音声言語情報処理 39－12
(2001. 12. 20)

〔招待論文〕

ロボット聴覚の課題と現状

奥乃 博[†‡] 中臺 一博[‡]

[†] 京都大学大学院 情報学研究科 知能情報学専攻
〒606-8501 京都市左京区吉田本町
[‡] 科学技術振興事業団 ERATO 北野共生システムプロジェクト
〒150-0001 東京都渋谷区神宮前 6-31-15 M31
e-mail: okuno@nue.org, nakadai@symbio.jst.go.jp

あらまし

ロボットが人間社会の中に入り込み, 共生していくためには, 混合音が扱えること, アクティブオーディション, 動きながら聞く機構, 未知環境での音の知覚, 画像処理などの他の処理の統合, 実時間処理が大きな課題であることを指摘した. 混合音の処理では, 音源定位が重要であり, 頭部伝達関数 (HRTF) を使わない手法が必要となる. 本稿では, これら課題に対して, マイクロフォン 2 本が必要であるという考えを述べ, 2 本のマイクロフォンで実現可能な機能について, 解説をした. 体を動かして聞くというアクティブオーディション, あるいは, 画像処理とモータ処理を統合して, 体全体で聞くという情報統合が重要である. そのために, 方向通過型フィルタや聴覚エピポーラ幾何学, 実時間処理方法を開発して, 複数の実験で有効性を確認した.

キーワード ロボット聴覚, 音環境理解, 音響エピポーラ幾何学, アクティブオーディション, 実時間処理

Research Issues and Current Status of Robot Audition

Hiroshi G. Okuno[†‡] Kazuhiro Nakadai[‡]

[†] Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
[‡] Kitano Symbiotic Systems Project, ERATO, Japan Science and Tech. Corp.
M31 6-31-15 Jingumae, Shibuya, Tokyo 150-0001, Japan
e-mail: okuno@nue.org, nakadai@symbio.jst.go.jp

Abstract In this paper, we present an active audition system which is implemented on the humanoid robot *"SIG the humanoid"*. The audition system for highly intelligent humanoids localize sound sources and recognize auditory events in the auditory scene. Active audition reported in this paper enables *SIG* to track sound sources by integrating audition, vision, and motor movements. Given the multiple sound sources in the auditory scene, *SIG* actively moves its head to improve localization by aligning microphones orthogonal to the sound source and by capturing the possible sound sources by vision. However, such an active head movement inevitably creates motor noises. The system adaptively cancels motor noises using motor control signals. The experimental result demonstrates that active audition by integration of audition, vision, and motor control attains sound source tracking in variety of conditions.

key words robot audition, computational auditory scene analysis, auditory epipolar geometry, active audition, sensor fusion

聴覚は人間にとって最も重要な感覚である．言語によるコミュニケーションが聴覚によって成立することは容易に理解されるが，「ヒトは聴覚によってのみ言語を獲得し，そこに文化が生まれ，継承される．書かれた言語は日によって伝承されるが，話す言葉は耳からしか得られない．話し言葉があって書く言葉が生まれる」ことを，多くの人が理解していないのは残念なことである．

鈴木淳一，小林武夫共著
『耳科学 — 難聴に挑む』(中公新書 1598, 2001)

1 はじめに

新千年紀に入り，ヒューマノイド (人間型ロボット) やペットロボットが数多く開発されるようになった．それまでは，早稲田大学で 30 年間以上にわたる一連のヒューマノイド研究で開発された二足歩行ロボット Wabian や人間との協調作業を行う Hadaly など，ホンダで 10 年間以上にわたり開発されてきた P3 に至る一連のヒューマノイドぐらいであった．今日では，ホンダの ASIMO や科学技術振興事業団北野共生システムプロジェクトの PINO はテレビ放送のコマーシャルにも登場している．科学技術振興事業団川人プロジェクトの油圧式制御による DB は，神経科学の成果を生かし，複雑な運動制御ができ，見る人を驚かせる．映画スターウォーズに登場する古典的なロボット R2D2 の形をした Papero や Robovie などに加え，エンターテインメント分野に目を転じれば，PINO だけでなく，犬型の AIBO，猫型のネコロ，熊型の Latte，カンガルー型などのロボットも次々と登場してきている．

多数登場しているロボットには，「耳」に相当する機能が備わっており，そのためにマイクロフォンが装備されている．その本数は，2 本，ないし，3 本のものが多いが，マイクロフォンアレイを搭載しているものもある．しかし，視覚機能が，ボールを認識し，追跡できるようなレベルにあるのに対して，聴覚機能は，例えば，音声によるインターラクションをパソコンで提供されている音声認識システムとして比較しても，相当機能が低いと言わざるをえない．

本稿では，ロボット音響についての研究課題について検討し，我々が現在行っている研究を報告する．

2 ロボット聴覚に求められる機能

我々は，ロボット聴覚やコンピュータオーディションを設計する前提として，入力音は複数の音源から到達した混合音であり，かつ，

「主たる音源の数よりも，ロボットやシステムが装備するマイクロフォンの数が少ない」

という状況を想定している．

2.1 何本のマイクロフォンを使用するか

混合音の分離には，マイクロフォンアレイが一般的に使用される．マイクロフォンアレイで得られたマルチチャネルのデータから特定方向の音だけを抽出するには，ビームフォーミングやがよく使われる．その原理は，『$N+1$ 本のマイクロフォンを使えば，N 個の音響的な死角を作成できる』というものである．ビーフフォーミング技法にはさまざまな手法があるが，音源分離でよく使用されるのは，遅延型加算 (delayed sum) である．

すべての音源が情報論的に互いに独立であるとすると，独立成分解析 (ICA, Independent Component Analysis) を使えば，N 本のマイクロフォンで N 個の音源を理論的に分離することができる [1]．実際，2 話者同時発話から調波構造を抽出し，方向情報でグルーピングし，さらに非調波構造部分は入力音から調波構造を取り去った残差で補填する手法により音声を分離する方法と比較して，独立成分解釈の方が，分離音の音声認識結果がよいという結果が得られている [2]．

Wang らも複数のマイクロフォンを使用することによって，音響ストリーム分離の精度の向上を達成している [3]．つまり，複数のマイクロフォンを使用すれば，単一マイクロフォンよりも音源分離の性能が向上することが理論的にも，実験的にも明らかになっている．

一般環境では，マイクロフォンの位置が変化したり，音源が移動したり，未知の音響環境に置かれた時など，音源の個数以上にマイクロフォンが用意されていたとしても，必ずしも理論だけでは解決できないことが多い．ロボットの場合，体が動くことが前提であるため，ロボットの体に装着したマイクロフォンは，頻繁に動くことになり，また，マイクロフォン間の相対位置も同様に頻繁に変化するであろう．また，ロボットの体に何十本というマイクロフォンを装着することも現実的ではない．もちろん，このような状況に対して，適用可能なマイクロフォンアレイを開発しようというプロジェクトが学術創成研究「言語理解と行動制御」で始まっている．

我々は，このような問題を「マイクロフォンの数が音源の数よりも少ない時に，音源分離を行うにはどのようにしたらよいのか」ととらえ，研究を進めてきた．視覚の例を取ると，片目でも頭を動かせば，3 次元位置が分かることは日常よく経験することである．さらに，眼球の一箇所をじっと見ているつもりでも，実は眼球は細かく動いているという「固視微動」により，単眼であっても奥行きを検知することができる．

本谷らは, イメージセンサを微動させることにより, 3次元情報を取得する固視微動型イメージセンサの開発を行っている [4]. 聴覚でも同様に, 頭を動かせば, 片耳でも音源方向情報の取得が可能である. 読者の方々も, 他方の耳にマスキング信号を入れて, 片耳で方向情報が分かることを実感していただきたい. なお, もう一方の耳にマスキング信号を入れず, 耳栓をするだけでは, 骨伝達があるので片耳で聞くという実験条件として不十分である. また, 動的に振幅やピッチが短時間で変化する音に対しては, 頭の回転に対する音源方向の感度が悪いので, 方向情報を取ることは難しい. 片耳では画像のような汎用的な微動型センサーの実現は難しいと思われる.

我々は, 2本のマイクロフォンでも体を動かせば, 多くの状況に対応できるのではないかと考えた. これは, 人間や動物の耳が2つであることが多いという事実を勘案している. 方向情報を安定して得るためには, 最低2本のマイクロフォンは必要である. 2本のマイクロフォンでは方向は分かっても, 前か後ろかは分からないという「前後問題」を解決するためにもう1本のマイクフォンを装備しているロボットもある [5]. 我々は, 前後問題は, 視覚情報を使用したり, 体を動かしたりすることにより解決しようと考えている.

2.2 具体的なロボット聴覚の研究課題

2本のマイクロフォンを使用したロボット聴覚の研究課題として, 我々は次のような項目に重点を置いてきた.

1. **音環境理解 (CASA)** — 特定の音に特化しない一般の音の理解・認知機構
2. **アクティブオーディション**— 動作を伴った音の知覚・認識,
3. **動きながら聞く機構**— ロボット自身の発生する音の抑制機構
4. **未知環境での音の知覚**— 頭部伝達関数を使用しない音源定位・音源分離
5. **画像処理や他の処理との情報統合**
6. **実時間処理**

音環境理解とアクティブオーディション： 従来の音の研究は, 音声や楽音に特化してきたが, 我々が日常聞く音は様々な音源から届く混合音である. 音による事象の認知である音環境理解 (Computational Auditory Scene Analysis, CASA) は, 音源分離や混合音の理解の機構を追求する工学的なアプローチである [6].

アクティブビジョンが焦点距離, ズームイン・アウト, 解像度, 虹彩, などのカメラパラメータをアクティブに制御し, 視覚情報を効率よく正確に取得する手法である [7] のと同様に, アクティブオーディションは, マイクロフォンの位置, 指向性, などをアクティブに制御することによって, 音環境理解の精度を向上させようと試みである [8].

実環境で動き回たり, 頭部が動かせるロボット, あるいは, 行動と知覚が結び付いたアクティブパーセプションでは, カメラやマイクロフォン自体が動き, それに伴って, モータ雑音や機械音が発生する. このような音は, たとえ小さくてもマイクロフォンに近いので, 相対的に大きな雑音となり, 外部からの音の信号雑音比が低下する.

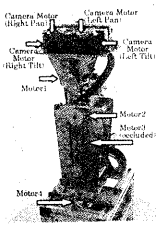
ロボットやシステムが発生する内部雑音を軽減する最も簡便な方法は, 動作を中断してから, 聞くこと, つまり, "*stop-perceive-act*" 法である. この方法は, マイクロフォンを搭載した大部分のロボットが採用している. また, アクティブオーディションでは, よく聞こうとして動いたところ, 自分の出す音が災いして, 返って聞こえ難くなるということも想定される. アグティブオーディションでは, 内部雑音抑制, あるいは, 自己生成音の抑制は極めて重要である. また, ロボットヒューマンインターラクションでは, 自分の発話を削除し, 相手の発話の信号雑音比を向上させることも必要である.

音源定位や音源分離の手法： 音源分離の一つの手法は, 頭部伝達関数 (Head-Related Transfer Function, HRTF) を使用することである. 左右の耳 (マイクロフォン) からの入力から両耳間位相差 (Interaural Phase Difference, IPD) や, 両耳間強度差 (Interaural Intensity Difference, IID) を求め, それぞれの角度の HRTF との相関を求めて, 方向を決定し, 分離を行う. 中谷らは, 混合音中の調波構造のピークを抽出し, それに対する IPD や IID を求めて, 調波構造を持つ音を無響室環境で分離している [9].

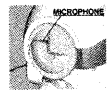
実環境で HRTF を使用するにはいくつかの問題点がある. 例えば, 部屋の伝達特性が必要であり, また, 離散的な点の HRTF しか測定されていないので, 動く音源には使うのが難しい. 我々は, HRTF の代用として**聴覚エピポーラ幾何学** (auditory epipolar geometry) を提案した [10]. これは, ステレオ画像処理でのエピポーラ幾何学の焼き直しであるが, マイクロフォン間の距離は, *SIG* の頭部上の距離を使用している.

さらに, 特定の音源方向から来る音だけを分離する方向通過型フィルタ (Direction-Pass Filter, DPF) を設計した. DPF における方向推定は, IPD と IID に関する仮説推論で行う. 具体的には, 入力音の各サブバンド (離散フーリエ変換, DFT の各点) で IPD と IID を求める. 一方, 所与の方向に対する IPD と IID を聴覚エピポーラ幾何学で求める. 次に, IPD と IID について両者の距離

a) 外装



b) 内部機構



c) マイクロホンの実装部

図 1: ヒューマノイド *SIG*

を計算し, 確率密度関数を用いて, それらを確信度に変換する. 最後に, 2 つの確信度を Dempster-Shafer 理論で統合し, 統合確信度の高いサブバンドだけを集めて, 逆離散フーリエ変換で分離音を再構成する. 無響室環境における DPF による 3 話者同時発話認識では、ほぼ単一音源並の単語認識率が達成されている [11]。

3 2本のマイクロフォンによる音響と画像を統合した実時間複数話者追跡

本稿では, 2 つの耳でどのような機能を実現できるかを実証するために, 図 1 に示したヒューマノイド *SIG* の上に実現されている複数話者実時間追跡システム [12] を紹介する. このシステムは, 繁華街の大通りに面したマンションの一室 ($3m \times 3m$) に置かれている.

3.1 ヒューマノイド *SIG*

SIG は, 4 自由度を有した上半身ロボットである. アクチュエータには, 図 1b) に示す内部機構のように, ポテンショメータによって速度・位置制御が可能な DC モータを使用している. 視覚センサは, CCD カメラ (Sony EVI-G20) であり, 聴覚センサは, 2 組の無指向性のマイク (Sony ECM-77S) である. 外側にある一組のマイクは, 図 1c) に示す耳の位置に設置されており, *SIG* 外部からだけの音を集めるように隔離されている. 内部に置かれたもう 1 組のマイクは, 主に *SIG* 内部の音を集める. 内部機構は 1a に示す外装に覆われ, ユーザフレンドリにするだけではなく, 内部音が外部に漏れるのを防いでいる.

3.2 実時間複数話者追跡システム

システム全体の構成を図 2 に示す. システムは, 音響処理部, 画像処理部, モータ駆動部, アソシエーション部, 対話管理部, 注意制御部, および, サーバから構成されている. これらは 5 つの PC 上に配置されており, Gigabit Ethernet と Fast Ethernet で接続されている. 音響処理部は DPF と同じ考え方で, 音源方向を抽出する. ただし, 画像からの音源方向が得られない場合には, すべての方向について仮説を生成することによって, 音源方向を抽出する. 抽出した方向情報は, 確信度とともにアソシエーション部に送られる. 画像処理部は, 肌色抽出により顔を発見し, ステレオ画像処理により, 3 次元情報を取得する. さらに, 抽出した顔ごとに顔認識を行い, 3 次元情報と顔 ID 情報をアソシエーション部に送る. モータ駆動部からは, 現在の *SIG* の体の向きについての情報をアソシエーション部に送る.

SIG では, 体内に有する 1 組のマイクロフォンから取得されるモータ音や機械音については簡単なモデルを持っており, モータが稼働中でモデルに合うような音が発生すると, ヒューリスティクスを用いて, 破壊されているサブバンドを破棄するようにしている. FIR フィルタを応用したアクティブノイズキャンセレーションを検討したが, IPD を計算するために必要な左右の位相特性の線形性が実データでは成立しないので, ヒューリスティクスによる方法を採用した.

アソシエーション部では, 各モジュールから得られる情報 (方向や顔) を同期させ, 音響ストリームと画像ストリームを構築する. 次に, 音響ストリームと画像ストリームを, 時間的継続性や距離的近さを基にグルーピングをし, アソシエーションストリームを構成する. 一定時間, 音響ストリームや画像ストリームが消失すると, アソシエーションは解除される.

このようにして, システムは, 内部的には複数の話者や人物の位置を常時把握している. 2 名の話者追跡のベンチマークに対して, 得られた音響ストリーム, 画像ストリーム, および, アソシエーションストリーム (太い線) を図 3 に示す.

3.3 注意制御部

注意制御部は, システムが保持する話者や人物情報をもとに, どの対象に注意を向け, 正対するかを制御する. この部分は完全にプログラム可能であるので, いくつかのシナリオで, *SIG* の挙動を紹介する.

受付嬢： 話している人に注意を向けるのが第一目的であり, そうでなければ, 音のする方向に振り向く. このため, 注意を置くべきストリームの優先順位は, アソシエーションストリーム $\succ$ 音響ストリーム $\succ$ 画像ストリームとなっている.

具体的な挙動を図 4 に示す. この例では, 来客が知らな

い人なので,「どちらさまですか」という応答をしているが, 既知の人の場合には,「こんにちは, XX さんでいらっしゃいますか」とその人の名前を呼び, 確認を行う. このように, 対話管理部は音声認識と音声合成を含んでいる. 音声認識には京都大学開発の Julian を使用し, 音声合成には市販のソフトを使用している.

コンパニオン： 音のする方に注意を払うように, **音響ストリーム ≻ アソシエーションストリーム ≻ 画像ストリーム** という優先順位で, 注意制御を行う. 図 5 a) にその実験風景を示す.

仮想音源： 追跡するものは, 人物だけではなく, 仮想的な音源であっても良い. 基本周波数が 100 Hz の調波構造を持つ音を左右のスピーカから流し, そのバランスコントロールを変化させることで, 仮想的に音源を左右にふる. *SIG* は, その変化に応じて, 仮想的な音源を正しく追跡する (図 5 b)).

実験のまとめ： これらの例から, 2 本のマイクロフォンでも, 画像処理と統合すると, 実時間で音源定位を行い, 話者追跡を行える. 2 本のマイクロフォンの音響処理だけでは, 方角は分かるが, それが前か後かが分からないという前後問題に対しては, 図 3 に示したおうに, 画像情報により曖昧性の解消ができるし, アクティブオーディションにより, 体を回転した時に, 音源がどちらに動くかで解消することができる.

4 おわりに

本稿では, ロボット聴覚に対して, マイクロフォン 2 本でできる機能について, 著者の考え方も含めて解説をした. 体を動かして聞くというアクティブオーディション, あるいは, 画像処理とモータ処理を統合して, 体全体で聞くという情報統合が重要である. そのために, 方向通過型フィルタや聴覚エピポーラ幾何学, 実時間処理方法を開発してきた. 今後の研究課題としては, 複数話者の話者認識や音声認識が挙げられる. それらのフトントエンドとしての精度の高い音源分離, 雑音の混入により復元できないミッシングデータをうまく扱うことのできる新しい話者認識や音声認識の研究が必要であろう.

最初に引用した節からも, 聴覚機能というのは人間生活にとって極めて重要であり, そのような原点に戻り, 音響処理の新しい枠組を打ち立てることが, 音声認識システムが一般的となってきた今求められてことではないであろうか. 本稿が, そのような新しい潮流の一助となれば幸いである.

謝辞 北野共生システムプロジェクト総括責任者北野宏明氏, 元同僚の松井龍哉氏, Tino Lourens 博士, 日台健一氏, および, 音声認識システム Julian を提供いただいた河原達也京都大学大学院助教授に感謝をする. 本研究の一部は, 科研費萌芽的研究・特定研究 (C) および, NTTCS 基礎研究所の援助を受けた.

参考文献

[1] N. Murata and S. Ikeda. An on-line algorithm for blind source separation on speech signals. In *Proceedings of 1998 International Symposium on Nonlinear Theory and its Applications*, pp. 923–927, 1998.

[2] 奥乃博. 音環境理解 — 混合音の認識を目指して. 情報処理, Vol. 40, No. 10, pp. 1096–1101, Oct. 2000.

[3] F. Wang, Y. Takeuchi, N. Ohnishi, and N. Sugie. A mobile robot with active localization and discrimination of a sound source. *Journal of Robotic Society of Japan*, Vol. 15, No. 2, pp. 61–67, 1997.

[4] 本谷秀堅, 来海暁, 安藤繁. 固視微動型イメージセンサとその応用. 研究報告「コンピュータビジョンとイメージメディア」, No. 118–002, 1999.

[5] J. Huang. Spatial sound processing for a hearing robot. In *Enabling Society with Information Technology*, Vol. Lecture Notes in Computer Science. Springer-Verlag, Nov. 2001.

[6] D. Rosenthal and H. G. Okuno, editors. *Computational Auditory Scene Analysis*. Lawrence Erlbaum Associates, Mahwah, New Jersey, 1998.

[7] Y. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International Journal of Computer Vision*, Vol. 1, No. 4, pp. 333–356, 1987.

[8] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano. Active audition for humanoid. In *Proceedings of 17th National Conference on Artificial Intelligence (AAAI-2000)*, pp. 832–839. AAAI, 2000.

[9] T. Nakatani and H. G. Okuno. Harmonic sound stream segregation using localization and its application to speech stream segregation. *Speech Communication*, Vol. 27, No. 3-4, pp. 209–222, 1999.

[10] K. Nakadai, H. G. Okuno, and H. Kitano. Epipolar geometry based sound localization and extraction for humanoid audition. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2001)*. IEEE, Oct. 2001.

[11] H.G. Okuno, K. Nakadai, T. Lourens, and H. Kitano. Separating three simultaneous speeches with two microphones by integrating auditory and visual processing. In *Proceedings of International Conference on Speech Processing (Eurospeech 2001)*, pp. 2643–2646. ESCA, Sep. 2001.

[12] K. Nakadai, K. Hidai, H. Mizoguchi, H.G. Okuno, and H. Kitano. Real-time auditory and visual multiple-object tracking for robots. In *Proceedints of the Seventeenth International Joint Conference on Atificial Intelligence (IJCAI-01)*, pp. 1424–1432. MIT Press, 2001.

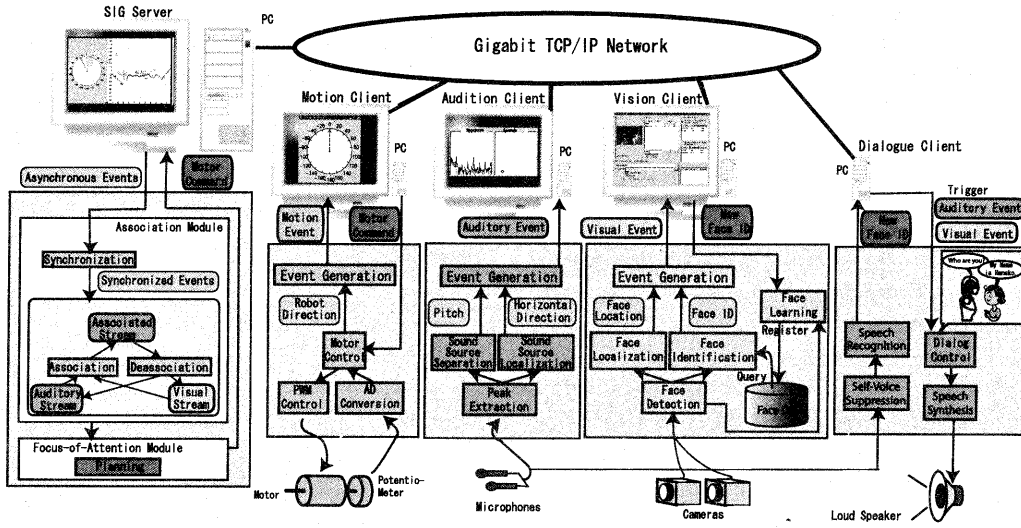


図 2: 視聴覚統合による実時間話者追跡システム

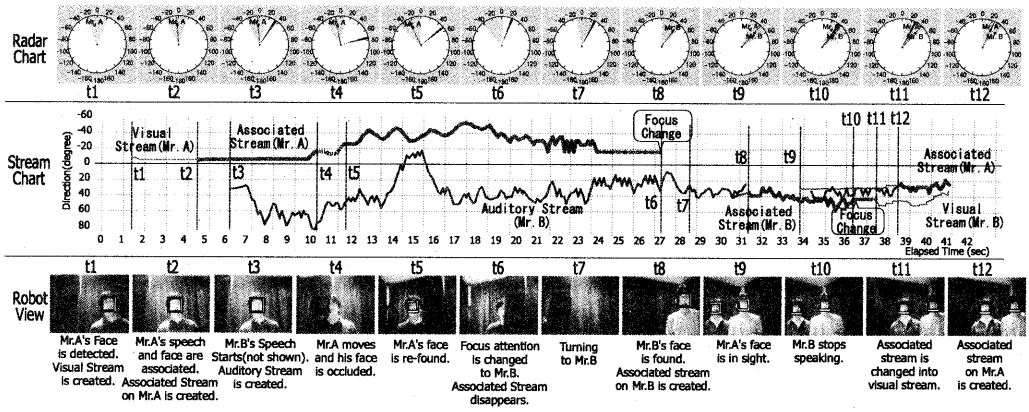
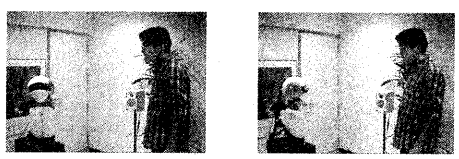
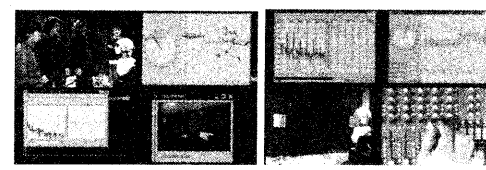


図 3: 音響と画像の統合による 2 話者追跡の時間経過: **Radar Chart** と **Stream Chart** は viewer のスナップショットである. radar chart 上の幅広い扇形は, カメラの視野を示し, 狭く鋭い扇形は音源方向を示す. stream chart では, 細い線は音響ストリームか画像ストリームを表し, 太い線はアソシエートされたストリームを示す.



a) 声のする方に振り向き, 顔認識をする. b) 顔認識に基づいて応答し, 来客者の確認をする.

図 4: 受付嬢としての *SIG* の振舞い



a) 4 人の中で声のする方を向く. b) ステレオバランスコントロールを変化に追従.

図 5: *SIG* の 音源追跡のさまざまな評価実験