかし，被験者の感想として，音声入力は意外と疲れるとのことであった．また，本人は「右」のつもりだが発声では「左」と誤ってしまうこともあった．（因に，かつて"Road Runner"というパソコンゲームをキーボード操作ではなく音声で操作するようにしたことがあるが，初めはスムーズに操作できる（逆にいうと，操作しているという感覚が乏しい）が切羽詰まってくると，無意味な声が出るだけで，最後は手に頼ることになった．）

ユーザインタフェースとしての音声の利用，特に音声入力は，認識そのものより，ユーザから発せられた音声から如何にユーザの意図を抽出するかということが重要であると思われる．すなわち，システムがこちらの日頃の行動や思考パターンを把握していれば，「風呂」，「飯」，「ビール」などのトリガー語だけで，こちらの意図が達成されることになる．

音声はシステムにある動作を起こさせるトリガーであり，そのトリガー語からシステムが如何に意図を理解できるかということが重要な問題である．つまり，システムが如何に気が利くかということになる．「ミロのビーナスのあの胸の曲線」と言うだけで，意図した曲線が得られれば音声入力は大歓迎である．しかし，これは音声認識というより人工知能の領域だからと，音声研究者には関係ないことであろうか？

システムが気が利くためには，音声以外の情報も重要である．たとえば，人間が何か物をつかもうとする時に，プリシェイピングと呼ばれる，手が対象物に届く前にその大きさや形，機能に応じて手の形を準備する無意識の行動が存在する．したがって，プリシェイピング情報から，操作者が長い棒状の物体を掴もうとしているのかそれとも水平の平板を掴もうとしているのかを推測できる可能性がある[13]．操作者の無意識の動作からその動作の意味を解釈できれば，操作者が意図を音声化する必要がなくなり，操作者にとっての認知的な負担の少ないより自然なインタフェースが得られるものと考えられる．

また，気が利くシステムは，人間の認知的特性も把握しておかねばならない．話は唐突だが，レストランで，料理を注文した際に，ウェイトレスが注文を復唱するが，復唱されることが店によっては何となく煩わしく感じることがある．ファミリーレストランなどで多く，小さな飲み屋では気にならない．お客が料理を注文する際，思い付いた順に注文するわけだが，その時，その順序も記憶に残っているということになる．ウェイトレスが注文を注文伝票に記入することを考えると，

☆チェーン型のファミリーレストラン

効率化を狙っているチェーン型のファミリーレストランでは，注文伝票には予め料理名が記入されており，客から注文を受けたときには，料理の数だけを記入すればよいようになっている．そし

て，注文の復唱の際には，その伝票の上から順に読み上げていく．

そのため，客の注文の順序と復唱される料理の順序は一致しないことになる．したがって，客の側では料理の復唱時に記憶との照合が煩わしい．

一方、

☆一杯飲み屋

注文伝票と言ってもただの紙切れの場合が多く，客から注文を受けると，受けた順にその紙切れに料理名を書き込んでいく．その際の習性として，上から書き込んでいく．また復唱時には，これも習性として上から読み上げる．従って，「注文を受けた順＝復唱順」が成り立ち，しかも「注文順≒客の記憶順」であるので，客の側では，料理の復唱時に記憶との照合が容易になり煩わしを感じない．

そこで，復唱時の最適順序についての実験をした結果は，系列記憶の初頭効果と新近性効果と同じパターンが見られた．すなわち，復唱時には，被験者が口に出した順ではなく，最初に口に出した単語と最後の単語を先に提示したほうが，記憶との照合時間が最も短くなった．

以上のことより，気の利くシステムが何かを復唱する際には、単に認識率100%でかつ美しい合成音声を発するだけでは不十分であり，確認を求める人間の認知的側面も考慮する必要があるということになる．

5．むすび

計算機により生成される仮想の三次元環境内で，異なる地点にいる参加者が協調作業を行なう3地点間の臨場感通信会議システムについて紹介し，仮想物体操作におけるマルチモーダルインタフェースについて述べた．

また，仮想物体操作に関わる音声の利用に対する筆者の思い込みについても吐露せてもらった．音声の専門家のお叱りを受けるかもしれない．見当はずれの点はご容赦願いたい．

従来のテレビ会議のような送り手主体の画像通信で行なえる協調作業は限られている．円滑な協調作業の実施を支援するためには，離れた利用者同士がお互いに対等の立ち場で情報を共有できかつ協調相手の動向も正しく把握できることが望ましい．本稿で述べた仮想空間を用いたコミュニケーション空間の共有はこのような形の協調を可能にする．

今後，共有空間の最適な設計については，心理的，生理的，社会的な面からの研究も重要であり，検討を進める予定である．

参考文献

1) 岸野，山下："臨場感通信のテレコンファレンスへの適用"，信学技報,IE89-35(1989)

2) 岸野："ヒューマンコミュニケーション-臨場感通信"，テレビジョン学会誌，46(6), 698-702, 1992

3) 宮里，岸野："3次元類似シーンの検索"，マルチメディアと映像処理ｼﾝﾎﾟｼﾞｳﾑ'94 (Apr. 1994)

4) 志和，岸野："臨場感通信会議のための広視野立体表示"，信学技報，IE94-112 (1995)

5) 大谷，北村，竹村，岸野："臨場感通信会議における3次元顔画像の実時間表示"，信学技報，HC92-61,pp.23-28(1993)

6) 吉田，ティヘリノ，安部，岸野："マルチモーダル3－D形状編集システムの実現"，94信学秋 A-165

7) 北村，竹村，アフジャ，岸野："仮想空間における協調作業支援を目的とした物体間の干渉チェック法に関する検討"，ヒューマンインタフェース N&R, 8, 247-254, 1993

8) 石淵，竹村，岸野："パイプライン型画像処理装置を用いた実時間手形状認識"，信学技報，HC92-14 (1992)

9) 鉄谷，岸野："視点追従型連続視域立体表示"，3D Image Conference '93, 39-44, 1993

10) 西本，宮里，岸野："連想記憶を用いた情報提供による発散的グループ思考支援の試み"，第2回発想支援ツールシンポジウム講演論文集(1993)

11) 森井，岸野，鉄谷："眼のCGアニメーションと視線の知覚に関する検討"，信学論誌 Vol. J78-A, No. 4, pp. 512-522 (Apr. 1995)

12) 竹村，伴野，岸野："仮想空間操作における力フィードバックに関する考察", 1991年信学春全大 A-248

13) 宮里："音声コマンド付フットマウスの評価について"，昭63信学春 D-301

14) 宮里，岸野："把持動作におけるプリシェイピングからの目標物体の推測"，1995年信学総大 A-259