

ヒューマノイドにおける聴覚機能の課題と アクティブオーディションによる音源定位

Issues in Humanoid Audition and Sound Source Localization by Active Audition

中臺 一博
Kazuhiro Nakadai

科学技術振興事業団 ERATO 北野共生システムプロジェクト
Kitano Symbiotic Systems Project, ERATO, JST
nakadai@symbio.jst.go.jp, <http://www.symbio.jst.go.jp/~nakadai/>

奥乃 博
Hiroshi G. Okuno

京都大学大学院情報学研究科
Department of Intelligence Science and Technology, Graduate School of Kyoto University
okuno@nue.org, <http://www.symbio.jst.go.jp/~okuno/>

北野 宏明
Hiroaki Kitano

科学技術振興事業団 ERATO 北野共生システムプロジェクト
Kitano Symbiotic Systems Project, ERATO, JST
kitano@symbio.jst.go.jp, <http://www.symbio.jst.go.jp/~kitano/>

keywords: active audition, robots, perception, noise cancelation, sensor fusion

Summary

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 localizes sound sources and recognizes auditory events in the auditory scene. Active audition reported in this paper enables SIG to track 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 and the cover acoustics. The experimental result demonstrates that active audition by integration of audition, vision, and motor control attains sound source tracking in variety of conditions.

1. はじめに

近年、ロボットは、RoboCup^{*1} など、AI を応用するためのプラットフォームとして盛んに研究が行われている。従来からの産業用ロボットだけではなく、ペットロボットやヒューマノイドが相次いで発表され、大衆に身近な存在となってきていることもこのような研究を促している一因であろう。しかし、実際にロボットに AI を応用して、例えば、周囲の状況をロボストに知覚することは、簡単な問題ではなく、長年にわたり様々なアプローチが行われてきた。このような問題に対して、アクティブパーセプションは、知覚と動作を統合し、情景分析 (Scene Analysis) の向上やロボストな知覚の実現への糸口を与える重要な研究である [Aloimonos 93, Bajcsy 88]。

ビジョンの分野では、アクティブパーセプションは、アクティブビジョンとして盛んに研究が行なわれている [Aloimonos 87, 久野 95]。アクティブビジョンは、例え

ば、焦点、ズームイン・アウト、解像度、虹彩、ステレオカメラの輻輳および開散といった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。つまり、視覚と動作を統合することによって視覚情景分析 (Visual Scene Analysis) を向上する枠組みを提供している。

これは、聴覚においても重要な概念である。例えば、カクテルパーティ効果として知られるように、人は複数の音源からの音声やノイズが混在した状況下でも、これを分離し、特定の音源に注意を向け続けることができる。また、音源方向を向き、視覚から得た音源情報を聴覚の向上に役立てている。さらに、頭部の動きが前後判断や音源定位の精度向上に寄与しているという知見も得られている [Makous 90, Perrett 97, 植松 01]。つまり、人間は常にアクティブな動作を伴うことによって聴覚情景分析 (Auditory Scene Analysis) を向上させている。

しかし、これまでロボットに対してアクティブな動作を積極的に用い、聴覚情景分析を向上させる試みはあまり見られなかった。そこで、本稿では、聴覚とモータ動

*1 <http://www.robocup.org/>

作を統合し、ロボットを対象にアクティブオーディションを提唱する。そして、アクティブオーディションを実現する上でのロボット聴覚処理における課題を整理し、アクティブオーディションに基づき、動作中の音源定位を実現するような実験的なシステムを構築し、その有効性を評価する。なお、本稿で扱うアクティブオーディションは、パッシブなセンサ(マイク)を対象としているため、蝙蝠やイルカが発する超音波のようにアクティブなセンサを用いるアクティブセンシングとは区別して扱うものとする。

以降、2章では、ロボットの聴覚処理について議論し、3章では、我々が扱うヒューマノイド SIG を紹介する。4章では、構築したアクティブオーディションシステムを説明し、5章でその評価を行う。6章では考察と課題について言及し、7章でむすぶ。

2. ロボットの聴覚処理における課題

従来のロボット、例えば、*ROBITA* [Matsusaka 99] や *Kismet* [Breazeal 99] では、音声認識機能を用いて、人間とのインタラクションをテーマとした研究が行われている。前者は音声認識と共に、相手の視線を追跡して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、後者は刺激に応じて様々な感情を仕草や言葉で表現することができる。しかし、これらのロボットは、ロボット自身に備わったマイクを利用せず、話者の口元に備えたマイクを使って、部屋の音響環境やロボット自身のノイズによる影響を避けている。

Hadaly [Kurata 95] は、頭部に設置された2本のマイクとカメラによる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と定位を実現している。また、Haung らは、球形の頭部を持ったロボットに3本のマイクを用いることで、2本のマイクを使用する際に生じる前後問題を解決し、音源のトラッキングを実現している [Huang 97]。しかし、これらは、単一音源を仮定しており、同時発話の扱いや雑音下での定位など音源分離が必要な場合の処理が困難である。また、動きながら定位を行うことも難しく、一旦、静止してから音を聴く“*stop-perceive-act*”原理に従っている。この制約は、マイクロホンアレイを用いて簡単なフレーズによる命令を認識する *Jijo-2* [Asoh 97] にも当てはまる。

SmartHead [Ando 95] は頭部の4本のマイクとステレオカメラを用いて低レベルな情報を抽出し、これらを統合することによって、複数音源の定位・追跡を可能にした。しかし、理論上、頭部形状は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)を考慮する必要があるロボットには適用することが難しく、分離における最大音源数も制限されている。また、動作時のノイズ問題は触れられていない。

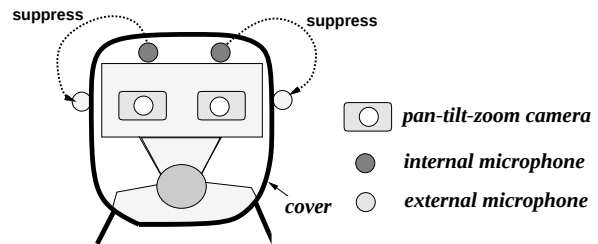


図1 内部ノイズ抑制に使用する内外のマイク

WA-2 [高西 94] では、ロボットに備わったマイクを使用する際のモータノイズの問題に対して取り組み、機構的な改良を行い、無響室で単一音源の連続定位に成功している。しかし、実環境アプリケーションを考えると、自身の声の認識・抑制や、何かと衝突した場合のノイズ抑制など、機構的な改良だけでは完全な対処は難しい。

以上をまとめると、ロボットの特徴であるアクティブな動作による恩恵を最大限に享受し、アクティブオーディションを実現するためには、以下の4つの課題に取り組む必要があることがわかる。次節でそれぞれの課題について議論する。

- (1) ロボット自身が発生する音の抑制
- (2) 未知環境における音の知覚
- (3) 様々なセンサ情報の統合
- (4) 特定の音に特化しない一般の音の理解・認知機構(一般音理解)

2.1 内部音の抑制

アクティブオーディションは、アクティブな動作によりロボット聴覚を向上させるが、このようなアクティブな動作は、モータ、ギア、ベルト、ベアリングの回転などによる動作時ノイズを発生させ、聴覚処理の妨げてしまうという問題を作り出してしまふ。さらに、このようなノイズ源は外部の音源と比較しマイクに近く減衰が小さい。このため、ノイズ源のパワーが小さくても、相対的にパワーが大きくなり、入力信号に大きな影響を与えてしまふ。以上より、従来のロボットの多くは、前述したように、その影響を無視できるようにパワーの大きい音源をターゲットにしたり、話者の口元など音源の側にマイクを設置して対処するか、“*stop-perceive-act*”原理に従わざるを得ない。しかし、動きながら聴いたり、動くことによって聴覚を向上させるためには、内部ノイズを抑制し、外部音を強調することが必要である。

このような問題に対し、図1に示すように、外装によりロボットの内部と外部を隔て、内部の音が外部に漏れるのを防ぎ、ロボットに音響的な身体性をもたせることは、内部音を抑制する有効な手段である。

2.2 未知環境における音源定位

未知環境における音の知覚に関しては、ロボット頭部に 2 本のマイクを設置するという前提で、音源定位問題を扱う。

一般に、両耳聴の音源定位には、頭部伝達関数 (HRTF) が用いられることが多い。HRTF は、頭部形状による音響特性の変化を表す伝達関数であり、通常は、無響室に設置したダミーヘッド等に対し、各方向からのインパルス応答を測定することによって取得する。しかし、ロボットでは、音源定位に無響室で計測した HRTF を利用した場合、以下の 2 つの問題が生じる。このため、ロボットにおける音源定位では、HRTF を使用しない音源定位の手法が有効である。

- (1) HRTF に基づく音源定位。無響室で測定した HRTF を実環境の音源定位に適用すると、部屋の反響により、著しく定位精度が低下する。さらに、実環境では頭部の形状によって生じる純粋な HRTF と部屋の反響などによる伝達関数を区別して計測することは難しく、部屋の反響まで考慮した広義の HRTF が必要である。ところが、このような HRTF の測定は時間を要するばかりでなく、部屋に新しい家具が設置されたり、ロボットの位置や向きが変わったりするなど、環境が変わるたびに再測定が必要となる。
- (2) HRTF に基づく音源追跡。HRTF は測定関数であるため測定点は離散値にならざるを得ない。そのため、HRTF による音源定位を利用して、連続的に動く音源の追跡を実現することは難しい。

2.3 聴覚処理におけるセンサフュージョン

知覚を向上させるため、様々なセンサ情報を統合することは、聴覚処理においても本質的である。例えば、Aarabi らは、3 本のマイクロホンによる静止したアレイを用いて、聴覚情報の時間方向の統合、ステレオカメラによる視覚定位情報と統合することにより、S/N 比が 0.5dB という雑音下で 15 cm 程度の誤差で複数音源の定位を実現している [Aarabi 01]。また、浅井らは、視聴覚情報と聴覚による音源定位誤差を取得するために対象音源を視覚的に捕らえるような頭部の回転動作を組み合わせ、音限定能力を学習によって獲得するモデルを提案し、その有効性を示している [浅井 00]。

このような情報統合は、音源定位だけではなく、音源分離にも有効である。例えば、それぞれ独立した音源からの 2 本の音響ストリームが存在し、この 2 本の音響ストリームの周波数が交差するような場合の音源分離を考えてみよう。このような場合、倍音構造のみを分離の手がかりにしたのでは、2 本の音響ストリームの周波数が交差しているのか、交差せず近づいて離れているのかという曖昧性を解消することが困難である。中谷らは、このような倍音関係だけでは解決できないような音源分離問題に対して、HRTF によって得られた音源方向を統合

して、その曖昧性を解消し、聴覚情報統合の有効性を示している [Nakatani 95]。さらに、中川らは、音響処理よりも画像処理によって得られる方向情報の方が正確であることに着目し、Bi-HBSS を改良した [Nakagawa 99]。具体的には、音響ストリームを抽出する際に、画像処理による正確な方向と HRTF による方向通過型フィルタを使用して音源分離の精度を向上させた。また、実際に 2 本のマイクを使用して 3 話者同時発話の分離を行い、音声認識システムのフロントエンドに適用し、分離の正確さを示している。

この他、Hershey らが行った信号レベルでの相関を利用した視聴覚統合に関する研究 [Hershey 00]、池田らの相互情報量を計算する際の時間窓長を適応させた視聴覚統合の研究 [池田 01] でも視聴覚統合により、システムのロバスト性が向上できることが示されている。

以上の研究から、聴覚処理に関して、モノラル音からバイノーラル音、そして視覚情報といったように、モダリティ数を増やし、情報統合を行うことは、ロバストな聴覚システムを構築するための重要な鍵であるといえる。

2.4 一般音理解

日常的な環境では、複数の音源からの音響信号が混在した混合音が聞こえてくることが一般的であり、そうした環境での動作が期待されるロボットもこうした一般的な音、つまり混合音を聞き分ける機能を有することが求められる。このような一般的な音の理解に関する研究は、ロボットの分野より、むしろ聴覚情景分析や、その工学的実現を目指した音環境理解の分野で行われてきた。実際に、単一音源ではなく混合音を対象として、音声だけではなく、音楽など非音声を含めた音響信号を理解するために様々な研究が行われてきた [Brown 92, Cooke 93, Nakatani 94, Rosenthal 98]。近年では、音声認識の分野でも、一般的な雑音下環境でのロバストな音声認識に関心が払われるようになり、音環境理解は注目を浴びている。

特に、混合音から、同一音源に由来するなど一貫した属性を備えた音響ストリームを抽出・分離する音源分離については、音環境理解における主要な研究テーマの一つとして、これまで様々な手法が試みられてきた。例えば、人間の聴覚モデルを仮定して蝸牛の処理をシミュレーションしているシステム [Brown 92, Slaney 94]、入力音に対し様々な聴覚マップを構築し、それらを統合することによって入力音から音声を分離するシステム [Brown 92]、倍音関係と音源方向を分離の手がかりとして利用し、インクリメンタルな処理が可能である音源分離システム (Bi-HBSS) [Nakatani 95] など様々な音響的特徴を用いた音源分離が行われている。しかし、音環境理解の研究のほとんどはシミュレーション環境で行われており、実時間処理、ノイズ対策や動的に変化する音響環境への対応など実環境への配慮が十分ではなかった。実時間処理という面から、青木らは、2 本のマイク間の強度差と位

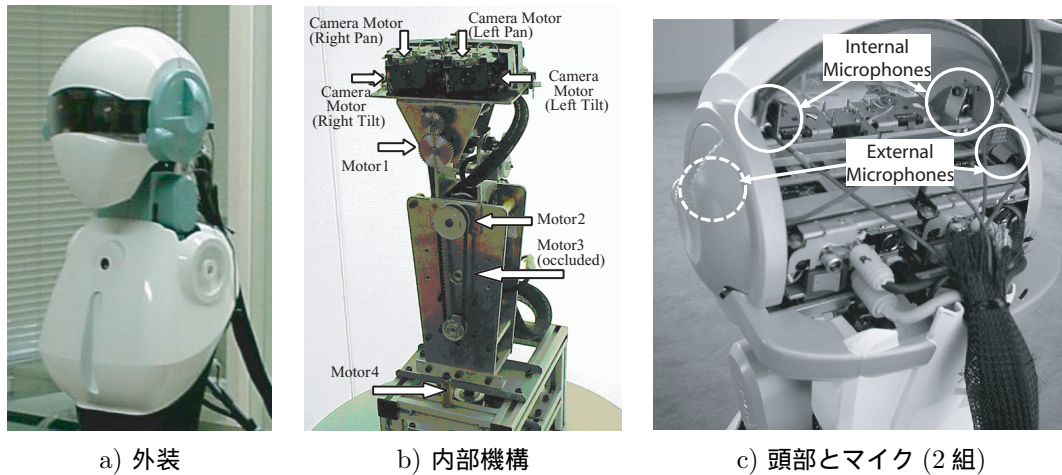


図 2 ヒューマノイド SIG

相差を利用し、18dB 以上の実時間音声強調を実現している [Aoki 01]。しかし、強度差については予め計測しておく必要があるため、未知環境への適応が難しく、マイクや音源も静止していることを前提としていることから、ロボットへの適用は難しい。近年では、マイクロホンアレイ [Saruwatari 99, Kim 96, Sturim 97] や独立成分分析 (ICA) [Okuno 99] を利用したアプローチも盛んに行われている。しかし、これらの手法は比較的システムが大きくなりがちであり、計算コストが高く、雑音を含む音源数が動的に変化したり、アレイや音源が移動する場合には、検出誤差が大きくなることなどから、ロボットへの搭載に適しているとは言えない。

3. ヒューマノイド SIG

研究のテストベッドとして、我々は図 2 に示すヒューマノイド SIG (以後、単に SIG) を開発した [Kitano 00]。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 2b 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。視覚センサには、左右一組のオートフォーカス機能付の CCD カメラ (Sony EVI-G20) を用いており、RS232C 経由でパン、チルト、ズーム制御および、方向情報などのカメラパラメータの取得が可能である。また、聴覚センサには図 2c に示すように、二組の無指向性のマイク (Sony ECM-77S) を用いている。一組のマイクは、耳の位置に設置されており、SIG 外部からの音を集め、内部からの音をできるだけ拾わないように内部とは隔離されている。また、マイク自体は無指向性であるが、外装の耳部の形状から、前方に指向性を有している。外装内部にあるもう一組のマイクは、SIG 内部音収音用であり、外部音収音用マイクと外装を挟んで対になるように設置されている。通常、内部機構は図 2a に示す外装に覆われている。外装は、その審美性や内部機構の保護だけではなく、音響的にも SIG 内部の音が

外部に漏れるのを防ぎ、音響信号処理の複雑性を緩和するという意味で重要である。

4. アクティブオーディションシステム

提案したアクティブオーディションを実現し、その有効性を示すため、図 3 に示すような実験的なシステムを構築した。このシステムは、2 章で挙げたロボット聴覚の 4 つの課題のうち、(1)–(3) の課題を解決すべく、動作時の音源定位を行うシステムである。

入力は、複数の純音とし、これを 4 本のマイクで、同期を取りながら、48 kHz, 16 bit サンプリングで収音する。各チャンネルに対し、窓長 1,024 ポイントの FFT による周波数解析を行った後、内外のマイクの強度差を利用して内部音抑制を行う。さらに、ピーク抽出、周波数値によるグルーピング、音源定位を行うことによって、各純音の定位を行う。以下にシステムの内部音抑制機能と音源定位機能を説明する。

4.1 SIG 内部音抑制システム

まず、内部音を抑制するために SIG の内部音の特性を計測した。その結果から、SIG の内部音は、次のような特性を有していることが判明した。

- カメラのモータ動作音は無視できるレベルであるが、カメラの定常ノイズは、左右それぞれ、3.7 dB である。
- ロボット駆動モータによる定常時、動作時のノイズは、それぞれ 5.6 dB, 23 dB である。
- ストッパー、ケーブルや外装の摩擦、カバーの結合部分の軋みなどの原因で動作時に、バースト的なノイズが発生し、広域に渡って大きな影響を与えている。

そこで、どのような方法が SIG の内部音抑制に効率的なのかを調べるため、独立成分分析 (ICA)、適応フィルタ、バーストノイズキャンセルフィルタという 3 種類の方法によるノイズキャンセルを実装し、比較検討した。

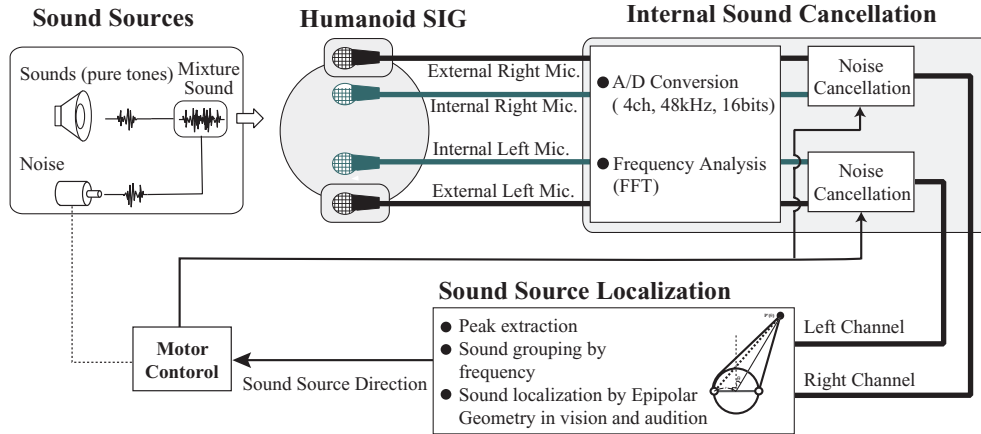


図3 アクティブオーディションシステム

ICA は、近年、盛んに音源分離を始め様々な分野に適用されている信号分離の手法である [Murata 98, Okuno 99]。ICA は統計的に独立な信号が混合した観測信号から、その混合割合や方向などの情報がない場合でも、元の独立な信号を推定することが可能である。しかし、SIG の場合では、マイク間距離が短い、マイク数 (4 本) に比べノイズ源数が多く、かつ、すべてのノイズ源が常に音を発しているわけではないなどの理由により、効果的なノイズ抑制を行うことは困難であった。さらに、分離の結果、得られた分離信号は、音源定位の際に重要な位相情報が歪んでしまい、SIG の内部音の抑制には適さなかった。

次に、アクティブノイズコントロール (ANC) で使用される適応フィルタを適用した。適応フィルタは、音源定位における IPD 算出誤差を極力抑えるため、線形位相フィルタを用いた。具体的には、次数 100 の FIR (Finite Impulse Response) フィルタを使用し、パラメータの更新は、最小二乗法の適応アルゴリズムを用いた。しかし、適応フィルタを使用しても効果的な内部音の抑制を行うことができなかった。これは、適応フィルタは、純粋にノイズを取得できる場合には有効に働くが、外装はロボット内部を完全に密閉しているわけではなく、ノイズのみを収音することが難しかったためと考えられる。

そこで、動作時に広帯域に渡って大きな影響を与えるバーストノイズをキャンセルするフィルタをヒューリスティクスに基づいて構築した。このフィルタはバーストノイズと見なされるサブバンドをキャンセルし、そのようなサブバンドは以後の処理には用いないものとした。サブバンド単位でのフィルタリングを用いたのは、ICA や適応フィルタのように、ノイズ抑制の計算誤差で生じる信号の位相情報の歪みを防ぐためである。特に、低周波域では、小さな歪みでも音源定位の結果に与える影響は大きくなる。バーストノイズを判定するための具体的な条件は以下の通りであり、各数値は実験的に求めた。

- (1) 内部音のパワーが外部音よりも強い
- (2) 20 以上の連続したサブバンドで一定値 (30 dB) 以

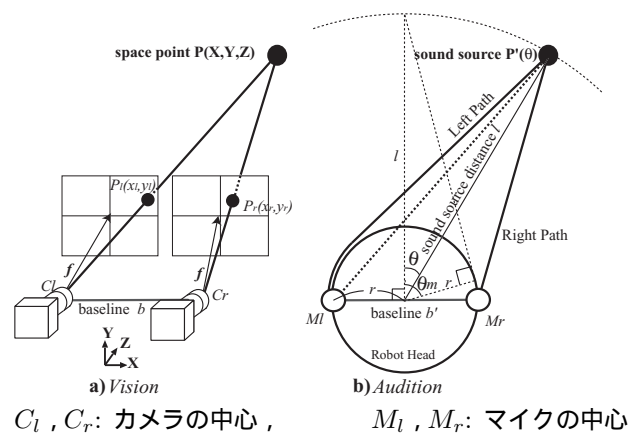


図4 エピポラ幾何による定位

上のパワーが観測される

(3) モータ駆動のコマンドが発行されている

この方法は、良好なノイズキャンセル効果を示したため、内部音抑制システムでは、バーストノイズキャンセルフィルタを採用している。

4.2 3 種類のセンサによる定位

視聴覚およびモータ情報による定位手法を説明する。聴覚処理による音源定位方法として、HRTF に依存しない手法である聴覚用のエピポラ幾何を用いる [Nakadai 00]。これは、ステレオビジョンの基本的な定位法であるエピポラ幾何を音源定位に適用したものであり、IPD から、水平方向の定位を行う手法である。

エピポラ幾何による視覚情報の定位

図 4a に示すような同一の焦点距離を持ち、光軸が並行でレンズ面が同一平面状にある 2 台のカメラを使った単純なステレオカメラを考える。空間上の点 $P(X, Y, Z)$ の左右のカメラに対するそれぞれの投影面上の座標を $P_l(x_l, y_l)$, $P_r(x_r, y_r)$ とすると、 P の座標は、

$$X = \frac{b(x_l + x_r)}{2(x_l - x_r)}, Y = \frac{b(y_l + y_r)}{2(x_l - x_r)}, Z = \frac{bf}{x_l - x_r}$$

として得られる [Faugeras 93]．ここで f はカメラレンズの焦点距離， b はベースラインを示している． SIG の左右のマッチングには，コーナ抽出アルゴリズムを利用する [Lourens 00]．この手法でコーナとエッジの集合を抽出した後， (P, C_l, C_r) によって構成されるエピポーラ面と投影面の交線であるエピポーラ線上に，左右の画像の対応点が存在するという性質を利用して，ステレオマッチングを行う． SIG では， $y_l = y_r$ という関係が保たれるため，マッチングは比較的容易である．

エピポーラ幾何を使用した音源定位

聴覚用のエピポーラ幾何では，収録したデータに対し周波数解析，ピーク抽出を行い，左右のピークの位相差情報を利用して定位を行う．ピーク抽出には，スペクトラルサブトラクション [Boll 79] に基づく高速で精度のよい手法 [Nakadai 99] を使用している．位相差情報による定位方法は以下の通りである．

まず，同時刻における左右のチャンネルのスペクトルをそれぞれ S_l, S_r とすると，IPD は，

$$\Delta\varphi = \arctan\left(\frac{\Im[S_r(f)]}{\Re[S_r(f)]}\right) - \arctan\left(\frac{\Im[S_l(f)]}{\Re[S_l(f)]}\right)$$

により得られる．ここで f はスペクトルにおけるピーク周波数， $\Re[S]$ ， $\Im[S]$ は，それぞれスペクトル S の実数部と虚数部を示す．この時，音源から左右のマイクへの距離差（両耳間距離差） Δd は，音速を v とすれば，

$$\Delta d = \frac{v}{2\pi f} \Delta\varphi \quad (1)$$

と表すことができる．

次に，図 4b に示すように SIG の頭部を球体とみなし，頭部形状による回折を考慮すれば，式 (2)-(4) に示すように Δd は， θ と l の関数 D として表すことができる．

$$D(\theta, l) = \begin{cases} r\left(\frac{\pi}{2} - \theta - \theta_m\right) + \delta(\theta + \pi, l) & (0 \leq \theta + \frac{\pi}{2} < \theta_m) \\ -2r\theta & (|\theta| \leq \frac{\pi}{2} - \theta_m) \\ -r\left(\frac{\pi}{2} + \theta - \theta_m\right) - \delta(\theta, l) & (0 \leq \frac{\pi}{2} - \theta < \theta_m) \end{cases} \quad (2)$$

$$\delta(\theta, l) = \sqrt{l^2 - r^2} - \sqrt{l^2 + r^2 - 2rl \sin \theta}, \quad (3)$$

$$\theta_m = \arccos \frac{r}{l}. \quad (4)$$

ここで，式 (2)-(4) で， θ, l を変化させた時の D の振舞いをシミュレーションした結果を図 5 に示す． θ が大きくなるほど l の影響は大きくなるが，50 cm 以上では l による影響は小さくなる．そこで， l を無限大と仮定すると D は θ のみの関数として以下のように定義される．

$$D(\theta) = \lim_{l \rightarrow \infty} D(\theta, l) = r(\theta + \sin \theta). \quad (5)$$

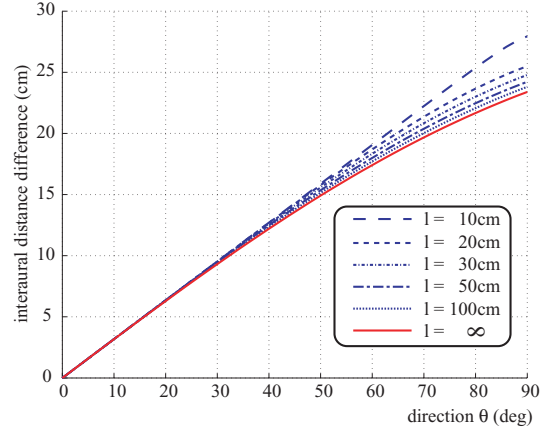


図 5 IPD and distance from sound source

近接学 (proxemics) の観点からも 50cm 以下は “intimate distance” と呼ばれ [Hall 66]，人間とのインタラクションなど実際の利用にあたり， l が無限大という仮定をおくことは妥当であろう．最終的に，音源方向 θ は，式 (1),(5) を用いて，以下の式により求めることができる．

$$\theta = D^{-1}(v/(2\pi f)\Delta\varphi). \quad (6)$$

音源を視覚用のエピポーラ幾何によって定位できれば，図 4 において， $P = P'$ となり，ビジョンによる定位結果 $\vec{P}$ を式 (7) を用いて，容易に音源方向 θ に変換できる．

$$\theta = \arcsin\left(\frac{\vec{P} \cdot \vec{M}_r}{|\vec{P}| |\vec{M}_r|}\right). \quad (7)$$

これにより，視覚情報が利用可能な場合，音源方向 θ の精度を高めることができる．また，音源方向を向くようなアクティブな動作により，音源定位の精度を高めることも可能となる．

サーボモータによる定位

SIG の体の向きはサーボモータのポテンシオメータから得ることができる．水平回転のモータを使用した場合 SIG の水平方向の精度の誤差は $\pm 1^\circ$ である． SIG の方向情報を使用して，精度よく聴覚，視覚の方向をワールド座標に変換することができる．

5. 動作中の音源定位実験

アクティブオーディションシステムにおいて， SIG 動作中に生じるノイズをキャンセルし，動きながらの定位が可能となることを示す．また，視覚，聴覚，ポテンシオメータから得られる SIG の方向情報から情報が欠如するような場合でも，これを補完して音源定位の曖昧性が解消できることを示す．さらに，動作によって音源定位が向上することを示すことによってアクティブオーディションの有効性を示す．

5.1 実験のシナリオ

音源には、500 Hz の純音を発生するスピーカ A (音源 A)、600 Hz の純音を発生するスピーカ B (音源 B) の 2 台のスピーカ (B&W Nautilus 805) を使用する。スピーカを設置した部屋は、約 10 平方メートルの一般的な住居用アパートの一室であり、残響時間が約 0.4 秒、暗騒音が 37.7 dB 程度の部屋である。

図 6 に示すように、初期状態では、SIG は 53° の方向を向いており、音源 A、B はそれぞれ SIG の 5°、69° 左に置かれている。また、SIG と各音源との距離は約 210 cm である。SIG の視野は水平方向で 45° であるため、初期状態では、69° 左に置かれている B は視野外となり見ることはできない。実験のシナリオは以下の通りである。

- (1) 最初に A が SIG の 5° 左で音を発する。
- (2) SIG は、抽出した画像と音の方向が同じであることから両者を同一音源に由来するものとして、アソシエーションする。
- (3) A が音を発した後、しばらくして B が音を発する。この時、B は SIG の視野外にあり、音源方向は聴覚情報によってのみ得られるため、SIG は B に関するアソシエーションを行なわない。
- (4) SIG は、視野外の音源 B の方向を向く。
- (5) 回転中、SIG の向きが 81° から 104° の間は、両方のスピーカが視野外となるため、SIG は聴覚情報のみで音源方向を推定する。
- (6) SIG が 104° まで回転すると、B が視野内に入るため、B を新たな物体として認識し、B に関して音と画像のアソシエーションを行なう。

評価には 4 種類のベンチマークを用いた。これらは、SIG の回転速度が速い場合 (秒速 68.8°)、遅い場合 (秒速 14.9°)、内部ノイズと同程度の弱い音響信号を入力した場合、および、内部ノイズを無視できる程度の強い音響信号を入力した場合の組み合わせである。

本稿では、500 Hz と 600 Hz の純音分離は、予め周波数情報を与え、周波数値に応じたグルーピングにより行っている。また、本実験では、視聴覚の定位はオンライン処理は行わず、予め定めた SIG の動作中に、音響データと画像データを同期を取って収録し、収録データに対してノイズキャンセルおよび視覚、聴覚による定位の処理を行った。

5.2 実験結果

弱い音響信号を入力した時のスペクトログラムを図 7 に示す。また、モータノイズをキャンセルしない場合の音源定位結果を図 8 に、キャンセルした場合の定位結果を図 9、10 に示す。図 8 - 10 では、X 軸が時間、Y 軸が音源方向であるが、音源方向は SIG 座標系でプロットされているので、図中の 0° は SIG のその時刻における

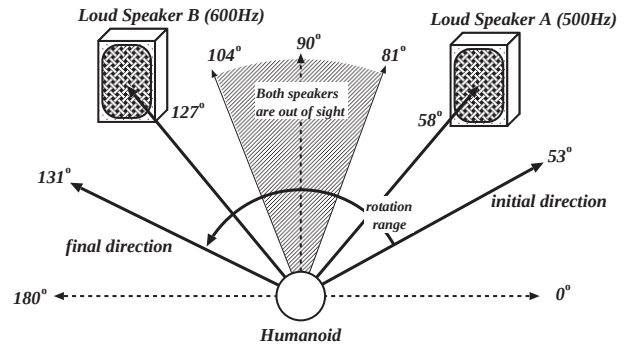
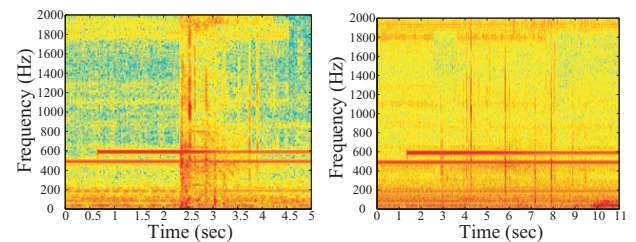
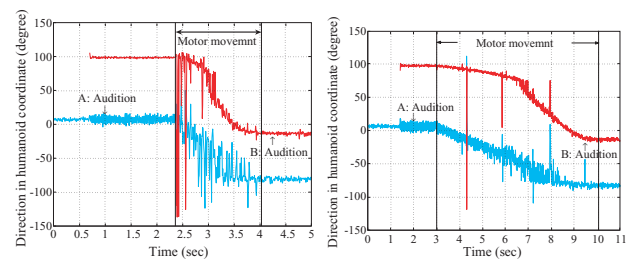


図 6 実験: SIG 動作中の視聴覚による定位



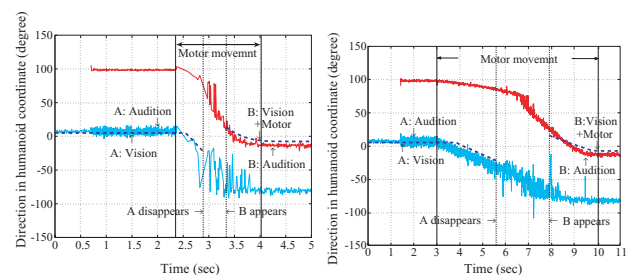
a) SIG の動作が速い場合 b) SIG の動作が遅い場合

図 7 入力音のスペクトログラム



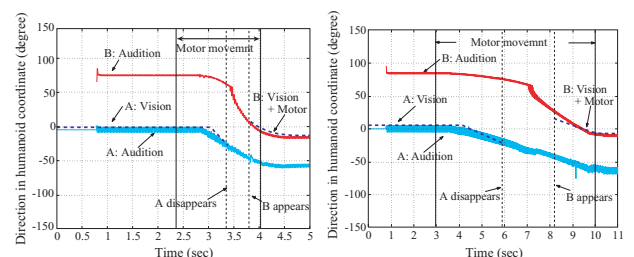
a) SIG の動作が速い場合 b) SIG の動作が遅い場合

図 8 内部ノイズ抑制なしの音源定位



a) SIG の動作が速い場合 b) SIG の動作が遅い場合

図 9 視聴覚による音源定位



a) SIG の動作が速い場合 b) SIG の動作が遅い場合

図 10 パワーの強い入力音の定位 (50dB)

正面方向、負の値は右方向、正の値は左方向を示す。図 9, 10 における点線は視覚による定位結果を示す。

実験の観測結果は以下の通りである。図 8a では、動作時の 2.5–3.5 秒付近で定位結果に深刻な影響が見られる。図 8b でも、4, 6, 8 秒付近で同様の現象が確認できる。両者とも、動作開始前の最初の 2 秒間の定位結果からは、入力音のパワーの強弱に関係なく音源定位は安定している。図 9, 10 では、動作時でもノイズキャンセルにより、バーストノイズが取り除かれ正確な音源定位が達成されている。

これらの定位結果より、まず、ノイズが少ない静止時には、聴覚用エピソード幾何は効果的であるが、動作時には、定位を不可能にするようなパワーの大きいモータノイズが発生することがわかる。実際にスペクトログラムでは、動作時に広帯域のモータノイズが観測されている。

提案したノイズキャンセルは、動作が遅い場合には、バーストノイズをきれいに取り除いている。動作が速い場合でも聴覚用エピソード幾何で音源定位を行うことができる程度に、ノイズの影響を軽減している。また、入力信号が強い場合は、より正確な定位を達成している。

視覚による定位結果は、聴覚による定位の正確さを指示している。音源 B が視野内に入った際に B に関するアソシエーションが行われることにより、視覚情報とモータ情報からの補完が可能になり、定位精度の向上を図ることが可能になる。

次に、音源がロボットの正面方向に近づくに従い、定位誤差が小さくなることがわかる。実際、動作前の状態で正面方向に近い音源 A の定位誤差はほぼ 0° であるのに対し、遠い音源 B の定位誤差は約 30° である。音源 B の方向を向くに従い、定位誤差は小さくなり、動作終了時には、ほぼ 0° となる。これは、音源方向を向くという動作が定位精度を向上させることを示しており、アクティブオーディションの有効性を指示している。一般に、パワーの小さい音や未知環境での定位など利用できる情報や信頼できる情報が少ない場合には、定位誤差が大きくなる。このような場合には、アクティブオーディションの効果はより大きくなる。特に純音のように他の倍音成分からの情報を得られない場合は、部屋の音響環境の影響を受けやすいので、アクティブな動作は定位向上に不可欠といえる^{*2}。

6. 考察と課題

構築したアクティブオーディションシステムは、動作時のノイズをキャンセルし、残響時間 0.4 秒、暗騒音 37.7dB 程度の反響のある部屋での音源定位が可能であることを示した。より反響の強い環境では、精度が悪くなることが予想されるが、そのような場合の音源定位に

ついては、今後の課題である。実環境では、道路を走る自動車や実験室外からの人の声、ロボット内部の音など複数の音源が存在しており、入力音は混合音であるという音環境理解の仮定は本質的である。多くのロボットは単一音源を仮定した聴覚処理を行っているが、*SIG* では、同時に 2 つの純音が存在する場合における定位を達成した。調波構造を有する音や互いの周波数成分が重なり合う場合については、本稿では扱わず、別途報告する。

エピソード幾何は、視聴覚の処理を統合し、音源を定位する有効な手法である。これまで、聴覚を備えたロボットは HRTF を明示的に、または暗黙のうちに使用しているが、HRTF による音源定位は、特定の環境下でのみ有効であり、動いている音源を追跡することも難しい。これに対し、*SIG* では、HRTF を用いず、未知環境下での音源定位を可能にした。さらに、定位精度は音源方向を向くというアクティブな動作によって向上することを示し、アクティブオーディションの有効性を示した。ただし、聴覚用のエピソード幾何では IPD のみを定位情報として用いているため、波長の短い高周波数域 (*SIG* では 1500 Hz 以上) では、定位に曖昧性が生じ、定位精度が低下してしまう。このような状況に対処するために、IID を用いることは効果的である。HRTF を利用せずに、IPD と IID の両方から、ロバストな音源定位を行う方法については、別途報告する。また、本稿で提案したエピソード幾何は、音響ストリーム分離に有効であるという結果も得られているが、音響ストリーム分離にエピソード幾何を用いる方法についても別途報告する。

アクティブな動作は、知覚をロバストにし、インタラクションを豊かにする有効な手段である反面、モータノイズが聴覚処理を困難にするという側面を持っている。例えば、Cog プロジェクトは、ソーシャルインタラクションに焦点を当てているが、聴覚処理への影響は言及されていない [Brooks 98]。*SIG* では、ロボットが音響的な身体性を獲得できるように外装の内外にマイクを設置することで、内部音の認識・抑制を可能とした。そして、外装をノイズキャンセルに利用することはアクティブな動作を有効利用するために重要であることを示した。ただし、外装内部での共鳴は無視できないため、共鳴が少ないデザインを考慮する必要があるだろう。

7. まとめ

本稿では、ロボット聴覚を実現するために、アクティブオーディションを提案し、ヒューマノイド *SIG* を対象に実験システムを構築・評価した。構築システムは、エピソード幾何と外装の音響効果を利用することにより、内部音抑制、複数純音の音源定位を実現した。エピソード幾何は、HRTF を使用しない未知環境での音源定位、および視聴覚の統合を可能とし、*SIG* の外装は、動作時のノイズキャンセルを実現し、ロボット聴覚における “stop-

*2 純音定位は一見簡単に思われるが、Hornbostel によれば、一般人では純音定位は調波構造を伴う音よりも定位が難しい。

perceive-act” という制約を緩和する上で大きな役割を果たすことを示した。また、実環境における音源定位実験を通じて、音源方向を向くというアクティブな動作は、聴覚による音源定位の精度向上、視覚による定位情報の利用を可能にするという点でロバストな知覚を実現するために有効であることを示し、アクティブオーディションの有効性を示すことができた。

今後、聴覚処理では重複周波数成分や調波構造音への対処、アプリケーションの面からは視聴覚の実時間処理、ロバスト性向上の面からは、他のモダリティの追加、階層的な知覚処理の導入が課題である。

謝 辞

北野共生システムプロジェクトのメンバに感謝する。また大阪大学石黒浩教授の視聴覚統合に関する助言と議論に感謝する。

◇ 参 考 文 献 ◇

- [Aarabi 01] Aarabi, P. and Zaky, S.: Robust sound localization using multi-source audiovisual information fusion, *Information Fusion*, Vol. 2, No. 3, pp. 209–223 (2001)
- [Aloimonos 87] Aloimonos, Y., Weiss, I., and Bandyopadhyay, A.: Active Vision, *International Journal of Computer Vision*, Vol. 1, No. 4, pp. 333–356 (1987)
- [Aloimonos 93] Aloimonos, Y.: *Active Perception*, Lawrence Erlbaum Associates, New Jersey (1993)
- [Ando 95] Ando, S.: An autonomous three-dimensional vision sensor with ears, *IEICE Transactions on Information and Systems*, Vol. E78–D, No. 12, pp. 1621–1629 (1995)
- [Aoki 01] Aoki, M., Okamoto, M., Aoki, S., Matusi, H., Sakurai, T., and Kaneda, Y.: Sound source segregation based on estimation of incident angle of each frequency component of input signals acquired by multiple microphones, *Acoust. Sci. and Tech.*, Vol. 22, No. 2, pp. 149–157 (2001)
- [浅井 00] 浅井 泰行, 中島 弘道, 山村 毅, 黄 捷, 大西 昇: 運動を介した視聴覚による物体定位能力の獲得, *IEICE Trans. D-II*, Vol. j83-D-II, No. 7, pp. 1676–1684 (2000)
- [Asoh 97] Asoh, H., Hayamizu, S., Hara, I., Motomura, Y., Akaho, S., and Matsui, T.: Socially Embedded Learning of the Office-Conversant Mobile Robot *Jijo-2*, in *Proceedings of 15th International Joint Conference on Artificial Intelligence (IJCAI-97)*, Vol. 1, pp. 880–885, AAAI (1997)
- [Bajcsy 88] Bajcsy, R.: Active perception, *Proceedings of the IEEE, Special issue on Computer Vision*, Vol. 76, No. 8, pp. 996–1006 (1988)
- [Boll 79] Boll, S. F.: A Spectral Subtraction Algorithm for Suppression of Acoustic Noise in Speech, in *Proceedings of 1979 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-79)*, pp. 200–203, IEEE (1979)
- [Breazeal 99] Breazeal, C. and Scassellati, B.: A context-dependent attention system for a social robot, in *Proc. of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99)*, pp. 1146–1151 (1999)
- [Brooks 98] Brooks, R., Breazeal, C., Marjanovic, M., Scassellati, B., and Williamson, M.: The Cog Project: Building a Humanoid Robot, in *C. Nehaniv, ed., Computation for Metaphors, Analogy and Agents*, Vol. 1562 of *Springer Lecture Notes in Artificial Intelligence*, pp. 52–87, Springer-Verlag (1998)
- [Brown 92] Brown, G. J.: *Computational auditory scene analysis: A representational approach*, University of Sheffield (1992)
- [Cooke 93] Cooke, M. P., Brown, G. J., Crawford, M., and Green, P.: Computational Auditory Scene Analysis: Listening to Several Things at once, *Endeavour*, Vol. 17, No. 4, pp. 186–190 (1993)
- [Faugeras 93] Faugeras, O. D.: *Three Dimensional Computer Vision: A Geometric Viewpoint*, The MIT Press, MA. (1993)
- [Hall 66] Hall, E. T.: *The Hidden Dimension*, Anchor books doubleday (1966)
- [Hershey 00] Hershey, J., Ishiguro, H., and Movellan, J. R.: Audio Vision: Using Audio-Visual Synchrony to Locate Sounds, in *Neural Information Processing Systems*, Vol. 12, pp. 813–819, MIT Press (2000)
- [Huang 97] Huang, J., Ohnishi, N., and Sugie, N.: Building Ears for Robots: Sound Localization and Separation, *Artificial Life and Robotics*, Vol. 1, No. 4, pp. 157–163 (1997)
- [池田 01] 池田 徹志, 石黒 浩, 浅田 稔: 相互情報量最大化に基づく異種センサの信号の適応的統合, 第 19 回ロボット学会学術講演会, 第 2F23 巻, pp. 691–692 (2001)
- [Kim 96] Kim, H. Y., Asano, F., Suzuki, Y., and Sone, T.: Speech Enhancement Based on Short Time Spectral Amplitude Estimation with Two-Channel Beamformer, *IEICE Trans. fundamentals*, Vol. E79–A, No. 12, pp. 2151–2158 (1996)
- [Kitano 00] Kitano, H., Okuno, H. G., Nakadai, K., Fermin, I., Sabish, T., Nakagawa, Y., and Matsui, T.: Designing a Humanoid Head for RoboCup Challenge, in *Proceedings of Agent 2000 (Agent 2000)* (2000)
- [久野 95] 久野 義徳: アクティブビジョン—歴史と展望—, *人工知能学会誌*, Vol. 10, No. 4, pp. 493–499 (1995)
- [Kurata 95] Kurata, T., Chang, D., and Hashimoto, S.: Multimedia sensing system for robot, in *Proceedings of 4th IEEE International Workshop on Robot and Human Communication (RO-MAN '95)*, pp. 83–88, IEEE (1995)
- [Lourens 00] Lourens, T., Nakadai, K., Okuno, H. G., and Kitano, H.: Selective Attention by Integration of Vision and Audition, in *Proceedings of Humanoids 2000*, IEEE/RSJ (2000)
- [Makous 90] Makous, J. C. and Middlebrooks, J. C.: Two-dimensional sound localization by human listeners, *Journal of the Acoustical Society of America*, Vol. 87, No. 5, pp. 2188–2200 (1990)
- [Matsusaka 99] Matsusaka, Y., Tojo, T., Kuota, S., Furukawa, K., Tamiya, D., Hayata, K., Nakano, Y., and Kobayashi, T.: Multi-person Conversation via Multi-modal Interface — A Robot who Communicates with Multi-user, in *Proceedings of Eurospeech*, pp. 1723–1726, ESCA (1999)
- [Murata 98] Murata, N. and Ikeda, S.: 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)
- [Nakadai 99] Nakadai, K., Okuno, H. G., and Kitano, H.: A Method of Peak Extraction and Its Evaluation for Humanoid, in *SIG-Challenge-99-7*, pp. 53–60, JSAI (1999)
- [Nakadai 00] Nakadai, K., Lourens, T., Okuno, H. G., and Kitano, H.: Active Audition for Humanoid, in *Proceedings of 17th National Conference on Artificial Intelligence (AAAI-2000)*, pp. 832–839, AAAI (2000)
- [Nakagawa 99] Nakagawa, Y., Okuno, H. G., and Kitano, H.: Using Vision to Improve Sound Source Separation, in *Proceedings of 16th National Conference on Artificial Intelligence (AAAI-99)*, pp. 768–775, AAAI (1999)
- [Nakatani 94] Nakatani, T., Okuno, H. G., and Kawabata, T.: Auditory Stream Segregation in Auditory Scene Analysis with a Multi-Agent System, in *Proceedings of 12th National Conference on Artificial Intelligence (AAAI-94)*, pp. 100–107, AAAI (1994)
- [Nakatani 95] Nakatani, T., Okuno, H. G., and Kawabata, T.: Residue-driven architecture for Computational Auditory Scene Analysis, in *Proceedings of 14th International Joint Conference on Artificial Intelligence (IJCAI-*

- 95), Vol. 1, pp. 165–172, AAAI (1995)
- [Okuno 99] Okuno, H., Ikeda, S., and Nakatani, T.: Combining Independent Component Analysis and Sound Stream Segregation, in *Proceedings of IJCAI-99 Workshop on Computational Auditory Scene Analysis (CASA'99)*, pp. 92–98, IJCAI (1999)
- [Perrett 97] Perrett, S. and Noble, W.: The effect of head rotations on vertical plane sound localization, *Journal of the Acoustical Society of America*, Vol. 102, pp. 2325–2332 (1997)
- [Rosenthal 98] Rosenthal, D. and Okuno, H. G. eds.: *Computational Auditory Scene Analysis*, Lawrence Erlbaum Associates, Mahwah, New Jersey (1998)
- [Saruwatari 99] Saruwatari, H., Kajita, S., Takeda, K., and Itakura, F.: Speech Enhancement Using Nonlinear Microphone Array Based on Complementary Beamforming, *IE-ICE Trans. fundamentals*, Vol. E82-A, No. 8, pp. 1501–1510 (1999)
- [Slaney 94] Slaney, M., Naar, D., and Lyon, R. F.: Auditory Model Inversion For Sound Separation, in *Proceedings of 1994 International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 77–80 (1994)
- [Sturim 97] Sturim, D. E., Brandstein, M. S., and Silverman, H. F.: Tracking Multiple Talkers Using Microphone-Array Measurements, in *Proceedings of 1997 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-97)*, pp. 371–374, IEEE (1997)
- [高西 94] 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, pp. 793–796 (1994)
- [植松 01] 植松 尚, 加藤 正晴, 柏野 牧夫, 平原 達也: 頭外音像定位における自発的な頭部回転の影響, 日本音響学会講演論文集, pp. 501–502 (2001)

〔担当委員: 岡 隆一〕

2002 年 7 月 22 日 受理

著 者 紹 介



中臺 一博(正会員)

1970 年生。1993 年東京大学工学部電気工学科卒, 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。主にロボット聴覚, 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受賞, IROS 2001 BEST Paper Nomination Finalist 受賞, 2002 年電気通信普及財団賞奨励賞受賞など。日本人工知能学会, 日本ロボット学会, 日本音響学会会員。



奥乃 博(正会員)

1972 年東京大学教養学部基礎科学科卒業。NTT, 科学技術振興事業団, 東京理科大学理工学部情報科学科を経て, 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。1990 年度論文賞, IEA/AIE-2000 最優秀論文賞, IROS-2001 論文賞佳作。音環境理解, ロボット聴覚の研究に従事。著編書: 『インターネット活用術』(岩波書店, 1996), 『Computational Auditory Scene Analysis』(LEA, 1998) 等。情報処理学会, 日本ロボット学会, ACM, AAAI 等会員。



北野 宏明

1984 年国際基督教大学教養学部理学科(物理学専攻)卒業後, 日本電気(株)に入社。1988 年より米カーネギー・メロン大客員研究員。1991 年京都大学博士号(工学)取得。1993 年ソニーコンピュータサイエンス研究所入社, 2002 年同取締役副所長。1998 年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総括責任者兼務。ロボカップ国際委員会ファウンディング・プレジデント。国際ロボットデザイン委員会(IRoDA)理事。1993 年 Computers and Thought Award 受賞。