学術・技術論文

ヒューマノイドを対象にした視聴覚統合による実時間人物追跡
— アクティブオーディションと顔認識の統合 —

中臺 一博[*1*2] 日台 健一[*1*3] 溝口 博[*4]
奥乃 博[*1*5] 北野 宏明[*1]

Real-Time Human Tracking by Audio-Visual Integration for Humanoids
— Integration of Active Audition and Face Recognition —

Kazuhiro Nakadai[*1*2], Ken-ichi Hidai[*1*3], Hiroshi Mizoguchi[*4],
Hiroshi G. Okuno[*1*5] *and* Hiroaki Kitano[*1]

This paper describes a real-time human tracking system by audio-visual integrtation for the humanoid *SIG*. An essential idea for real-time and robust tracking is hierarchical integration of multi-modal information. The system creates three kinds of streams – auditory, visual and associated streams. An auditory stream with sound source direction is formed as temporal series of events from audition module which localizes multiple sound sources and cancels motor noise from a pair of microphones. A visual stream with a face ID and its 3D-position is formed as temporal series of events from vision module by combining face detection, face identification and face localization by stereo vision. Auditory and visual streams are associated into an associated stream, a higher level representation according to their proximity. Because the associated stream disambiguates parcially missing information in auditory or visual streams, "focus-of-attention" control of *SIG* works well enough to robust human tracking. These processes are executed in real-time with the delay of 200 msec using off-the-shelf PCs distributed via TCP/IP. As a result, robust human tracking is attained even when the person is visually occluded and simultaneous speeches occur.

Key Words: Active Audition, Sensor Fusion, Sound Source Localization, Face Recognition, Stereo Vision, Humanoids, Sound Source Separation, Real-Time Tracking

1. はじめに

昨今，Sony AIBOに代表されるペットロボットから，HONDA ASHIMO，Sony SDR-4Xといったヒューマノイドに至るまで様々なロボットが世間を賑わすようになり，研究のみならず世間一般にロボットが脚光を浴びている．これらのロボットはスムーズな2足歩行や愛らしい仕草をするが，現時点では，"人間のパートナー"として人間と知的なソーシャルインタラクションを行うことは難しい．

知的なソーシャルインタラクションを行うロボットを実現するためには，単なるアクチュエータ制御だけでなく，視覚や聴覚といったセンサ情報の処理を適切に行い，ロバストな知覚や認識を行う必要がある．例えば，声のする方向を向いたり，特定の人物に注意を向けるためには，捉えた画像や音響信号から顔や声を抽出，同定して名前や位置を認識することが必要である．また，追跡中に注意を向けていた人物が，後ろを向いてしまったり，物陰に隠れたり，他の人物と交差したりするような場合にも，継続して追跡を行えるように複数のセンサ情報を統合してロバストな状況把握を行うことが必要である．さらに，同時に複数の人を判別するためには，複数の顔を識別したり，複数の声やノイズが混在している状況下でも，カクテルパーティ効果として知られるように特定の音源に注意を向けつづける能力も求められる．視覚，聴覚，およびこれらを統合した処理は，ソーシャルインタラクションを行うための最低限必要な機能といえる[2]．

ロボットでは，視覚処理，および視覚とモータ制御を統合する処理は，アクティブビジョンやビジュアルサーボに代表されるように盛んに研究が行われている分野である．しかし，聴覚は人間では視覚と同様に主要なセンサであるにもかかわらず，あまり取り組まれていないのが現状である．この原因として，聴

原稿受付 2002年3月27日
*1科学技術振興事業団 ERATO 北野共生システムプロジェクト
*2現在は株式会社ホンダ・リサーチ・インスティチュート・ジャパン所属
*3現在はソニー株式会社所属
*4東京理科大学理工学部機械工学科
*5京都大学大学院情報学研究科
*1Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Corp.
*2Currently with Honda Research Institute Japan, Co., Ltd.
*3Currently with Sony Corp.
*4Faculty of Science and Technology, Science University of Tokyo
*5Graduate School of Informatics, Kyoto University

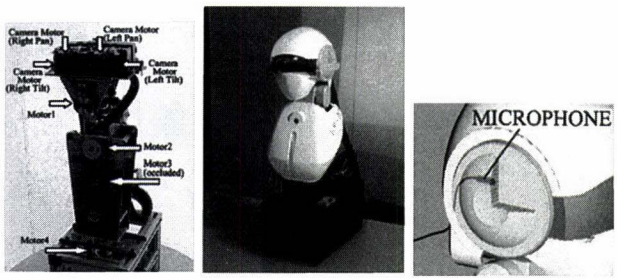
覚のセンサ情報としての扱いにくさが考えられる．

実環境では，指向性マイクロホンを使用しても，他の複数の音源からの音の混入を防ぐことは難しいことに加え，聴覚情報は部屋の反響や環境の変化に敏感であるため，視覚ほど高精度の処理を行うことが難しい．特にロボットにおいては，マイクロホンが他の音源と比較しモータに近い位置に配置されることが多いため，自分自身が発生するモータノイズが大きなノイズ源となる．このため，多くのロボットは，いったん停止してから音を聞く“stop-perceive-act”原理に従って行動せざるを得ない．

例えば，聴覚機能を備えたロボットとして，早稲田大学の *WA-2*[11]，*Hadaly*[8]，*ROBOITA*[5]，MIT AI Lab の *Kismet*[1]，東京大学の *SmartHead*[7] が挙げられる．*WA-2* では，頭部動作を利用した純音に対する三次元空間の定位機能を実現している．しかし，調波構造を使用することにより，定位の精度を向上させる機能を欠いているので，混合音中の複数の主要な音の定位や動きながら定位をすることができていない．*Hadaly* では，頭部に設置された 2 本のマイクロホンとカメラを用いた動き抽出を組み合わせて，複数の話者が存在している状況下で話者発見や定位が実現されている．しかし，基本的に 1 対 1 のコミュニケーションを念頭において設計されているため，音源分離能力を備えておらず，同時発話や動きながらの定位が困難である．*ROBOITA* は音声認識と共に，相手の視線を追跡することで，話者チェンジを抽出し，話者の方向を向き対話を行うことができる．また，*Kismet* は音声認識により，話者と対話を行うとともに，刺激に応じて様々な感情を仕草や言葉を用いて表現することができる．しかし，これらのロボットは，話者の口許に備えたマイクロホンを使って，部屋の音響環境による影響や“stop-perceive-act”原理を避けている．*SmartHead* は 4 本のマイクロホンとステレオカメラを用いて抽出される低レベルな情報を統合し，複数音源の定位，ならびに追跡を可能にしている．しかし，低レベルの情報のみを扱っているため，話者が交差するような複雑な状況への対応は難しい．また，頭部が音響信号を妨げないことが前提となっているため，頭部形状による音の歪み（頭部伝達関数）を考慮しなくてはならない形状をもったロボットには適用することが難しく，理論上，分離における最大音源数も制限されている．

このような問題点を解決し，一般的なロボット聴覚を実現するため**アクティブオーディション**が提案されている[6]．アクティブオーディションでは，聴覚，視覚，およびモータ制御を統合することによって，複数音源の追跡を向上させることができる．また，アクティブな動作により発生するモータノイズをモータ制御信号とヒューリスティクスを利用してキャンセルし“stop-perceive-act”原理を回避している．予備実験では，ロボットに搭載されたマイクロホンを使用した複数音源の追跡に有効であることが示されているが，純音のみを扱っていたため，実時間・実環境で動作をさせることは難しかった．

そこで，本稿では，複数の聴覚的な手がかりを統合し，実時間・実環境で複数音源を定位できるロバストなアクティブオーディションを実現した．さらに，アクティブオーディションに複数顔同時認識とステレオビジョンによる視覚処理を統合することによって，聴覚の曖昧性を解消するだけではなく，視覚の視野の狭さ，オクルージョンも解決できるロバストな実時間複数人物追跡システムをヒューマノイド *SIG* 上に実現した．



(a) Mechanical structure (b) Cover (c) Microphone

Fig. 1 *SIG* the humanoid

以下，2，3 章では，本システムに用いたヒューマノイドとシステムの詳細について述べる．4 章でシステムの実験と評価を行い，5 章でまとめる．

2. 実時間人物追跡システム

構築した実時間人物追跡システムは，研究のテストベッドとして使用しているヒューマノイド *SIG* と，*SIG* のセンサ情報を処理し，その行動を制御する処理部からなっている．以下にこれらを説明する．

2.1 ヒューマノイド *SIG*

研究のテストベッドとして，**Fig. 1** に示される上半身ヒューマノイド *SIG* を使用している[6]．外装には FRP を用い (Fig. 1 (b))，音響的に内部と外部の世界を区別できるようデザインされている．*SIG* は外界からの音響信号を収音するよう耳の位置に一対のマイクロホンを，またモータ動作音などの内部ノイズを収音し，ノイズキャンセルに利用するため，外装の内部に一対のマイクロホンを備えている (Fig. 1 (c))．マイクロホンはすべて無指向性マイクロホン (Sony ECM-77S) を使用している．左右の目の位置には，一対の CCD カメラ (Sony EVI-G20) を備えており，顔認識およびステレオ視による顔の定位を行うことができる．また，*SIG* は，4 自由度を有し，各モータには，DC モータを用い，ポテンシオメータによって位置，速度の制御が可能になっている (Fig. 1 (a))．なお，本稿では，実際に追跡で使用するモータは，ロボットの水平方向の回転用のモータ (Fig. 1 (a) の Motor4) のみである．

2.2 処理部の構成と処理の概略

処理部のシステム構成を **Fig. 2** に示す．システム内のモジュール群や情報は，*SIG* デバイス層，プロセス層，特徴層，イベント層，ストリーム層の五つに分けられており，Fig. 2 の上にいくほど，高次の情報や処理を扱うような階層的な枠組みを備えている．

実装上は，「聴覚モジュール」，「視覚モジュール」，「アソシエーションモジュール」，「注意制御モジュール」，「モータ制御モジュール」，「ビューワモジュール」 の大きく六つのモジュールから構成されている．各モジュールは複数のサブモジュールから構成され，モジュールの内部およびモジュール間では様々なレ

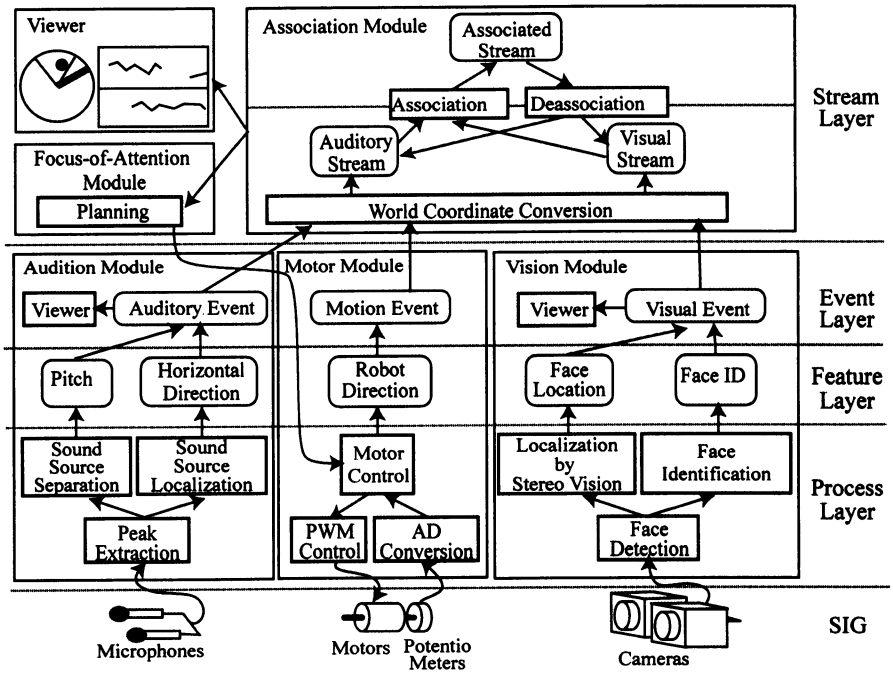


Fig. 2 The system architecture for real-time human tracking

ベルの情報の通信が非同期に発生する．システムを実時間で動作させるために，これら六つのモジュールは Gigabit Ethernet で接続された 3 台の Pentium III 1GHz Linux ノードに分散されている．モジュールの配置については，視覚モジュール，聴覚モジュールは CPU パワーを必要とするため，それぞれ別のノードに配置し，その他のモジュールは，残りのノードに配置している．また，ノード間の情報通信は，TCP/IP によって行われるが，ノード間で正確な同期をとるために，各ノードは NTP（Network Time Protocol）によって同期を行った上で，各モジュールの起動時に再同期を行い，モジュール間の誤差が 100 [μs] 以内になるようにしている．

聴覚および視覚モジュールは，マイクロホンおよびカメラの入力信号に対し，特徴抽出を行い，観測時刻とともに，位置，人名，音高（ピッチ）情報などを含んだ聴覚および視覚イベントを生成し，アソシエーションモジュールへ送出する．また，モータ制御モジュールはモータ方向情報をイベントとしてアソシエーションモジュールへ送出すると共に注意制御モジュールの要求に従い，PWM（Pulse Width Modulation）信号を生成し，DC モータを駆動する．

アソシエーションモジュールは，聴覚，視覚モジュールから送られたイベントを時間的なつながりを考慮して接続し，聴覚および視覚ストリームを生成する．さらに結びつきの強い聴覚ストリームと視覚ストリームを結びつけ，より高次の表現であるアソシエーションストリームを生成する．逆に，アソシエーションストリームを形成する聴覚ストリームと視覚ストリームの結びつきが弱くなれば，アソシエーションは解除される．

注意制御モジュールは，ストリームの状態やアソシエーションストリームが存在するかどうかに基づいて，*SIG* の動作のプランニングを行う．プランニングの結果，モータを動作させる必要があれば，モータ制御モジュールへモータ駆動用のモータイベントを送出する．

ビューワモジュールでは，聴覚，視覚，アソシエーションストリームをレーダーチャートとストリームチャートの 2 種類のビューワを使って表示することができる．また聴覚モジュール，視覚モジュール，モータ制御モジュールのそれぞれに，各モジュールで生成されるイベントを表示できるビューワを用意し，内部状態を把握しやすいインタフェースを提供している．

次章では，各モジュールの詳細な処理について説明する．

3. 各モジュールの詳細

システムの中核をなす聴覚，視覚，アソシエーション，注意制御モジュールについて詳細な処理を説明する．

3.1 聴覚モジュール—アクティブオーディション

一般に，両耳聴における音源定位には，頭部伝達関数（HRTF）を計測して得られる両耳間位相差（IPD）と両耳間強度差（IID）が用いられる．HRTF は無響室において各方向からのインパルス応答を計測することによって得られる離散的な関数であるため，HRTF に基づいた方法を実環境でロボットに適用する場合，以下の点を考慮する必要がある．

- 実環境では，環境の伝達系による影響を考慮し，無響室で得られた HRTF に対して，その環境の伝達関数を畳み込む必要性がある．
- ロボットの場合は，部屋を移動したり，向きを変えたりと動的に環境が変化するため，環境の伝達関数もその都度再計測を要する．
- HRTF は計測によって得られるため離散的な関数であるが，動作中における音源定位を円滑に行うためには，連続的な定位が必要であり，そのためには，測定点間での HRTF の

補完が必要である．

したがって，実環境アプリケーションには，HRTF を用いず，連続的に音源定位を行うことができる方法が望ましい．アクティブオーディションでは，そのような音源定位法の一つとして，聴覚エピポーラ幾何が提案されている [6]．聴覚エピポーラ幾何は，ステレオビジョンにおけるエピポーラ幾何を聴覚に適用した音源定位手法であり，音源から左右のマイクロホンまでの距離の差を位相差として検出することで水平方向の定位を行う．この際，**Fig. 3** に示されるように，頭の形状による音の回りこみを考慮に入れている．ロボット頭部が半径 r の球体であること，音源が無限遠であること（Fig. 3 における l が無限大）を仮定し，音速を v，入力信号の周波数を f，入力から得られた IPD を $\Delta\varphi$ とした場合，音源方向 θ は，式 (1) によって求めることができる．なお，$D(\theta)$ は音源方向が与えられた際の音源から左右のマイクロホンまでの距離差を表している．

$$\theta = D^{-1}\left(\frac{v}{2\pi f}\Delta\varphi\right),$$
$$D(\theta) = r\,(\theta + \sin\theta) \qquad (1)$$

このように，聴覚エピポーラ幾何を使用すると，事前測定を必要としない音源方向推定が可能である．しかし，定位に使用する特徴は IPD のみで，単一周波数を扱っていたため調波構造を有した音を扱えないなど汎用性や精度の面で乏しかった．そこで，本稿では，倍音成分すべてについて，IPD，IID という複数の聴覚的情報を Dempster-Shafer 理論により統合するように拡張することにより，定位のロバスト性を向上させた．

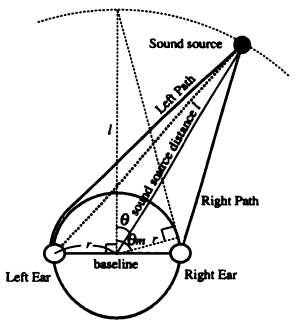


Fig. 3 Auditory epipolar geometry

聴覚モジュールにおける処理の詳細を **Fig. 4** に示す．聴覚モジュールでは，入力信号は，異なる方向からの混合音を仮定しており，48 [kHz]，16 ビットでサンプリングされる．その後，高速フーリエ変換（FFT）による周波数解析を行い，左右のチャネルごとにスペクトルを生成する．

3.1.1 ピーク抽出と音源分離

左右のチャネルのうち，パワーの大きい方のスペクトルを取り出し，パワーが閾値（部屋の暗騒音の音圧レベルに，感度パラメータ 10 [dB] を加えた値）以上のローカルピークを抽出する．ここで，感度パラメータは実験的に求めた値であり，部屋の暗騒音は，システムにより自動的に計測される．この際，抽出されたローカルピークのうち低周波ノイズとパワーの小さい高周波域をカットするため，90 [Hz] から 3 [kHz] の周波数を持ったローカルピークのみを抽出する．次に，周波数が低い方から順番に抽出されたローカルピークを取り出し，その周波数を $F0$ として，式 (2) を満たすような周波数 Fn を持つローカルピークを，$F0$ の倍音としてクラスタリングを行う．

$$\left| \frac{Fn}{F0} - \left\lfloor \frac{Fn}{F0} \right\rfloor \right| \leq 0.06 \qquad (2)$$

ここで，定数 0.06 は，実験的に求めた値である．これにより最終的に集められた n 個のピーク周波数の集合を

$$\{f_h(i) | i = 0, 1, \cdots, n-1\} \qquad (3)$$

とすると，f_h を分離した一つの音とみなすことができる．ここで便宜上，f_h に含まれる周波数値は昇べきの順に並んでいるものとする．f_h に含まれる周波数に対応するスペクトル成分を抽出して，逆FFTを適用すれば，混合音から音源ごとの音響信号を分離・再合成することができる．

3.1.2 音源定位

抽出した調波構造を有した音に対し，IPD と IID を用いて音源方向（方位角）を計算する．

まず，f_h の各倍音ごとに IPD $\Delta\varphi_s$ を計算する．$\Delta\varphi_s$ は，左右のチャネルから同じ $F0$ 値を持つローカルピークの集合 f_h を選択し，対応する各倍音の位相差を計算することによって得

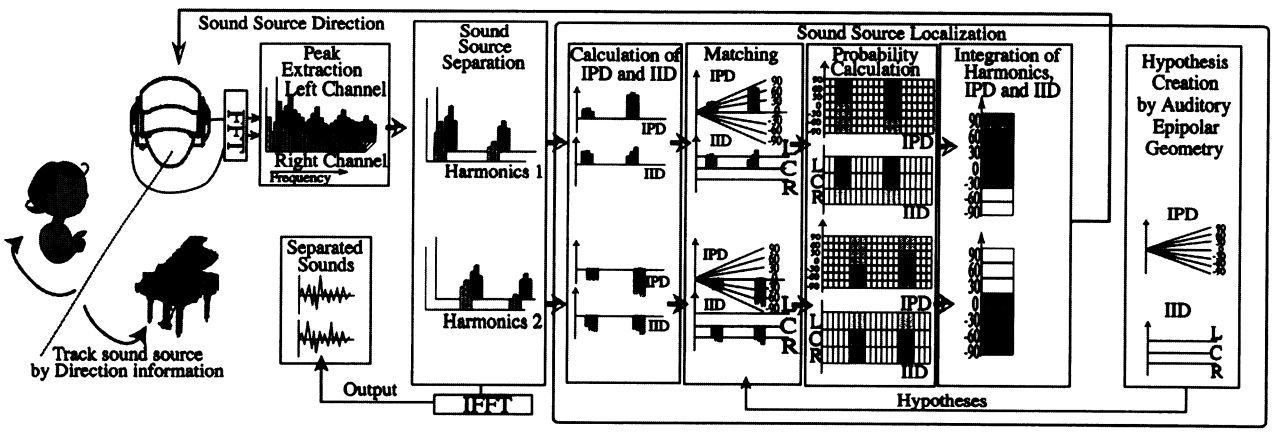


Fig. 4 Sound source separation and localization in auditory module

Table 1 Belief factor of IID, $B_{\mathrm{IID}}(\theta)$

θ		90° ~ 35°	30° ~ −30°	−35° ~ −90°
S_I	+	0.35	0.5	0.65
	-	0.65	0.5	0.35

ることができる．ただし，IPD による定位は，回り込みが発生するような高い周波数域では有効性が薄れてしまう．*SIG*では，左右のマイクロホン間距離が約 18 [cm] であることから，計算上，1.2～1.5 [kHz] の間にその閾値が存在するため，本稿では，IPD は，1.5 [kHz] 以下の倍音についてのみ扱うものとする．

次に，式（1）に示される聴覚エピポーラ幾何を変形すると，IPD が音源方向 θ と周波数 f の関数になることを利用して，±90°（*SIG* 正面の方向を 0° とし，右手が正，左手が負の値とする）の範囲で 5° おきに，$\Delta\varphi_s$ に対応する IPD 仮説 $\Delta\varphi_h$ を生成する．最後に，式（4）に定義された距離関数により，入力音と各仮説間の距離（$d(\theta)$）を計算する．ここで，n_{th} は周波数が 1.5 [kHz] 以下である倍音数である．

$$d(\theta)=\frac{1}{n_{th}}\sum_{i=0}^{n_{th}-1}\frac{(\Delta\varphi_h(\theta,f_h(i))-\Delta\varphi_s(i))^2}{f_h(i)} \quad (4)$$

IID に関しては，IPD と同様に入力音の各倍音の左右チャネル間パワー差から求める．ただし，IID については，仮説推論ではなく，式（5）で示される判別関数を用いて，音源が*SIG*の正面方向か左側か右側かのみを判定するものとする．つまり，周波数が f である倍音の IID を $\Delta I_s(f)$ とした場合，S_I が正なら音源は*SIG*の左方向に存在し，0 に近ければ正面方向，負であれば右方向に存在するものとする．

$$S_I=\sum_{i=n_{th}}^{n-1}\Delta I_s(f_h(i)) \quad (5)$$

IID の仮説生成には，頭部の形状を考慮した膨大な計算が必要となり，実時間性を考慮して IPD と同様の仮説推論は行わない．

3.1.3 Dempster-Shafer 理論による IPD と IID の統合

IPD，IID から得られる音源方向を支持する値から，これらを Dempster-Shafer 理論によって統合しロバストな音源方向を推定するために，それぞれの値を確信度に変換する．IPD の確信度 B_{IPD} は式（4）によって得られた距離に対し，式（6）で定義される確率密度関数を適用することによって得られる．

$$B_{\mathrm{IPD}}(\theta)=\int_{-\infty}^{\frac{d(\theta)-m}{\sqrt{\frac{s}{n}}}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)dx \quad (6)$$

ここで，m と s は，それぞれ，$d(\theta)$ の平均と分散であり，n は d の個数である．

IIDの確信度 $B_{\mathrm{IID}}(\theta)$ は，式（5）の S_I の値に応じて，**Table 1** のように定義した．

これらによって得られた値をそれぞれ IPD と IID から音源方向を支持する確信度と捉え，式（7）で示される Dempster-Shafer 理論によって，IID と IPD の確信度を統合し，IPD と IID の両方から音源方向を支持する新しい確信度を生成する．

$$B_{\mathrm{IPD+IID}}(\theta)=B_{\mathrm{IPD}}(\theta)B_{\mathrm{IID}}(\theta)+\big(1-B_{\mathrm{IPD}}(\theta)\big)B_{\mathrm{IID}}(\theta)+B_{\mathrm{IPD}}(\theta)\big(1-B_{\mathrm{IID}}(\theta)\big) \quad (7)$$

最終的に，聴覚モジュールは入力音の音源方向として尤度の高い順に 上位 20 個の $B_{\mathrm{IPD+IID}}$ と方向（θ）のリストをピッチ（$F0$）と共に，聴覚イベントとして生成する．

3.2 視覚モジュール—実時間複数顔認識・定位

視覚モジュールは，複数の顔の発見・定位・同定を行い，その結果を視覚イベントとしてアソシエーションモジュールへ送信する．顔発見サブモジュールは，肌色抽出と相関演算に基づくパターンマッチングの組み合わせにより，顔の位置，大きさ，明るさの動的変化にロバストな抽出，および顔が複数存在する場合でも 200 [ms] 以内にすべての顔領域検出が可能である [3]．個人同定サブモジュールは，切り出された顔領域画像を判別空間に射影し，事前登録された顔データとの距離を求め，これを不完全ガンマ関数を用いて確信度 P_v に変換する．判別空間の基底となる判別行列は，オンライン LDA によって求める [4]．オンライン LDA では，通常の LDA と比べ少ない計算で判別行列の更新が可能であり，実時間に顔を登録することが可能である．顔定位サブモジュールでは，ステレオ視によって各発見顔の定位を行う．まず，事前にアフィン変換を用いた補正を施した上で，局所領域マッチングによる対応点探索に基づいて視差画像を生成する．この際，PC 上で実時間処理を達成するため，再帰相関演算手法と Intel アーキテクチャ固有の最適化 [9] を用いている．定位は，視差画像中の顔領域に対してエピポーラ幾何を用いて行われる．

最終的に視覚モジュールは，各顔ごとに，上位五つの確信度付きの顔 ID（名前）と位置（距離，方位角，仰角）からなる視覚イベントを生成する．

3.3 アソシエーションモジュール—情報の統合

アソシエーションモジュールは，*SIG* がロバストに周囲の状況を把握するために，様々なイベント情報を統合し，ストリーム，およびアソシエーションストリームを生成する．ストリームはイベントを時間方向に接続することによって生成され，アソシエーションストリームは，ストリーム間の状態によって発生するアソシエーションによって生成される高次のストリームである．アソシエーションのプロセスを **Fig. 5** に示す．

3.3.1 イベントの絶対座標変換

Fig. 5 (a) は各モジュールから送られてくるイベントの座標を絶対座標に変換するために使用する 2 秒間の短期記憶を示している．Fig. 5 (a) において [M], [S], [V] はそれぞれ各モジュールから非同期で送られるモータ，聴覚，視覚イベントを示している． 聴覚，視覚イベントに含まれる位置情報は，イベント観測時刻のロボットから見た座標系（*SIG* 座標系）における情報である．そこで，モータイベントを利用してこれらのイベントを絶対座標へ変換する．しかし，**Table 2** に示されるように各イベントは遅延時間，到着周期が異なる非同期イベントであるため，各イベントをいったん，2 秒間の短期記憶に格納する．まず，座標変換すべき聴覚，または視覚イベントの発生時刻に近接する二つのモータイベントを一次線形補間することによって該当イベント発生時刻のモータ方向を推定し，その方向，時刻情報を含んだモータイベント [M] を生成する．次に，[M] のモータ

方向を利用して，該当イベントの位置情報を絶対座標系に変換し，変換後の聴覚イベント [S]，視覚イベント [V] を生成する．

3.3.2 ストリームの生成，消滅

ストリームは，Fig. 5 (b) に示すように，短期記憶から座標変換したイベントを取り出し，時間方向に接続することによって生成される．図の横軸は経過時間，縦軸はロボットを中心にした絶対座標での水平角を示している．また，黒点は聴覚イベント，白丸は視覚イベントを示しており，これらをつなげた1本の線が1本のストリーム（太線，細線はそれぞれ視覚，聴覚ストリーム）に対応している．短期記憶からイベントを取り出す際は，ネットワークや処理のレイテンシを考慮して，200 [ms] の遅延を持つようにイベントを取り出す．このため，Fig. 5 (b) では，最新時刻から 200 [ms] の間にはストリームやイベントは存在しない．短期記憶から取り出されたイベントとストリームの接続には，次のアルゴリズムを用いる．

- **聴覚イベント**：音高が，同等もしくは倍音関係にあり，方向が ±10° 以内で最も近い既存の聴覚ストリームに接続される．この値は，聴覚エピポーラ幾何の精度を考慮し定めた値である．また，倍音関係を許容するのは，基音の抽出誤りを補完するためである．
- **視覚イベント**：視覚イベントは，共通の顔IDをもち，40 [cm] の範囲内で最も近い既存の視覚ストリームに接続される．この値は，4 [m/s] 以上で人間が移動しないことを前提にして定めている．

すべての既存ストリームに対して探索を行った結果，ストリームと接続できないイベントが存在した場合，そのイベントから新規にストリームが生成される．ただし，視覚イベントに関しては，すでに存在している視覚ストリームの顔ID と同じ顔 ID を持ったストリームは生成せず，第二候補以降の顔 ID を持ったストリームが生成される．また，既存ストリームは，接続するイベントがまったく存在しない場合でも，最大 500 [ms] 間は存続が可能であるが，500 [ms] 以上イベントが接続されない場合，そのストリームは消滅する．このような時間の流れを考慮したストリーム形成により一時的な音や顔の抽出ミスを前後の情報によって補完できるという利点がある．

3.3.3 ストリームのアソシエーション

次に，ストリームのアソシエーション判定の前処理として，ストリーム間同期を行う．具体的には，Fig. 5 (b) におけるストリームに対し，100 [ms] 周期で再サンプリングを行う．サンプリング時刻のストリーム情報は，サンプリング時刻に近接するストリーム内の二つのイベント情報を一次線形補間して得ている．Fig. 5 (c) の□，○は 100 [ms] 周期で再サンプリングしたストリーム中の視覚，聴覚イベントを示している．この処理により，ストリーム間距離の算出処理を簡潔している．

同期後，ストリーム間距離を算出し，視覚，聴覚ストリームが同一の人物に由来すると判断された場合，二つのストリームがアソシエーションされ，高次のストリーム表現であるアソシエーションストリームが形成される．Fig. 5 (c) の太線は，視覚，聴覚ストリームからアソシエーションストリームが形成されたことを示している．アソシエーション条件は，『方位角が ±10° 以内に近接する状況が1秒間のうち 500 [ms] 以上観測される』ことであり，実験的に求めた．アソシエーションストリームを形成する視聴覚ストリームの一方が消滅した場合，アソシエーションストリームはデアソシエーションされ，存在するストリームのみが残る．また，アソシエーションストリームを構成する二つの視聴覚ストリームが 3 秒間以上にわたり，水平角の差が 30° 以上になった場合，誤ったアソシエーションであると判断され，デアソシエーションにより 2 本のストリームに分割される．

方位角（θ）は双方のストリームに含まれるが，センサの性質上，視覚情報は聴覚情報より精度が高い．このため，一見，精度の高い視覚情報だけ利用すれば，十分であるように考えられるが，実際には，情報を互いに補い合うことで，そのロバスト性を高めている（disambiguation）．例えば，視野外にいる人

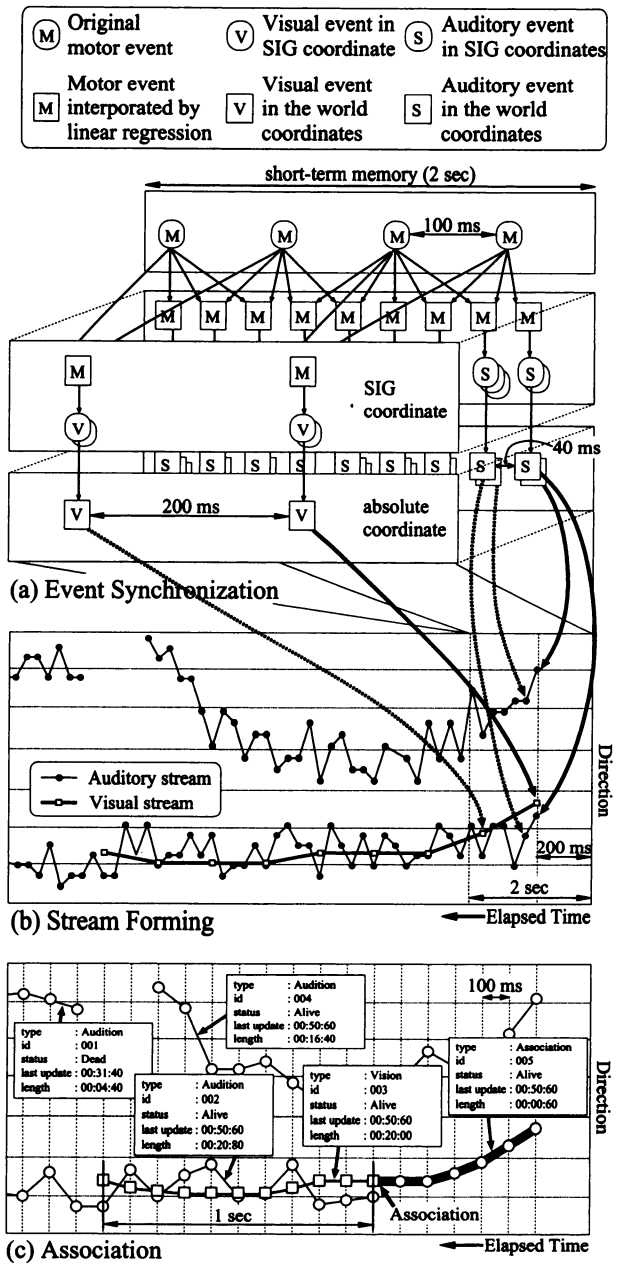


Fig. 5 Stream formation in association module

Table 2 Event generation cycle and latency

visual event	200 ms
auditory event	40 ms
motor event	100 ms
network latency	10 ～ 200 ms

からの声，物陰に隠れている人からの声は，視覚だけでは定位することができない．また，視覚モジュールでは壁にかけてある写真を，人物として抽出してしまうなどの誤抽出があり，アソシエーションによってこのようなエラーを訂正することができる．さらに，方位角しか分からない聴覚ストリームであっても，視覚ストリームとアソシエーションできれば，正確な方位角が取得できるばかりか，距離，仰角，顔IDといった情報を取得することができる．

3.4 注意制御モジュール—視聴覚サーボ

注意制御モジュールでは，アソシエーションモジュール内の任意のストリームを選択して*SIG*の行動を決定し，モータ制御モジュールへモータイベントを送出する．このモータ制御は，音や顔が抽出できる間は，ロボットの体の動きに伴うセンサの運動によって目標値が逐次更新されるので，センサ情報をフィードバックした制御系といえる．このフィードバック情報には，視聴覚両方の情報が含まれるため，この制御により**視聴覚サーボ**が実現されている．

注意制御のアルゴリズムはプログラマブルであり，状況に応じて変更が可能であるが，本稿では，以下のアルゴリズムにしたがって，注意を向けるストリームを決定している．

(1) アソシエーションストリームの追跡を最も優先する．
(2) アソシエーションストリームが存在しない場合，聴覚ストリームのトラッキングを優先する．
(3) アソシエーションストリームも聴覚ストリームも存在しない場合，視覚ストリームの追跡を優先する．

これは，人物の追跡に焦点を絞り，以下の2点を考慮に入れて採用したものである．

(A) アソシエーションストリームの存在は，*SIG*に正対して喋っている人物が現在も存在している，もしくは近い過去に存在していたことを示している．したがって，一般にそのような人物に対して，高い優先度でアテンションを向け，追跡を行うのは妥当である．
(B) マイクロホンは無指向であるためカメラのような視野角は存在せず，広範囲な情報を得ることができる．このため，視覚より聴覚のストリーム優先度を高くすべきである．

4. 実験と評価

本システムの評価を行うため，**Fig. 6**に示すシナリオをベンチマークとして使用した．このシナリオでは，2話者が約40秒にわたって様々なアクションを行う．具体的な2話者のアクションを以下に示す．なお，Fig. 6の Radar Chart，Stream Chart の方向値は絶対座標系での *SIG* の水平角を示し，Radar Chart，Robot View の t_n は Stream Chart の対応する時刻を示す．

t_1： A氏が*SIG*の視野内に入る．
t_2： A氏が*SIG*に話を始める．
t_3： B氏が*SIG*の視野外で話を始める．
t_4： A氏が動き，物陰に隠れる．
t_5： A氏が再び，物陰から現れる．その後，A氏は話を止め再び物陰に隠れる．
t_7： *SIG*が話をしているB氏の方向を向く．
t_8： *SIG*がB氏を視野内に捉える．
t_9： A氏も話しながら*SIG*の視野内に入ってくる．
t_{10}：B氏が話を止める．

Fig. 6より，このシナリオに対してシステムは以下のような特徴をもった動作を行った．

(1) 新しいアソシエーションストリームが生成されると優先的に*SIG*の注意が新アソシエーションストリームに向けられる［Fig. 6の t_1 と t_8］．
(2) オクルージョンにより，アソシエーションストリームの視覚情報が欠如してしまったが，アソシエーションが行われているため，聴覚情報により追跡が継続できる［Fig. 6の t_4，t_5 間］．
(3) アソシエーションストリームが消滅したため，アソシエーションストリームの次に優先度の高い聴覚ストリームに注意が向けられる［Fig. 6の t_6，t_{11}］．
(4) シナリオの26[s]以降は，2話者は同時にカメラの視野に移る程度（約20°）まで近づく．この場合でも，うまく話者の追跡が達成されている．

このシナリオにおける*SIG*の体の方向を**Fig. 7**に表す．Fig. 7は，2話者の場合においても注意制御モジュールおよびモータ制御モジュールが，適切なPWM信号を生成し，*SIG*の追跡動作をの制御に成功していることを示している．

視覚情報による追跡を**Fig. 8**に示す．Fig. 8は，Fig. 6の実験の際に視覚モジュールが生成した視覚イベントの第一候補のみを使用して作成したものである．そのため，モータ動作はFig. 7と同一である．Fig. 8の前半部分では，オクルージョンにより t_4 と t_5 の間では，視覚ストリームが分断されてしまう．また t_6 から t_7 までの間は，人が*SIG*の視野外にいるため視覚からは何も情報が得られない．このようにオクルージョンや視野外といった視覚だけでは解決できない場合でも，アソシエーションによって，簡単に解決できることをFig. 6は示している．

Fig. 9は視覚情報による追跡の場合と同様の方法で生成した聴覚による追跡結果を示している．聴覚モジュールは，t_3 から23[s]付近まで，および34[s]付近から t_{10} までの間，正しく2本の聴覚のストリームを分離することができているが，t_8 および t_9 の周辺では，誤ったストリームが生成されていることが分かる．また，11[s]（t_5）から17[s]までの間は，A氏の移動および*SIG*の体の回転が同時に起きているため，Fig. 9の2人の話者の定位は，それほど正確ではない．つまり，話者の移動やモータノイズ，とその反響音（エコー）が，音源定位の品質を悪化させている．このような場合でも，視覚情報によって聴覚情報の曖昧性が解決できていることをFig. 6は示している．

5. 結　　論

本稿では，視覚情報，聴覚情報，モータ制御を統合し，複数の顔を実環境で実時間で追跡できるシステムを構築した．実時

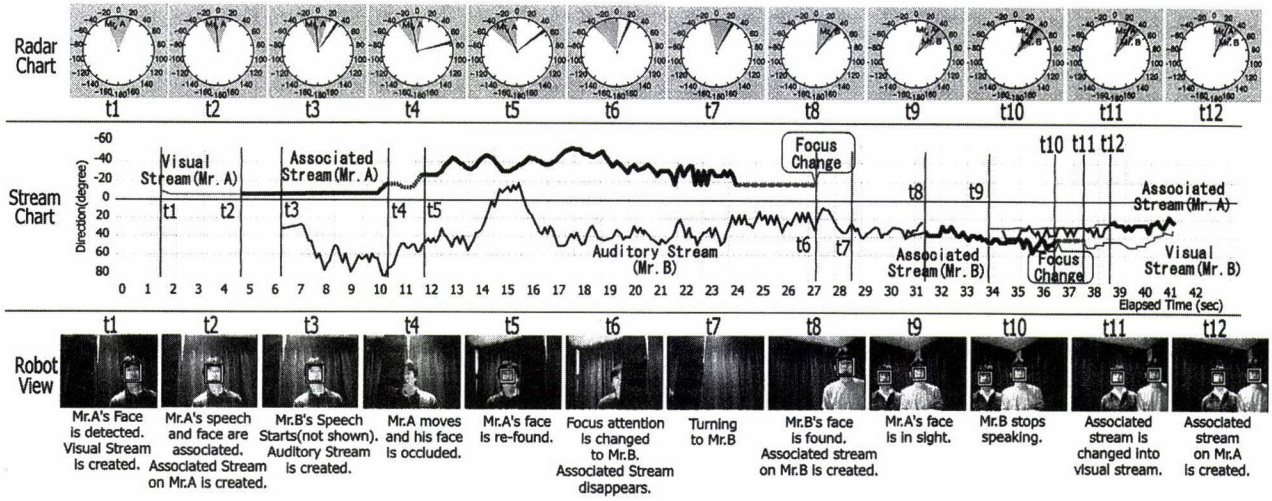


Fig. 6 Temporal sequence of auduitory and visual tracking of two speakers: **Radar Chart** and **Stream Chart** are screen shots of the viewer. In radar chart, a wide-light and a narrow-dark sector indicate the camera view field and sound source direction, respectively. In stream chart, a thin line indicates auditory or visual stream, while a thick line indicates an associated stream

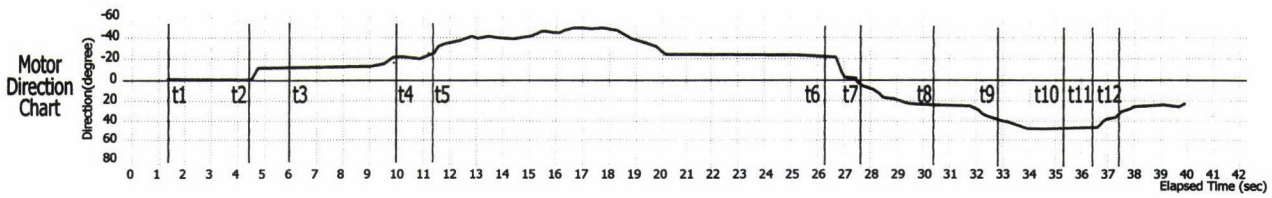


Fig. 7 Temporal sequence of body direction controlled by motor movement in the same scenario as Fig. 6

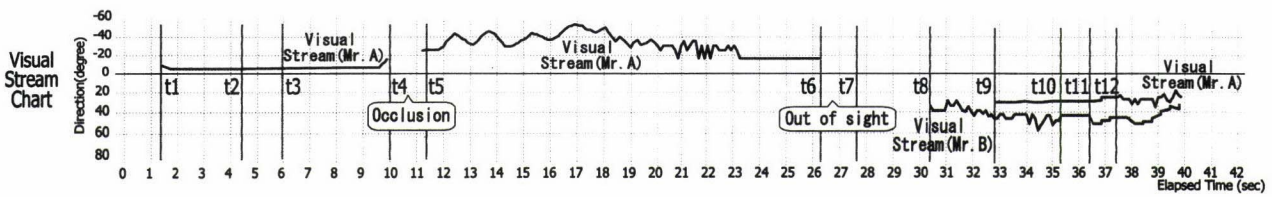


Fig. 8 Temporal sequence of visual tracking of two speakers in the same scenario as Fig. 6

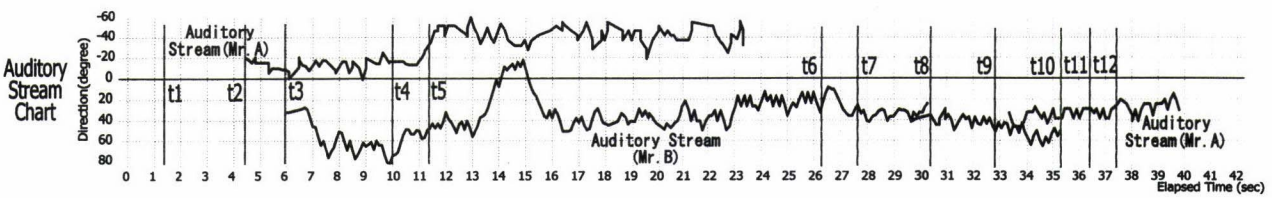


Fig. 9 Temporal sequence of auditory tracking of two speakers in the same scenario as Fig. 6

間で実環境で動作させるため，個々の処理は，速度を優先した実装を行っている．この速度優先の実装と実環境であるがゆえの難しさのため，各処理はある程度の処理精度の不足が生じてしまう．本稿では，この問題を様々な情報を統合することによって，解決している．例えば聴覚モジュールでは各周波数ごとの音源定位を正確に行うのではなく，音の倍音構造，IPD，IIDといった複数の聴覚情報を利用して，精度の不足を補っている．また，正確な顔識別や定位を行う代わりに聴覚情報，モータ情報などマルチモーダルな情報を イベントの流れ（ストリーム）に基づいた統合を行うことにより，精度不足を補っている．

実環境の実時間アプリケーションにおける情報統合の有効性は，実験を通じて，様々な状況に対応できるロバストな処理が実現できたことによって証明できた．特に，音環境理解（CASA）に基づいた聴覚情報処理は，ロボットに対してそれ自体有効であるばかりか，視覚情報の視野の不足を補うという意味でも有効であることを示すことができた．

将来的に，ロボットに人間との知的なソーシャルインタラクションを求めるためには，話者同定や触覚情報といったさらに

多くのセンサ情報を統合することにより，知覚のロバスト性を高めることが必要であろう．そのためには，様々な情報を透過的に，また階層的に扱う統合の仕組みも必要であろう．さらに，未知環境や動的に変化する環境へ対応するための学習の枠組み，より複雑な状況に対応できる注意制御ポリシーも必要であろう．

謝　辞　科学技術振興事業団 ERATO 北野プロジェクトのメンバに感謝する．元北野プロジェクトの研究員 Dr. Tino Lourens に感謝する．

参 考 文 献

[1] C. Breazeal and B. Scassellati: "A context-dependent attention system for a social robot," In Proc. of 16th Int'l Joint Conf. on Artificial Intelligence (IJCAI-99), pp.1146–1151, 1999.

[2] R. A. Brooks, C. Breazeal, R. Irie, C. C. Kemp, M. Marjanovic, B. Scassellati and M. M. Williamson: "Alternative essences of intelligence," In Proc. of 15th National Conf. on Artificial Intelligence (AAAI-98), pp.961–968. AAAI, 1998.

[3] K. Hidai, H. Mizoguchi, K. Hiraoka, M. Tanaka, T. Shigehara and T. Mishima: "Robust face detection against brightness fluctuation and size variation," In Proc. of IEEE/RAS Int'l Conf. on Intelligent Robots and Systems (IROS-2000), pp.1384–1397. IEEE, 2000.

[4] K. Hiraoka, S. Yoshizawa, K. Hidai, M. Hamahira, H. Mizoguchi and T. Mishima: "Convergence analysis of online linear discriminant analysis," In Proc. of IEEE/INNS/ENNS Int'l Joint Conf. on Neural Networks, III–387–391. IEEE, 2000.

[5] Y. Matsusaka, T. Tojo, S. Kuota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano and T. Kobayashi: "Multi-person conversation via multi-modal interface — a robot who communicates with multi-user," In Proc. of 6th European Conf. on Speech Communication Technology (EUROSPEECH-99), pp.1723–1726. ESCA, 1999.

[6] K. Nakadai, T. Lourens, H. G. Okuno and H. Kitano: "Active audition for humanoid," In Proc. of 17th National Conf. on Artificial Intelligence (AAAI-2000), pp.832–839. AAAI, 2000.

[7] S. Ando: "An Autonomous Three-Dimensional Vision Sensor with Ears," IAPR Workshop on Machine Vision Applications (MVA-94), pp.417–422, 1994.

[8] T. Kurata, D. Chang and S. Hashimoto: "Multimedia Sensing System for Robot," IEEE Int'l Workshop on Robot and Human Communication, pp.83–88, 1995.

[9] 岡田慧，加賀美聡，稲葉雅幸，井上博允："PCによる高速対応点探索に基づくロボット搭載可能な実時間視差画像・フロー生成法と実現"，日本ロボット学会誌，vol.18, no.6, pp.896–901, 2000.

[10] 中臺一博，日台健一，溝口博，奥乃博，北野宏明："顔認識とアクティブオーディションを利用した実時間人物追跡"，第11回AIチャレンジ研究会，SIG-Challenge-01-5, pp.27–34, 2001.

[11] 高西淳夫，升川聡，森寧，小川太郎："人間形聴覚ロボットの研究"，第11回日本ロボット学会学術講演会予稿集，pp.793–796, 1994.

中臺一博（Kazuhiro Nakadai）

1970年10月21日生．1993年東京大学工学部電気工学科卒業，1995年同大学大学院工学系研究科情報工学専攻修了．同年日本電信電話株式会社入社，1997年NTTコムウェア（株）出向後，1999年退職．同年7月より科学技術振興事業団 ERATO 北野共生システムプロジェクト研究員．2003年5月より，株式会社ホンダ・リサーチ・インスティチュート・ジャパンシニア・リサーチャー．主にロボット聴覚，実時間情報統合，音環境理解の研究に従事．2001年度日本人工知能学会研究奨励賞，IROS 2001 BEST Paper Nomination Finalist，2002年電気通信普及財団賞奨励賞など受賞．（日本ロボット学会正会員）

溝口　博（Hiroshi Mizoguchi）

1956年9月22日生．1980年東京大学工学部計数工学科卒業．1985年同大学大学院工学系研究科情報工学専攻博士課程修了．工学博士．同年（株）東芝入社．東京大学先端科学技術研究センター，埼玉大学工学部を経て，2002年より東京理科大学機械工学科．人間と機械との実世界中でのインタラクションの研究に従事．（日本ロボット学会正会員）

北野宏明（Hiroaki Kitano）

1961年3月16日生．1984年国際基督教大学教養学部理学科（物理学専攻）卒業．同年NEC入社．1988年CMU客員研究員．1991年京都大学博士（工学）．1993年ソニーCSLリサーチャー．1996年同研究所シニアリサーチャー．1998年科学技術振興事業団ERATO北野共生システムプロジェクト総括責任者．2002年ソニーCSL取締役副所長．ロボカップ国際委員会ファウンディング・プレジデント．国際レスキューシステム研究機構理事長．慶應義塾大学大学院理工学研究科非常勤教授．1993年IJCAIより Computers & Thought Award 受賞．著編書：『ロボカップレスキュー』（共立出版，2000），『大人のための徹底！ロボット学』（PHP研究所，2001）等．

日台健一（Ken-ichi Hidai）

1976年5月10日生．1999年埼玉大学工学部情報システム工学科卒業．2001年埼玉大学大学院理工学研究科情報システム工学専攻修士課程修了．同年4月より，科学技術振興事業団 ERATO 北野共生システムプロジェクト技術員．同年10月より Sony（株）勤務．

奥乃　博（Hiroshi G. Okuno）

1950年1月12日生．1972年東京大学教養学部基礎科学科卒業．日本電信電話公社，NTT，科学技術振興事業団北野共生システムプロジェクト，東京理科大学理工学部情報科学科教授を経て，2001年より京都大学大学院情報学研究科知能情報学専攻教授．博士（工学）．この間，スタンフォード大学客員研究員，東京大学工学部客員助教授．音環境理解，ロボット聴覚の研究に従事．人工知能学会1990年度論文賞，IEEE/RSJ IROS-2001中村賞finalist．著編書："Advanced Lisp Technology"（共編，Taylor & Francis，2002），"Computational Auditory Scene Analysis"（共編，LEA，1998）『インターネット活用術』（岩波書店，1996）等．（日本ロボット学会正会員）