博士論文

アクティブオーディションによる自然な ヒューマン・ロボットインタフェース の実現に関する研究

中臺 一博

平成 15 年 3 月 7 日 提出

内容梗概

近年, ヒューマノイドを中心にロボットが注目されている. この延長線上には, 将来, ロボットが人間社会で人と共生し, ソーシャルインタラクションを通じて, パートナーとしての役割を果たすことへの期待があろう. しかし, 現状のロボットは, そうした人間とのソーシャルインタラクションを可能にするほどのロバストな知覚処理は難しい. 特に, 人間とのソーシャルインタラクションで最も重要である聴覚機能にフォーカスした研究は, これまで, ロボットを対象としては, あまり行われていない. また, 実環境・実時間でのロボット聴覚を実現するために, いくつかの問題点は指摘されてきたものの, これらの課題を体系的にまとめた報告はなかった. そこで, 本研究では, ロボット聴覚機能の課題を体系的に整理し, これらを解決するための具体的な方法について議論を行う. そして, アクティブな動作を様々なセンサ情報と統合することにより, 知覚を向上できるアクティブパーセプションはロボット聴覚の向上にも本質的であると捕らえ, これをロボット聴覚に適用したアクティブオーディションを提案する. また, 複数の聴覚情報の統合, 聴覚情報以外の感覚情報との統合を行うことによる知覚向上, および, より一般的な処理を目指したロボットによる一般的な音 (混合音) の理解についてもあわせて議論する.

実際に構築したロボット聴覚システムは, ロボットに特有な動作時のノイズをキャンセルすることで, アクティブな動作の利用を可能とした. また, アクティブな動作を効果的に用いることにより, システムが有している視聴覚を統合した話者の定位・追跡, 注意を向けた方向の音源を実時間で抽出するアクティブ方向通過型フィルタによる音源分離, 分離音の音声認識といった機能を向上した. システムは, 聴覚センサとして, ロボット外装内部と外装の耳位置に取り付けた 2 組のマイクロホン, 視覚センサとしてステレオカメラを備えた 4 自由度の上半身ヒューマノイドロボット *SIG* 上に構築されている. 外装内部のマイクロホンは動作時のノイズキャンセル用に使用し, 音源の分離・定位・認識といった聴覚機能は, 左右の耳の位置に取り付けた 2 本のマイクロホンにより実現されている.

システムの各機能の個別評価やシステム全体を通した統合評価を通じて, アクティブオーディション, 感覚情報の統合, 一般音理解の有効性・ロバスト性を評価し, ヒューマン・ロボットインタフェースとしての有効性を示す.

目次

表目次

図目次

第1章

序論

1.1 論文の目的

本論文では, 将来, ロボットが人間社会に入り込み, 人間と共生する上で重要なヒューマン・ロボットインタフェースである聴覚について論じる. 特に, 人間が知覚向上のために行うアクティブな動作, 他の感覚情報との統合に着目し, ロボットの音源定位・分離・認識を向上させるモデルを提案する. また, その工学的な実現やヒューマン・ロボットインタラクションへの応用を通して, 有効性を明らかにする.

1.2 本論文の背景

20 世紀末から, 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. それまでロボットといえば, 古くから産業用途で NC (Numerical Control) 機械と共にファクトリーオートメーション (FA) 化を図るために使用されてきたものが一般的であった. 研究用途でも, 早稲田大学で 30 年間以上にわたって開発されてきた二足歩行ロボット Wabian や人間と協調作業を行う Hadaly といったヒューマノイド, あるいは, 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, Sony AIBO, ホンダの ASIMO や Sony SDR-4X といったメディアを大きく騒がしているロボットを筆頭に様々なロボットが登場している. 研究分野では, 人間と感情豊かなインタラクションを行う MIT AI Lab. の Kismet, 科学技術振興事業団 ERATO 川人プロジェクトで開発された脳神経科学の成果を生かした複雑な運動制御が可能な油圧式制御式ロボット DB, 日常活動型のロボットとして開発が進められている ATR の Robovie, 経済産業省のヒューマノイドロボットプロジェクト (HRP) のヒューマノイドなどが相次いで登場し, 一部のロボットは, 将来的にはビジネスを目指すなど, 従

来では見られなかった活発な動きが見られる.

一方, エンターテインメント分野に目を向ければ, AIBO だけではなく, 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 NEC Papero, キャラクタービジネスとして成功を収めている科学技術振興事業団 ERATO 北野共生システムプロジェクトの PINO, 猫型のネコロ, 熊型の Latte, カンガルー型などといった本格的なロボットからおもちゃと区別することが難しいほど簡単なロボットまで多種多様なロボットが次々と登場してきている. この百花繚乱ともいうべき急激な多様化をカンブリア紀に生物の飛躍的な多様化が起きた現象を指すカンブリア爆発になぞらえて, ロボットのカンブリア爆発と呼ぶ人もいる [111] ほどである. このことは, 図 1.1 に示されるように, AIBO が発売された 1999 年以降, 急激に朝日・毎日・読売・産経紙といった大衆紙でロボットが取り上げられるようになったことからも明らかである.

また, 図 1.1 からは, このような傾向が, 同じくロボットが頻繁に取り上げられた 1980 年代とは違った意味合いをもっていることを読み取ることができる. 1980 年代は, 日本の製造業の競争力を示す産業用ロボットが日経など産業紙で取り上げられていたが, 大衆紙はあまり関心を示さなかったことがわかる. そして, 産業紙においても産業用ロボットが珍しい存在でなくなるにつれ掲載回数は減少の一途を辿っていた.

これは, 近年登場しているロボットの多くは, 従来の研究や産業用途に特化することを目的としているわけではないことを示唆している. つまり, 大衆紙の関心をひく, 個人向けの新たなロボット産業の市場を開拓することを狙って開発されているといえよう.

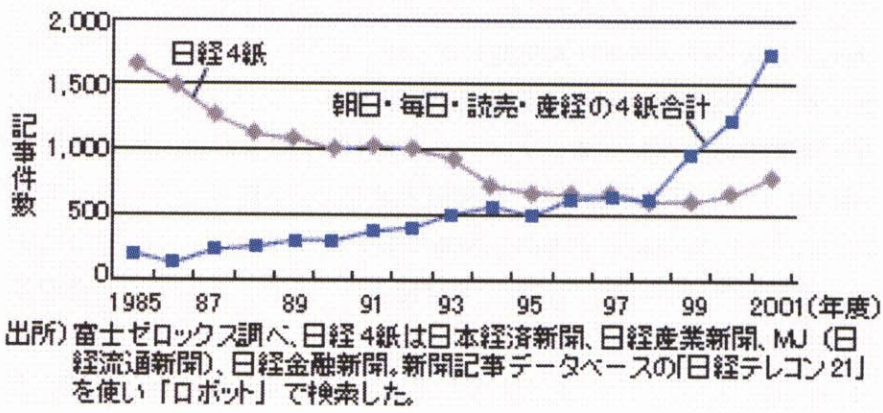


図 1.1: 新聞における「ロボット」という単語の出現頻度

こういった現象の背景の一つには, 社会的なロボットへのニーズの高まりが挙げられる. 例えば, 先進諸国の世界的な高齢化に対応して, 病院での看護, 独居老人の話し相手など実際に福祉や介護の現場で使用されているロボットも存在するし, 複雑さを増す社会を効率的に運営する上で, ロボットを使用することも有効であろう. また, 別の背景として, 映画

や SF といったメディアの影響もあり, ロボットが人間の社会に入ることに対して, 世間の人々の抵抗感が薄れてきていることも挙げられるだろう*1

昨今, 盛り上がりを見せているロボットへの社会的な関心を保ち, 新たなロボット産業の萌芽を促進するためには, 人間の社会に入って, 我々人間のパートナーとして人間と共生していくことができるロボットが必要である. 単に, 現状のロボットのように, 歩いたり, 可愛らしい仕草をするだけであれば, 高価な動くおもちゃ以上のものにはなりえず, 人々の興味を失い, 一過性のものとなってしまう可能性もある.

しかし, このようなロボットを実現するにあたり, 現状のロボットの知覚機能はあまりにも低いといわざるをえない. 特に, 聴覚機能は, 我々人間がいとも簡単に音を聞き分けてしまうため, 簡単に実現できる問題であると思われがちである.

例えば, 1950 年に出版された, Issac Asimov 著のサイエンスフィクション "I, Robot" では, 最初に登場したロボット「ロビー」は, 無声ロボットであり, しゃべることができないが, 人の言うことを聴いて子守をすることができるのである [14]. つまり, 当時は, 音声認識の実現は, 音声合成よりもやさしいと思われていたことを示している. 現代においてはどうであろうか. Asimov から 50 年以上たった 2002 年に出版された瀬名秀明著のサイエンスフィクション "明日のロボット" では, 捨てられていた旧型のロボットは, 「たどたどしい音声合成とすばやい音声認識」を行うと書かれているのである [111]. 実は, あまり変わっていないのである.

このような世間の認識とは裏腹に, 90%以上の認識率を持っているという触れ込みで販売されている音声認識システム一つをとってみても, システムの仮定している条件が少しでも満たされないと, 簡単に十数% 以下にまで認識率が落ちてしまうなど, 実環境でのロバスト性には大きな問題を抱えている.

そこで, 本研究では, ロボットが人間社会で, 人間と共生するための知覚機能に着目し, 特に, 人間との知的なインタラクションを行う上で, 最も基本的で重要な機能である聴覚に焦点を当て, ロボット聴覚を実現するための課題・手法について検討する.

1.3 論文の意義

本論文の意義は, 1 点目として, 従来行われていなかったアクティブな動作を伴った聴覚処理 "アクティブオーディション" を提唱し, ロボット聴覚を新たな研究分野として確立したことにある. 聴覚情景分析やその工学的な実現を目指す音環境理解では, 長年に渡り, 混合音の分離を中心に音源定位や認識に及ぶ広範囲な聴覚処理の研究が行われてきたが,

*1 実際, ホンダが P3 の製作に際し, ローマ法王に認めてもらうため, バチカンに許可を取りに行ったというエピソードがあるほど, 特にキリスト教圏では, 人が人に近い物を創造する事への反発があった.

音源やマイクの移動を前提とした研究は, ほとんど行われていない. 人間では, 二つの耳を使って, 自分や音源が移動する場合でも, 音源定位・分離・認識といった能力や, カクテルパーティ効果として知られるような選択的に注意を向ける能力は一般的であるが, 従来の研究では, このような視点が欠けていた. そこで, 本論文では, ロボット聴覚をロボティクス, AI, 信号処理を複合的に扱う新しい研究テーマとして定義し, その課題を明確にした.

2点目は, 応用的な観点として, 実際にシステムを実装して, より自然なヒューマン・ロボットインタフェースを実現したことにある. 従来の聴覚機能を備えたロボットの研究では, そのほとんどが, 混合音を扱っていなかったり, 話者の口元にマイクロホンをおいたり, 自分が発生するノイズで音声認識がうまくいかなかったりという問題を抱えたものであり, 実環境で聴覚によるロバストなヒューマン・ロボットインタフェースを実現する研究はあまり行われていなかった. 人間のようにアクティブな動作を利用した聴覚機能はロボットでも本質的であるにもかかわらず, これまでそのようなシステムは実装されてこなかった. 本論文では自然なインタラクションをロボットに備わったマイクロホンにより, 複数の人物 (音源) が同時に存在する場合や音源やマイクロホンの位置が動的に変化する場合においても積極的な動作を利用したフレンドリなインタラクションができることと定義し, ロボット聴覚システムの工学的な実現を通じて, その有効性を明らかにしている.

3点目として, アクティブな動作を伴った聴覚処理のモデル化とその定量的な評価という点が挙げられる. 聴覚心理の分野では, 動作による聴覚の向上が指摘されているが, これまで定量的な測定が難しかった. 本論文では, あえて2つのマイクを備えたヒューマノイドロボットによる音源定位・分離・認識のモデル化を行い, アクティブな動作の有効性を定量的に示している. また, 動作だけではなく, 視覚など他のセンサ情報との統合による知覚向上についても定量的な議論を行っている.

本論文におけるアクティブオーディションを利用した知覚の向上は, 動作が可能であるもの全般に適用できるため, ロボットを越えた様々な分野での応用が可能である. 本論文で示した考え方や手法は ロボットにとどまらず, 様々なヒューマンマシンインタフェースを高度化する要素技術としても発展し得るものである.

1.4 論文構成

本論文は全11章からなる.

第2章は, ロボット聴覚の課題と現状について述べる. 特に, ヒューマン・ロボットインタフェースとしてロボット聴覚に注目した場合に不可欠な機能として, 「アクティブオーディション」, 「自己が発生する音の認識と抑制」, 「視覚情報など他のセンサ情報との統合による曖昧性の解消」, 「一般的な音の理解」について論じる. また, 人間とのソーシャルインタラクションを念頭に置き, ロボット聴覚を応用した音声認識やロボットのパーソ

ナリティの必要性を議論する.

第 3 章は, アクティブオーディションに基づくロボット聴覚モデルを提案し, 本研究で用いたヒューマノイドロボット *SIG* と共に, 説明する.

第 4 章は, アクティブオーディションを実現するため, 自己が発生する音の認識と抑制について述べる. 外装の音響効果を利用し, 動作時のモータノイズを効率的にキャンセルする方法を述べ, 一般的なノイズキャンセル方法と比較する. また, ノイズキャンセル後の音源定位についてもあわせて述べる. 音源の定位では, 計測が不要で計算的に音源方向を推定できる聴覚エピポーラ幾何を提案し, 動作時の純音の定位を通じて, その有効性を述べる.

第 5 章では, 音源定位にフォーカスして, 実環境下での音源定位・追跡を論じる. 第 4 章で論じた純音の定位を拡張して, より一般的な音である調波構造を有する音の定位を行う. また, 同時に複数の音響的な手がかりを統合することによるロバスト性の向上についても述べる.

第 6 章では, ストリームベースのシンボリックな視聴覚統合について述べ, これに基づいて実装された実時間複数人物追跡システムを説明する. 特に, 分散処理を用いた実時間処理, カルマンフィルタを用いたストリームの形成, および階層的なアソシエーションにより, 曖昧性を解消し, ロバスト性を高めるメカニズムを述べる.

第 7 章では, 人間と同様に 2 つのマイクロホンを用いて, 実時間で音源を分離することができるアクティブ方向通過型フィルタについて述べる. この際に, 2 つのマイクロホンを用いた場合, 聴覚における中心窩ともいうべき現象が見られることを示し, この現象を積極的に利用することにより, 実環境で音源分離を向上させ, ロボットにおける聴覚情景分析の実現を目指す.

第 8 章では, 複数の音響モデルを利用した分離音の音声認識について述べる. 一般に従来の音声認識は分離処理やノイズの混入で歪んでしまった音声の認識が難しく, これを克服するために複数の音響モデルを利用した. 複数の音声認識結果や顔認識によって得られた話者情報を確率ベースの手法で統合することにより, 同時発話でもロバストな音声認識を実現した.

第 9 章では, 構築したロボット聴覚システムに対し, 自然なヒューマンロボットインタラクションを実現するためパーソナリティを導入する. 様々な状況における人間とのインタラクションを紹介することにより, その有効性を示す.

第 10 章では, 考察を述べ, 第 11 章で結ぶ.

また, 各章間の対応をより明確にするため, 各章と本研究で述べる技術の対応関係, および技術間の関係を示すチャートを図 1.2 に示す.

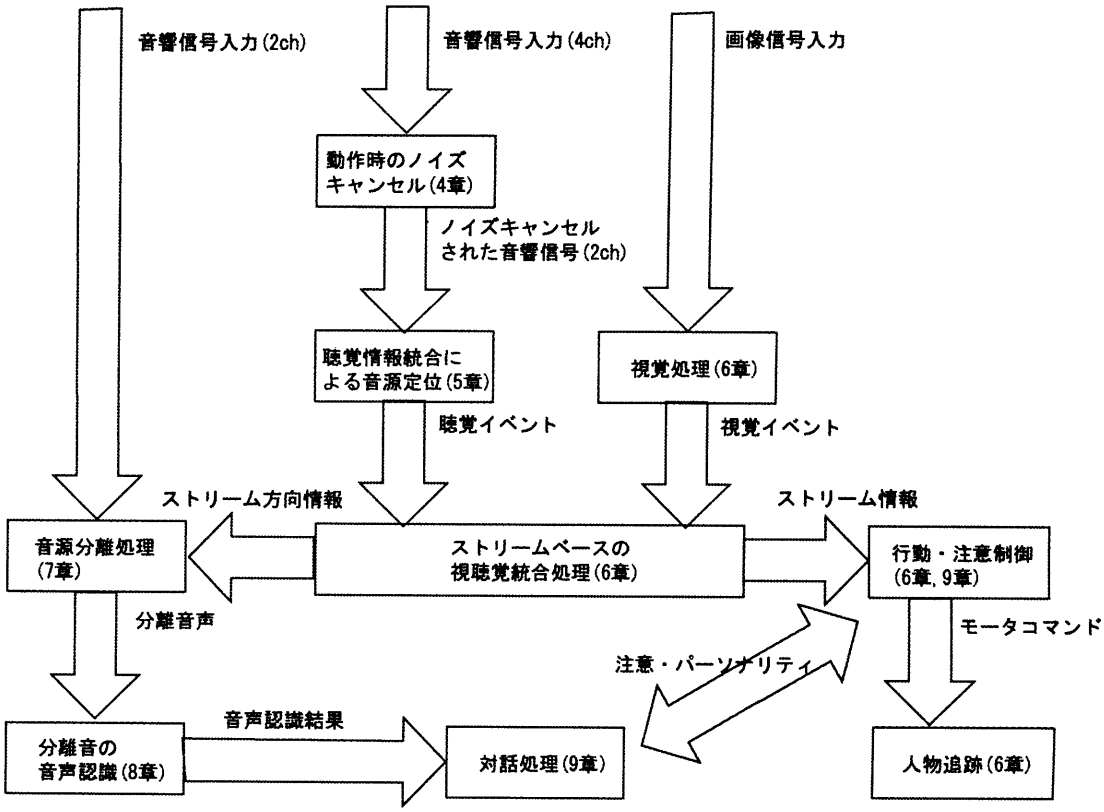


図 1.2: 本研究における技術間チャート

1.5 関連発表文献

発表論文と, 各章の対応関係を示す.

第 2 章 ロボット聴覚の課題と現状

[1]. 中臺 一博, 奥乃 博, 北野 宏明: 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. 人工知能学会論文誌 Vol. 18, No. 2-F, pp.104-113, 2003

[2]. 奥乃 博, 中臺 一博, 北野 宏明: ロボットの耳は2つで十分か. 日本音響学会誌 小特集 – なぜ耳は二つあるか? –, vol. 58, no.3, pp.205–210, 日本音響学会, 2002.

第 3 章 アクティブオーディションに基づくロボット聴覚処理モデル

[3]. Kazuhiro Nakadai, Hiroshi G. Okuno, Hiroaki Kitano: Exploiting Auditory Fovea in Humanoid-Human Interaction. Proceedings of the Eighteenth

National Conference on Artificial Intelligence (AAAI-2002), pp.431-438, Edmonton, Canada, Aug. 2002.

[4]. Kazuhiro Nakadai, Hiroshi G. Okuno, Hiroaki Kitano: Auditory Fovea Based Speech Separation and Its Application to Dialog System. Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2002), IEEE, Lausanne, Swiss, Oct. 2002.

第 4 章 動作時のノイズキャンセルと音源定位

文献 [1] の他に,

[5]. Hiroshi G. Okuno, Kazuhiro Nakadai, Lourens Tino, Hiroaki Kitano: Sound and Visual Tracking for Humanoid Robot. Applied Intelligence, Kluwer Academic Publishers, 2003(印刷中).

第 5 章 聴覚情報の統合による音源定位と追跡

[6]. Kazuhiro Nakadai, Ken-ichi Hidai, Hiroshi G. Okuno, Hiroaki Kitano: Epipolar Geometry Based Sound Localization and Extraction for Humanoid Audition. Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2001), pp.1395-1401, IEEE, Maui, Hawaii, Oct. 2001.

第 6 章 視聴覚統合による実時間複数人物追跡

[7]. 中臺 一博, 日台 健一, 溝口 博, 奥乃 博, 北野 宏明: 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 –. 日本ロボット学会誌, vol.21, no.6, 2003(印刷中)

[8]. Kazuhiro Nakadai, Ken-ichi Hidai, Hiroshi G. Okuno, Hiroshi Mizoguchi and Hiroaki Kitano: Real-Time Auditory and Visual Multiple-Speaker Tracking For Human-Robot Interaction. Journal of Robotics and Mechatronics, pp.479–489, JSME, 2002.

[9]. Hiroshi G. Okuno, Kazuhiro Nakadai, Ken'ichi Hidai, Hiroshi Mizoguchi, Hiroaki Kitano: Human-Robot Non-Verbal Interaction Empowered by Real-Time Auditory and Visual Multiple-Talker Tracking. Advanced Robotics, RSJ, 2003(印刷中).

第 7 章 アクティブ方向通過型フィルタによる音源分離

文献 [3] の他に,

[10]. Kazuhiro Nakadai, Ken-ichi Hidai, Hiroshi G. Okuno, Hiroaki Kitano: Real-Time Speaker Localization and Speech Separation by Audio-Visual Integration. Proceedings of IEEE International Conference on Robotics and Automation (ICRA 2002), pp.1043-1049, Washington D.C., May.

2002.

第 8 章 複数の音響モデルを利用した音声認識文献 [4] の他に,

[11]. Kazuhiro Nakadai, Hiroshi G. Okuno, Hiroaki Kitano: Auditory Fovea Based Speech Enhancement and Its Application to Human-Robot Dialog System. Proceedings of 7th International Conference on Spoken Language Processing (ICSLP-2002), Denver, USA, Sep. 2002.

第 9 章 ヒューマンロボットインタラクションへの応用文献 [8] の他に,

[12]. Hiroshi G. Okuno, Kazuhiro Nakadai, Hiroaki Kitano: Realizing Audio-Visually triggered ELIZA-like non-verbal Behaviors. Seventh Pacific Rim International Conference on Artificial Intelligence (PRICAI-2002), pp.552-562, Tokyo(NII), Japan, Aug. 2002.

第 10 章 考察

第 11 章 結論

第2章

ロボット聴覚の課題と現状

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．書かれた言語は目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)

本章では，まず，ロボット聴覚という研究を定義し，その上でロボットが人間のパートナーとして社会で人間と共生していくために，現状，ロボットの聴覚機能で問題となっている点の整理，最終的に達成しなければならない課題について議論を行う．次に，研究の方向性をより明確にするために，ロボットの聴覚を実現するために必要なマイクロホンの本数について議論を行い，2 本のマイクロホンが必要であるという考えを述べる．そして，2 本のマイクロホンでの処理を前提に，ロボット聴覚を実現する上で，アクティブオーディション，混合音が扱えること，動きながら聞く機構，未知環境や実環境での音の知覚，画像処理などの他の処理との統合，実時間処理といった事項が大きな課題であることを指摘し，従来の手法との比較を通じ，実現可能性について議論する．さらに，このようなロボット聴覚の工学的な実現を通じて，従来のヒューマンロボットインタラクションと比較し，自由度の高い，より自然なインタラクションの実現について“自然な”という言葉の定義を含め議論する．

2.1 ロボットにおける知覚処理

近年，ロボットは，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, 周囲の状況をロバ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.

こうした知覚機能のうち, 視覚機能は, 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比較しても, 相当機能が低いと言わざるをえない.

実際, 聴覚機能は, 将来, 我々の社会で日常的に人間と知的なインタラクションを行う上で, ロボットに求められる最も基本的で重要な機能である. 以下では, 人間とコミュニケーションを行う上で最も重要な機能の一つである聴覚機能を中心に議論を行う.

2.2 ロボット聴覚とは

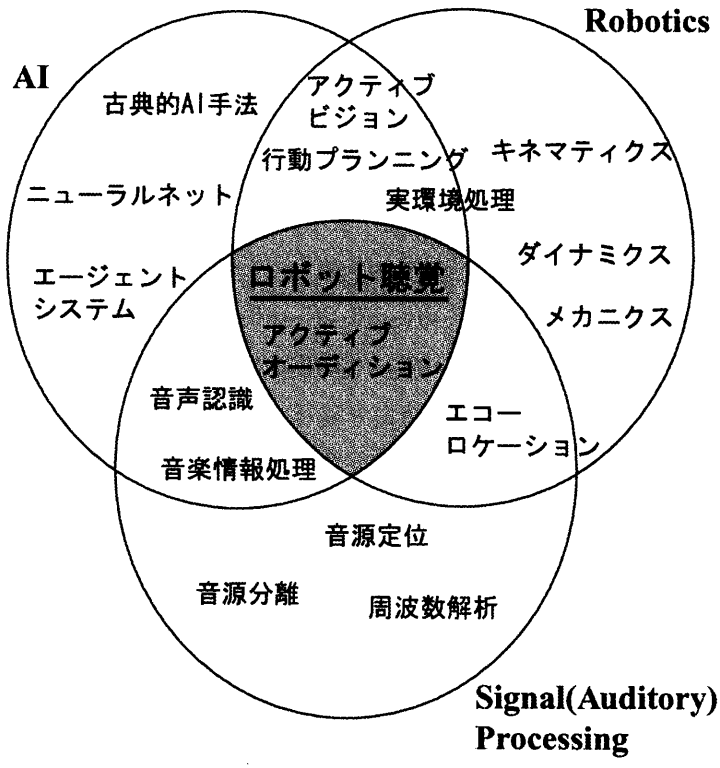


図 2.1: ロボット聴覚の位置づけ

一般に聴覚処理の工学的な実現には様々な課題がある. まず, 入力に音響信号が存在するかどうかの判断が必要である. これは, ゼロクロス法や振幅値に一定の閾値を設ける方法がよく行われるが, 汎用的な手法についてはあまり議論されていない. 入力信号に対して, どのような周波数解析が適しているかという点についても様々な議論がある

[53, 2]. また, 周波数領域では, 調波構造を持つ音なのか, バースト的な音なのかという判断, ピッチ情報の抽出, オンセット, オフセット, アタックの抽出, 倍音の抽出, 位相や強度の抽出, 倍音間の共有振幅変調 (共有 AM), 共有周波数変調 (共有 FM) など様々な特徴量を抽出する必要がある. 特に, Bregman が聴覚心理学の見地から "Auditory Scene Analysis"[24] を著してからは, 音環境理解 (*Computational Auditory Scene Analysis*, CASA) の分野を中心に盛んにその抽出法や利用法についても様々な議論が行われてきた [27, 83, 127, 32, 84, 106]. 音源方向, 音源距離といったパラメータを推定する音源定位も様々な研究が行われている [113, 47, 1]. さらに, 複数の音が混在している場合には, どの音がノイズで, どの音が目的とする信号なのかを判断した上で, ノイズを抑制, もしくはキャンセルしたり, 音源を分離したりする必要があるだろう. 音源定位や分離については, 上述の聴覚の特徴量を利用する方法だけでなく, ビームフォーミングや独立成分分析といった信号処理的な手法による音源分離, 強調を行う方法も盛んに行われている [57, 9, 109, 113, 13, 91, 95, 49] . また, 音源同定も楽音の採譜などを考えた場合に重要な問題である. これは, 音楽に限れば, 楽器同定に関する研究 [61, 62, 67], 音声に限れば話者同定に関する研究で盛んに行われている. また, その前段階で, 音声なのか, 楽音なのか, それとも他の音なのかといった音の種類を判断する研究も重要である. これに対しては, 様々な音を統一的に扱う枠組みとして, 音オントロジーが提唱されている [82]. より高次な処理としては, 音声認識も重要な課題であり, 如何にノイズに強い認識を実現するかだけではなく, 未知語への対応や, マルチリンガルな音声認識など, 様々な課題が残されている.

このような信号処理, 聴覚処理では, 古典的な AI 手法を含め, 学習, ニューラルネットワークといった人工知能の分野で培われてきた手法が適用されている. 一般的な音声認識手法として隠れマルコフモデルが用いられているのは, その典型的な例であるし, 近年盛んな脳科学と聴覚処理を結びつけるような研究 [124] でも, AI 的な手法の果たす役割は大きいと言える.

AI 的な手法は, 実環境処理や人間とのインタラクションという面から, ロボットにも積極的に利用されている. しかし, ロボットの世界では, 音響処理は超音波を用いたエコーロケーションなどを除き, あまり取り組まれてこなかった. もちろん, ロボットで聴覚処理を扱った研究も見受けられるが [9, 115, 121, 22], 明示的に, 人間や動物が行うような, 動作を積極的に利用して聴覚を向上させる研究, つまり, 図 2.1 に示すような, 聴覚処理, ロボット, 人工知能を複合的に組み合わせたロボット聴覚の研究は行われてこなかった.

そこで, 本研究では, ロボット聴覚研究を聴覚処理, ロボット, 人工知能を複合的に組み合わせた新しい研究と位置づけ, 特に実環境で動作するロボット特徴を生かし, 如何にアクティブな動作を行うことにより聴覚を向上させられるか (アクティブオーディション) という課題について扱う. 具体的には, 後述するが, ロボット動作時のノイズキャンセル,

および複数音源の定位・分離・認識を通じて, より自然なロボットと人間のインタラクションを実現することにより, その有効性を評価する.

2.3 マイクロホンの数に関する検討

ロボット聴覚の課題を議論する前に, 研究の方向性をより, 明確にするために用いるマイクロホンの数に関する検討を行う. ロボットに用いるマイクロホンの数には制限がないため, マイクロホンの数が多いほど, 精度の高い定位や分離処理ができるという考え方がある. 実際には, ロボットに搭載されるマイクロホンの本数は, 2 本, ないし, 3 本のものが多いが, 中には 8 本のマイクロホンで構成されたマイクロホンアレイを搭載している [15] ものもある. 本研究では, ロボット聴覚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ホンの数が少ない」**

という状況を想定している. このような条件の下で,「耳が2つある」ことがロボット聴覚において本質的であるかを検討する. なお, ヒトの聴覚については生理学や心理物理学などで数多くの知見 [24, 20] が得られているので, そのような研究は他に譲る.

混合音の分離には, 一般的にマイクロホンアレイが使用される. マイクロホ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がよく使われる. ビームフォーミング技法については, 2.5.3 節で述べる. 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 (*Independent Component Analysis, ICA*) を使えば, N 本のマイクロホンで N 個の音源を理論的に分離することができる [78]. 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, 独立成分解析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 [94]. 独立成分分析については, 関連研究として, 2.5.3 節で簡単な説明を行う.

Wang らも複数のマイクロホンを使用することによって, 音響ストリーム分離の精度向上を達成している [121]. つまり, 複数のマイクロホンを使用すれば, 単一マイクロホ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.

つまり, マイクロホンの数を増やせば, 音源分離の性能が向上することは, 理論的にも, 実験的にも明らかである. では, マイクロホンは何本用意すればよいのだろうか. 「想定される音源の数より多ければよい」が正解なのであろうか.

一般環境では, マイクロホ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ホ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ホンは, 頻繁に動くことになり, マイクロホン間の相対位置も同様に頻繁に変化するためである. また, ロボットの体に何十本というマイクロホンを装着することは現実的ではない. マイクロホンが移動するような状況に対して, 適用可能なマイクロホンアレイを開発しようというプロジェクトも始まっている [116].

本研究では, このような問題を「マイクロホ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」ととらえている. この場合, 次のように問題を整理することができよう.

1. 体が固定されたときに, 耳がいくつ必要か.
2. 逆に, 自由に体の動きが許されるときに, 耳がいくつ必要か.

前者の問題については, 方向情報を得るためには, 最低 2 本のマイクロホンが必要であると考えている. 後者の問題については, 最少解は 1 本のマイクロホンであろう. しかし, 上述したような理由から, 実用的な意味での最少解は, 2 本のマイクロホンではないかと予想している.

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, 眼球で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 [46]. 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. これは, 他方の耳にマスキング信号*1 を入れ, 片耳のみで方向情報が分かることで実感することができる. しかし, 動的に振幅やピッチが短時間で変化する音に対しては, 片耳で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. 従って, 1 本のマイクロホンでは画像のような汎用的な微動型センサーの工学的な実現は難しいと思われるが, 2 本のマイクロホンを用いれば, 体を動かすことにより, 多くの状況に対応できよう. これには, 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. 例えば, 聴覚の有用な情報の一つである方向情報を安定して得るためには, 最低 2 本のマイクロホンが必要である. 本研究では, 視覚情報を使用したり, 体を動かしたりすることにより, 2 本のマイクロホンを用いてこの問題の解決を図る.

*1 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.

2.4 アクティブオーディション

ロボットに周囲の状況をロバストに知覚させるための研究は, 簡単な問題ではないが, 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) の向上やロバストな知覚の実現への糸口を与える重要な研究である [6, 17].

アクティブパーセプションは, ビジョンの分野では, 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 [5, 69]. 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 [73, 103, 118]. 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, ASA*) を向上させている.

そこで, 本研究では, ロボットを対象に聴覚とモータ動作を統合した**アクティブオーディション**を提唱し, 人間や動物を見習って, アクティブな動作を積極的に利用した, ロボット聴覚の向上を目指す. 特にアクティブオーディションに基づき, ロボット動作時のノイズキャンセル, および複数音源の定位・分離・認識を中心に扱う. また, ロボットと人間のインタラクションにおけるアクティブオーディションの効果についても議論する. ただし, 本研究で扱うアクティブオーディションは, パッシブなセンサ (マイクロホン) 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.

2.5 アクティブオーディションの実現に向けた課題

アクティブオーディションに基づいて, ロボットが, アクティブな動作による恩恵を最大限に享受するためには, 動きながら音を聴く能力が必要である.

従来の聴覚研究は, そのほとんどが実環境を対象としたものではなく, オフライン, かつ, シミュレーション環境で行われている. また, 静的な環境を対象にしているので, 実環境で動作を行うロボットへの適用は考慮されていない. 音響処理における動作問題への考慮が不十分なロボットでは, 例えば, ノイズの問題により, "*stop-perceive-act*" 戦略をとらざるを得ないなど, 大きな制約を抱えている. この制約を緩和し, アクティブオーディ

ションをロボットに実現するためには, 動作時のノイズ問題を含め, 以下に挙げる 4 つの課題に取り組む必要がある.

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. 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)
4. 様々なセンサ情報の統合

以下では, それぞれの項目について詳しく述べる.

2.5.1 ロボット自身が発生する音の抑制

ロボットの知覚向上を考える上で, アクティブオーディションという考え方への到達は, 自然な流れである. しかし, 人間では, 聴覚は視覚と同様に主要なセンサであるにもかかわらず, 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. この原因として, 聴覚情報にはノイズが比較的混入しやすく, 聴覚情報に大きな影響を与えてしまうことが挙げられる. 例えば, 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. また, 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も難しい. 特にアクティブな動作を行うロボットにおいては, モータ, ギア, ベルト, ベアリングの回転などにより, ノイズ問題が発生する. さらに, ノイズ源は外部の音源に比べてマイクに近いため減衰は小さく, パワーの絶対値が小さいノイズであっても, マイクには強いパワーをもった音として収音されてしまい, 入力信号に大きな影響を与える. このため, 多くのロボットは, 一旦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. そこで, ロボットにおけるノイズ問題がこれまでどのように扱われてきたか, 実際に, これまでに聴覚機能を有したロボットを例に挙げ, その問題点を解析する.

■動作ノイズ問題: "stop-perceive-act" ロボット まず, 比較的身近な Sony AIBO を挙げてみよう. このロボットは, 「写真を撮って」というと, 鼻の位置にある CCD カメラで写真を撮ってくれるなど, 簡単な音声認識機能を備えているが, ロボットが動作時に発するノイズは, その場にいる人がうるさいと感じるほど大きい. 実際, この音声認識機能は, 動作中には働かず, ロボットが静止している場合に限られている. つまり, AIBO の音声認識は "*stop-perceive-act*" 原理に従っている.

研究用途で開発されているロボットを見てみよう. 早稲田大学の *ROBITA* [74] や MIT AI ラボラトリの *Kismet* [23] 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, 顔の表情や音声で表現することができる. これらのロボットでは, 人間とのインタラクションにおける制約を緩和するため, "*stop-perceive-act*" 原理を避ける必要がある. そのため, ロボット自身に備わったマイクロホンを利用せず, 対話をする話者の口元にマイクロホンを備えている. これにより, 話者の音声情報の信号対ノイズ (S/N) 比を高め, 室内音響の影響やロボット自身が作り出すノイズの影響を回避している. このようなアプローチは, 音源情報を精度よく抽出するという意味で有効であるが, ロボットとインタラクションを行う人すべてがマイクロホンを口元につけていなければならず, 本研究で仮定しているようなロボットが人の社会で共生することを考えると, 大きな制約といえよう.

早稲田大学の *WE-4* [76] は, 視覚, 嗅覚, 聴覚情報を統合した行動制御により, 興味のある対象を追跡することができる. さらに, 刺激に対して感情豊かな表情を作ることもできる. 聴覚機能としては, ロボットの左右の耳の位置にマイクロホンを備えており, 音圧情報を利用して, 単一音源の追従が可能である. しかし, このロボットでは対象音源がロボットのノイズが無視できるくらい大きな音量であることを仮定している. このため, 仮定している音源に対しては, 追従が可能であるが, 実環境では, 一般にこの仮定が成り立たないため, 完全な "*stop-perceive-act*" の解決とはなりえない.

Huang らは, 球形の頭部を持ったロボットに 3 本のマイクロホンを搭載したロボットを用いて, 全方位の音源追跡を実現している [47]. 基本的には, 2 本のマイクロホンによる定位を行い, 2 本のマイクロホンを使用する際に生じる前後問題を, 3 本のマイクロホンのうち, パワーの大きい 2 本のマイクロホンを選択することで解決している. しかし, このロボットは, やはりノイズの問題から, 動きながらの定位は難しく, 音を聴くためには一旦静止する必要がある. また, 自分自身の動作ノイズを扱えないのと同様, 入力は単一音源であることを仮定しているため, 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も困難である.

この制約は, 産業総合研究所の *Jijo-2* [15, 13] にも当てはまる. このロボットは, 8 チャネルのマイクロホンアレーを搭載し, 一般的なオフィスで音源の定位・分離を行い, さらに簡単なフレーズによる命令を認識することができる. しかし, 動きながら音声を認識するためには, ノイズの影響や, マイクロホンアレーの性質として, そのパラメータを適応的に制御する必要があるため, 音声認識を行うためには静止する必要がある. また, 報告されている方法は, 予め多くの測定が必要であることから, 実環境への適用を考えた場合, 十分とはいえない. マイクロホンアレーを用いたロボットによる話者トラッキングについては西浦らの報告 [91] もあるが, 現時点ではシミュレーション環境での定位に留まっている.

ノイズ問題に対して, 機構的な改善により, 動作時のモータノイズを低減する試みもなされている. 早稲田大学の *WA-2* は, 機構的な改良により, ロボットに備わったマイクロ

ホンを使用する際のモータノイズの問題に対して取り組み, 無響室で単一音源の連続定位に成功している [114] . さらに, 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も実現している [115]. 確かに, 機構的な改良による動作ノイズ低減は, 人間や動物のアクチュエータが静音で動作することからも, 人間と共生するための要件であり, 静かなロボットを実現することは重要な研究課題といえる. しかし, たとえロボット自身は静かであっても, 自身の声, 何かと衝突した場合の音, 服や外装が擦れあう音などロボット自身からノイズが発生する場合があり, このような場合は, 機構的な改良だけで完全に対処することは難しい.

■内部音の抑制 ロボットは, 知覚処理を向上させるという意味において, アクティブな動作を積極的に行うべきであり, ロボットを実環境でのアプリケーションとして捉えた場合, 動作中であっても, ロボット自身に備わったマイクロホンで聞くことが求められる. しかし, 実環境で動作することが期待されているロボットでは, 何らかのノイズ対策を行わなければ, 聴覚処理が難しい.

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, 聞くことである. 実際, 上述のように, この "*stop-perceive-act*" 法を, マイクロホンを搭載した多くのロボットが採用している. また, 対象音をノイズが無視できるくらい大きくしたり, 話者の口元にマイクロホンを設置して対処しているロボットもあることも述べた. しかし, このような方法はロボットが発生するノイズの問題の本質的な解決にはならない. 機構的な改良を施すことにより対処する試みもされているが, 機構的な改良だけでは, やはりノイズ問題の完全な解決とはなりえない.

従って, 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, ソフトウェア的に内部ノイズを抑制し, 外部音を強調するようなアプローチが必要である. 特にアクティブパーセプションを実現する場合, 様々な可動部分の動きによって音が発生するので, 外部音の S/N 比を高めるため, 内部ノイズを抑制することが重要である.

このような問題に対し, ロボットの外装は有効である. ロボットの外装は, 内部の機構を保護したり, プロダクトとしてのデザイン性という意味だけではなく, 図 2.2 に示すように, 内部音を抑制する有効な手段となりうるからである. 図 2.2 では, 外装によって, 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ぐことができるようになっている. 内部と外部を音響的に隔離し, ロボットに音響的な身体性を持たせることにより, 例えば, 外装の内部と外部にそれぞれマイクロホンを設置すれば, モータ音による妨害を軽減することが期待できる.

このような場合に, 一般的なノイズキャンセル法として, アクティブノイズコントロール (*Active Noise Control, ANC*) や, 音源分離で用いられる独立成分分析やビームフォーミングを利用した手法が考えられる. アクティブノイズコントロールは, 騒音と逆位相の

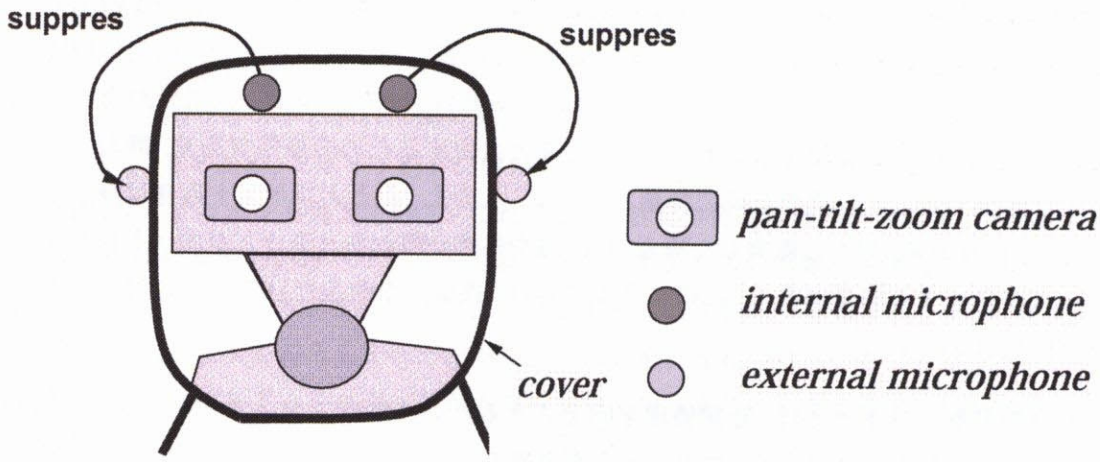


図 2.2: ノイズを抑制するための外装内外のマイクロホン

音を発生させることによって騒音を打ち消し, ノイズを抑制することができる [41, 87]. フィードフォワード制御, フィードバック制御, 適応予測制御など様々な制御方法が提案されているが, 図 2.3 は, フィードバック制御の構成例である. 音源からマイクロホンまでの伝達関数を $H_s(t)$, ノイズ源を $N_i(t)$ とし, マイクロホンで収音される信号を $N_o(t)$ とする. この時, $N_o(t)$ から, 目的の地点でノイズを打ち消すような逆位相の信号 $N_s(t)$ を予測し, ノイズを抑制することが ANC の基本的な原理である. 実際には, ノイズ源以外に

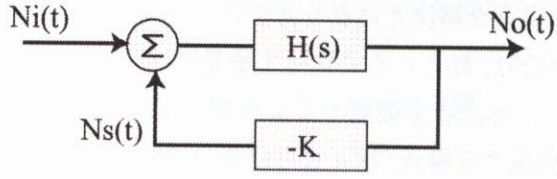


図 2.3: アクティブノイズコントロールの構成例

信号源が存在する場合や, 複数のマイクロホンを用いてノイズを抑制する場合など, 様々なケースが想定される. 実際に, ANC は近年工場や飛行機などの騒音低減のために実用化されており, 注目を集めている技術である.

独立成分分析やビームフォーミング手法については, 2.5.3 節で述べるが, これらは, ノイズ源を分離したり, ノイズ源方向に指向性の死角を作ることによって, ノイズの抑制を行うことができる.

しかし, このような一般的な手法では, 音源数が動的に変化したり, ノイズ源とマイクロホンの位置が相対的に変化したりするため, ロボット動作時のノイズをキャンセルすることは難しい. そこで, 本研究では, ロボット外装の音響効果を積極的に利用したヒューリス

ティックによるバーストノイズキャンセル法を提案している．実際に，従来手法との比較を含めたノイズキャンセルの詳細については 4 章で述べる．

2.5.2 未知環境における音の知覚

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 本のマイクロホンを設置するという前提で，音源定位問題を扱う．まず，2 本のマイクロホンを用いた一般的な音源定位の方法について述べる．

■従来の信号処理的な音源定位手法 信号処理的な手法を利用した音源定位について述べる．このような手法として，一本のマイクロホンで音源定位を行う方法も提案されている [117] が，一般には，複数のマイクロホンを用いる．2 本のマイクロホンを用いた場合はチャネル間の相関を求めるクロススペクトル法がよく用いられる [119]．この手法は，FFT を用いた高速な計算が可能であるため，音源数が 1 つの場合は，有効な手法であり，ロボットにも適用されている [108]．

2 本以上のマイクロホンを用いる場合にも適用可能な方法として，2.5.3 節で説明するビームフォーミングの一手法である遅延和法を用いて，パワーの大きい方向を音源方向とするノンパラメトリックな手法が挙げられる．この手法は，2 本のマイクによる遅延和型アレーを用いた場合，クロススペクトル法と等価である．Silverman や Branstein らは，マイクロホンアレーによるビームフォーミングに関して多くの研究を行っており，遅延和アレーとカルマンフィルタを組み合わせることにより複数話者の追跡も達成している [113]．また，音源数がマイクロ本数 $M-1$ であることを仮定して，よく知られたスペクトル推定法を用いて，音源方向を推定するパラメトリックな手法も挙げられる．以下でそのいくつかを紹介する．

- 線形予測法

  各マイクロホンへの受音信号の複素振幅系列を $X_1, X_2, \cdots, X_M$ とした場合，この系列に含まれる正弦波の周波数が，音源方向と対応している．従って，線形予測法を適用すると，線形予測係数 $\boldsymbol{\alpha}$ は，予測誤差 e

  $$e = X_1 + \sum_{i=2}^{M} \alpha_{i-1} X_i$$

  の 2 乗期待値を最小にするものとして，

  $$\boldsymbol{\alpha} = [1, \alpha_1, \alpha_2, \cdots, \alpha_{M-1}] = [1, 0, 0, \cdots, 0]\boldsymbol{R}^{-1}$$

  として計算される．R は相関行列である．従って，パワー推定式は，

  $$P(\theta) = \frac{1}{(\boldsymbol{\alpha d}(\theta))^2}$$

となり, $P(\theta)$ が最大となる方向 θ を音源方向とみなすことができる. なお, $\boldsymbol{d}(\theta)$ は, $\boldsymbol{d}(\theta) = [1, e^{-j\omega\tau}, e^{-j\omega 2\tau}, \cdots, e^{-j\omega(M-1)\tau}]$ と表される方向制御ベクトルである.

- 最小分散法

  最小分散法は, 注目方向以外から到達する音を最小にする推定方法である. この場合, パワー推定式は,

$$P(\theta) = \frac{1}{d(\theta)^* R^{-1} d(\theta)}$$

  となる. 各パラメータの意味は, 線形予測法で用いたものと同じである. この手法は, 音源方向以外の方向に現れる虚ピークが現れにくい, 不等間隔アレーに用いることができる, 音の到来方向に対する感度が高いといった利点がある.

- MUSIC 法

  MUSIC 法は, 近年, スペクトル推定によく用いられている方法であり, 上述の手法と比較し, 相関行列 R の性質を積極的に利用した方法である. MUSIC 法では, パワー推定式は,

$$P(\theta) = \frac{1}{d(\theta)^* R_n d(\theta)}$$

$$R_n = \sum_{q=K+1}^{M} v_q v_q^*$$

  と表される. K は到来音数, v_q は相関行列の固有ベクトルを意味している. 到来音波数 K を何らかの手段で取得しなければならないが, きわめて高い分解能で音源方向を推定することが可能である. 例えば, MUSIC 法を用いて, 頭部の両耳にマイクロホンを設置し, 頭部の回折係数を含めた方向制御ベクトルを利用した音源定位法も提案されている [37].

パワースペクトル推定法を利用した音源定位には, この他 Burg 法, Multitaper 法 (MTM), Welch 法, Yule-Walker 法といった手法が挙げられる.

その他の手法として, チャネル間の微分積和量を利用した時空間勾配法も提案されている [9]. この手法は, ロボットへの搭載が可能であり, ロボットが動作し, マイクロホンの向きが変わる場合でも 3 次元定位が可能な手法である. この手法では, 微分積和量の行列のランク値を利用して, 音源数の推定も可能である. 一辺の長さ W の正方形の頂点に 4 本のマイクを配置した場合を考える. マイクロホンを時計回りに, A,B,C,D とし, それらの出力を f_A, f_B, f_C, f_D とすると, 正方形の中心での音圧場 f と空間勾配 f_x, f_y, 時間勾配 f_t は,

$$f \simeq \frac{1}{4}(f_A + f_B + f_C + f_D)$$

$$f_x \simeq \frac{1}{2W}(f_A - f_B - f_C + f_D)$$
$$f_y \simeq \frac{1}{2W}(f_A + f_B - f_C - f_D)$$
$$f_t \simeq \frac{1}{4}\frac{\partial(f_A + f_B + f_C + f_D)}{\partial t}$$

と表すことができる．これらの値に対して，相互相関行列を求め，この固有値の数と対応する固有ベクトルによって音源数の推定と音源定位を行う．

4 本のマイクを搭載したロボットでは，音源数が 1 音源の場合，3 次元定位が可能であり，2 音源では，2 音源間を結ぶ共存直線として音源の定位が可能である．また，3 音源以上では，音が存在するか否かを判定することが可能となっている．さらに，この処理は 5 ms の分解能で 5 度以内の誤差での定位能力を持っており，かつ，毎秒 1000 回以上の定位が可能である．

しかし，一般に定位すべき音源数は未知であること，これらの手法では，マイクロホン数が多くないと定位精度が低いこと，ロボットでは，マイクロホンと音源との距離や位置関係が動的に変化することを前提としなければならないことなどから，純粋な信号処理的の手法では，実環境で十分な性能を得ることが難しい．そこで，次節では，実環境でロバストな音源定位が可能であり，2 つの耳でも複数音源の定位が可能な人間の聴覚処理に習った手法を紹介する．

■2 本のマイクロホンによる音源定位 人間の耳の聴覚機能として，音源定位のモデルとして知られている Jeffress の cross-correlator モデルは，2 本のマイクロホンから得られた入力音を遅延させながら相関をとり，最大の相関を与える遅延から両耳間時間差 (*Interaural Time Difference, ITD*) を求めるものである [56]．ITD は，両耳聴の工学的な研究では，しばしば音源定位に使用される．実際には，特徴的なオンセットを持つ音でない限りノイズの問題で正確な ITD を直接抽出することは難しく，クロススペクトル法を利用しているものが多い [119]．

しかし，クロススペクトル法では，人間と同じ大きさの頭部を仮定した場合，一般に 1500Hz 以上の高周波の定位は難しい．これは，ITD の代わりに両耳間位相差 (*Interaural Phase Difference, IPD*) を使用することにより，理解することができる．両耳間の顔の表面上の距離を，おおよそ 23cm とすると，IPD が初めて 2π になる，つまり，定位の際に，1 周期回り込むことによって曖昧性が簡単に解消できなくなるのが，1500Hz である．また，位相差が π となる 750Hz から，前後が分からなくなるという曖昧性が生じる．

実際，人間の場合でも，音源定位の際に ITD(IPD) が有効に働くのは，1500Hz までであり，それ以上の周波数帯域では，両耳間強度差 (*Interaural Intensity Difference, IID*)*2

*2 両耳間強度差 (*IID*) は interaural amplitude difference, *IAD* などとも表記する．

の方が有効に働くと言われている [77]. つまり, ITD と IID は周波数帯域によってその貢献度が異なり, Jeffress のモデルも, 後の研究では ITD だけではなく IID も音源定位に寄与するように修正されている.

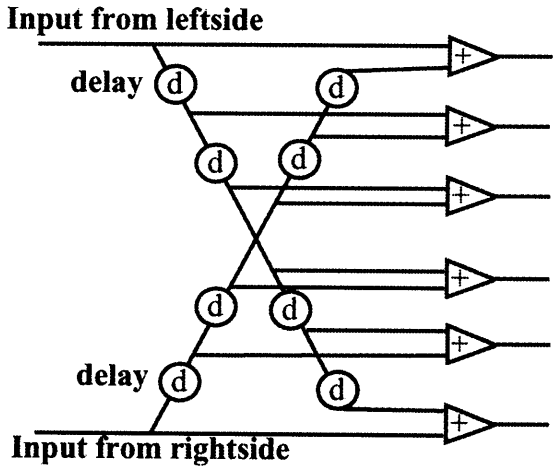


図 2.4: Jeffress の Cross-Correlator モデル

■頭部伝達関数の制約 両耳聴の研究では, IPD や IID を導出するために頭部伝達関数 (Head-Related Transfer Function, HRTF) が, しばしば利用される. HRTF は, 通常, 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される, 人間の頭部形状による音響特性の変化を表す伝達関数である.

中谷と奥乃らのグループは音源方向と音の調波構造に注目した音源分離システム Bi-HBSS を構築し, 無響室環境で 2 話者同時発話の音声分離を行っている [86]. このシステムは, 調波構造を抽出し, 左右の耳で同じ音源から来る調波構造を持った音のペアを抽出し, その音の各倍音ごとの IPD と IID を計算し, それらの値から 5 度刻みで測定した HRTF を基に方向情報を求める. そして, 調波構造断片を方向情報でグルーピングし, 調波構造ストリームを抽出する. この方法は, 基本周波数が重ならない限り, 理論的には音源数がいくつあっても, 音源定位とそれに基づいた音源分離が可能である. 実際には, 混合音中の調波構造抽出の精度は音源数が増えるにしたがって低下するので, 音源数が増えると精度のよい分離は難しくなる. 例えば, 2 話者同時発話の上位 10 位の単語認識率 75%が, 3 話者同時発話認識では 30%程度と大幅に分離性能が劣化する [97].

無響室以外では, 実際の環境での空間伝達特性を畳み込まないと, 得られる方向情報の精度が低下する. そのため, マイクロホンの位置が変化するようなシステムの場合には, HRTF の測定と同様に, スピーカ位置を変化させるだけでなく, マイクロホンの位置も変化させて, 空間伝達特性を測定する必要がある. また, 測定点数の制限から離散的な方向情

報しか使えないので, 移動音源への対処は難しくなる. 従って, ロボットで, 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必要である.

1. **HRTF に基づく音源定位**. 通常, HRTF は無響室で計測される. しかし, 無響室で測定した HRTF を実環境の音源定位に適用すると, 部屋の反響の影響により, 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. また, 音響ストリーム分離においても連続的な定位情報が得られないことは致命的である.

■頭部伝達関数を用いない音源定位手法 そこで, HRTF を用いない音源定位手法として, 聴覚エピポーラ幾何を提案する. 聴覚エピポーラ幾何は IPD を計算的に求めることができるため, 測定が不要であり, 連続的な定位が可能である. また, 基本的には, IPD を用いた定位手法であるため, 信号処理的な手法で問題となる音源数とマイクロホン数の間の理論的な制約がなく, パラメータのチューニングなども必要としない. また, 高速な処理が可能であり, ロボットに搭載して音源追跡を行うような用途に適している. 実際には, 提案した聴覚エピポーラ幾何は音源分離に対しても有効に働く手法であり, 本研究で音源分離を行うために使用するアクティブ方向通過型フィルタにも取り入れられている. 詳細については, 4 章で述べる. また, 複数の聴覚情報を統合して, 音源定位のロバスト性を高める手法についても本研究では扱っている. これについては, 5 章で述べる.

2.5.3 一般音理解

ロボットをリモートエージェントに適用する場合を考えてみよう. 遠隔会議に適用すれば, ロボットは身体性が利用できる分, テレビ会議などと比べて人間と円滑なコミュニケーションを行う, テレイグジスタンスが実現できよう.

しかし, このような状況では, 人間と同様, ロボットは複数のイベントを同時に聞き分ける必要がある. 会議では, 複数の人が同時に話していても, 特定の人に注意を向けたり, 極地作業では, 作業中でも身に危険を及ぼすようなイベントを検出するなど, 目的の音響信号に対して信号対ノイズ (S/N) 比の高い信号をオペレータに供給することが求められる.

そのためには, 他の音声や自分自身が作り出すモータノイズを抑制することに加え, 混合音をうまく扱えるように音源分離機能を備える必要がある.

このように音源分離は, リモートエージェントロボットでは, 特定の音響信号を強調するために重要な機能である. また, 一般に, 人間が日常生活で耳にする音は単一音源からの音ではなく, 複数の音源からの音が混じった混合音である [106]. 従って, 遠隔会議に限らず, 人間と共生するような日常活動型ロボットには, 必須の機能であるといえる. また, 音源分離は人間とコミュニケーションをとるために, 音声認識のフロントエンド処理としても有効な機能である.

つまり, ロボットにおける音源分離機能は「**実時間・実環境で音声認識のフロントエンド処理として使用することができる程度の分離能力**」という要件を満たす必要があるといえる.

■従来の音源分離処理 従来から研究されてきた音源分離処理について音環境理解, ビームフォーミング, 独立成分分析の 3 つのアプローチを説明し, ロボットへ適用した場合の問題点を明らかにする.

音環境理解によるアプローチ: このような一般的な音の理解は, ロボットの分野よりは, 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 (CASA) の分野で研究されてきた.

一般に音源分離問題は不良設定問題であり, 工学的には一意に解くことはできない. このため, 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 [27, 32, 84, 106]. 近年では, 音声認識の分野でも, 一般的な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ュミレーションしているシステム [27, 112], 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 [27], 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) [85], 音楽を対象とした音源分離システム [63] や, 調波構造を利用したマルチエージェントによる音源分離システム [83] 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, オフライン, かつ, 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・実時間処理への配慮が十分ではなかった.

実時間処理に関しては, 2 本のマイク間の強度差と位相差を利用し, 18 dB 以上の実時間

音声強調が報告されている [11]. 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.

また, 視聴覚統合によって, 音源分離を向上させようというアプローチも行われている [81]. しかし, 評価は, オフラインで, かつ無響室で計測した頭部伝達関数 (HRTF) を畳み込んで生成したデータを用いたシミュレーション環境で行われているため, やはり, 実環境や実時間での処理には問題がある.

ビームフォーミングによるアプローチ: 信号処理的な手法を利用した音源分離としては, 従来から盛んに行われているマイクロホンアレイを用いたビームフォーミング [109, 65, 113, 13, 91] が挙げられる. これは信号処理の分野では, 古くから音源分離の手法として知られてきた方法である [101]. 一般的な M 個のマイクロホンを用いたマイクロホンアレイを図 2.5 に示す. このマイクロホンアレーに音源方向 θ の平面波が到達すると

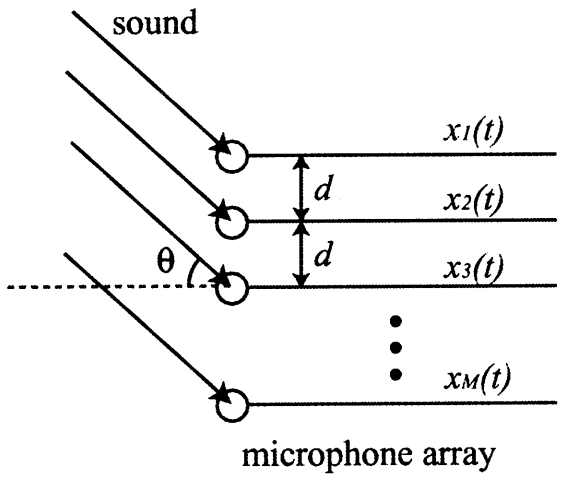


図 2.5: マイクロホンアレイ

仮定すると各マイクロホンに到達する音波は, 隣のマイクロホンに到達する音波と比べ,

$$\tau = \frac{dsin\theta}{v}$$

だけ遅れて到着する. ここで, d はマイクロホン間距離, v は音速である. つまり, 各マイクロホンで収音される信号は, 1 番目のマイクロホンで収音される信号を用いて,

$$x_i(t) = x_1(t - (i-1)\tau)$$

とあらわすことができる. 実際には, 振幅減衰も考慮すべきであるが, ここでは簡単のため時間遅れのみを考えるものとする. また, 信号の各周波数を ω とすれば,

$$x_i(t) = X_1 e^{jw(t-(i-1)\tau)}$$

と表すことができる．ただし，X_1 は1番目のマイクロホンで収音される信号の複素振幅を表す．

ビームフォーミング手法は，このマイクロホン間の時間遅れを利用して，音源方向を推定し，目的信号を抽出する手法である．用いられる手法は，遅延和型と適応型の2つに大きく分類される．

- 遅延和型マイクロホンアレー

  遅延和型マイクロホンアレーでは，図2.5の各 $x_i(t)$ に，

  $$D_i = d_0 - (i-1)\frac{dsin\theta_s}{v}$$

  という遅延を付加し，θ_s 方向から来る音波の位相を同相化する．式2.1で示されるように，これらの信号を足し合わせることにより，目的方向の信号を強調することができる．この場合，θ 以外から到来する音波については，位相がずれているため，信号を足し合わせても強調される程度が少ない．結果的に，目的方向 θ_s に対して強い指向特性を形成できる．

  $$s(t) = \sum_{i=1}^{M} x_i(t - D_i) \tag{2.1}$$

  遅延和型マイクロホンアレーの特徴は，素子係数 (D_i) により指向特性を制御するため，目的方向の制御が容易である．また，このマイクロホンアレーの指向性を表す主ローブ幅は

  $$\theta_r = sin^{-1}(v/fdM)$$

  で表すことができる．これにより，マイクロホン数やマイクロホン間距離をなるべく大きく取り，高い周波数を扱うほど，高い指向性が得られることがわかる．ただし，マイクロホン間距離が大き過ぎる (目的周波数の波長の半分以上) と，周波数の回り込みにより空間折り返し (spatial aliasing) が発生してしまう．また，信号を足しこむことにより，目的信号を強調するため，雑音も同時に拾ってしまい，サイドローブが高くなる傾向がある．

  遅延和型のマイクロホンアレーを用いた実用的レベルの研究としては，Flanaganらのアレー [35]，西らのアレー [88] 金森らのアレー [59]，東，打越らのアレー [16] が挙げられる．いずれも，サイドローブを低くするため，大規模なマイクロホンアレーとなっており，ロボットへの搭載には向かない．また，音源がアレーに近い場合は，平面波ではなく，球面波を仮定する必要があるため，計算量が大きくなるという問題もある．

- 適応型マイクロホンアレイ

適応型マイクロホンアレイは, 各収音信号に遅延を付加した後, 減算を行うことにより, 雑音の到来方向に指向特性を適応させる手法である. 雑音の到来方向を θ_N とし, マイクロホンが2本の場合を考えると, マイクロホンの遅延は,

$$\tau_N = \frac{dsin\theta_N}{v}$$

である. これを用いて, マイクロホンでの収音信号を同相化し, 減算すれば (式 2.2), θ_N 方向に死角を形成することができ, 結果として目的の信号を強調することができる. 実際には雑音方向が未知であっても, 減算出力で得られた信号のパワーが最小になるように τ_N を適応的に変化させることによって, 雑音除去が可能である. 目的信号は, 雑音と異なる方向から到達するため, 減算により, 歪みは生じるものの消去されることはなく, 結果として目的信号を取り出すことができる. 目的信号は, 減算後に補正フィルタを施し周波数特性を補償することができる.

$$s(t) = x_1(t - D_1) - x_2(t - D_2) \tag{2.2}$$

適応型マイクロホンアレイでは, 遅延和型に比べ, 高い S/N 比を得ることができ様々な手法が提案されている [39, 60, 38, 12]. 一般に, 特定の環境下では遅延和型に比べ, 高い S/N 比を得ることができるが制約も大きい. 例えば, 信号間相関が比較的高いので, フィルタの収束が遅いため, 雑音環境変化に追従することが難しい. また, 目的信号をキャンセルしないように雑音のみを観測する時間が必要である. さらに, 形成できる死角の数はマイクロホン数- 1 であることから, 音源数がマイクロホン数によって制限されてしまっている.

ビームフォーミングは, 信号処理的なアプローチにより音源の分離が可能であるが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, 背景雑音が動的に変化したり, マイクロホンアレイや音源が移動する場合では, 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.

独立成分分析によるアプローチ: 近年では, 独立成分分析 (*Independent Component Analysis*, ICA)[95, 49] を利用したアプローチも試みられている. 独立成分分析は, 1990 年代から盛んになった多変量解析の手法である. Jutten と Herault らによる研究 [57] が最初とされ, Comon により, Independent Component Analysis (ICA) という名がつけられた [31]. 理論的な体系は, 甘利や Cardoso らよって築かれた [8, 7]. 独立成分分析では, 複数の音源信号が混合されて観測された場合, 観測信号のみから音源信号を推定する技術であり, 目的音の方位・無音区間情報が不要であるという利点がある.

M 個の音源からの信号を $s_i(t)$, M 個のマイクロホンから得られる信号を $x_i(t)$ とし,

$\boldsymbol{s}(t) = (s_1(t), s_2(t), \cdots, s_M(t))$, $\boldsymbol{x}(t) = (x_1(t), x_2(t), \cdots, x_M(t))$ と定義すると,

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t)$$

と表すことができる. ここで, 行列 A の各要素 a_{ij} は i 番目の音源から j 番目のマイクロホンへの伝達関数を表す. この場合, A の逆行列 W を用いて

$$\boldsymbol{s}(t) = A^{-1}\boldsymbol{x}(t)$$

とすることにより元信号 $\boldsymbol{s}(t)$ を推定する方法である.

しかし, この逆問題では伝達関数は未知であるため観測された $\boldsymbol{x}(t)$ のみを用いて元信号 $\boldsymbol{s}(t)$ を推定する必要がある. 従って, W を直接求めることもできない. このため独立成分分析は blind separation とも呼ばれる. 独立成分分析では, この問題を元信号の独立性を利用し, 適応的に W を学習する手法を用いてこの問題を解いている. つまり, 適当な W によって,

$$\boldsymbol{y}(t) = W\boldsymbol{x}(t)$$

を求め, $\boldsymbol{y}(t)$ を元に, 下記の式によって, W を A^{-1} に収束させる.

$$W_{t+1} = W_t - \alpha H\{\boldsymbol{y}(t), W_t\}$$

ここで, α は学習のステップサイズである. Jutten と Herault[57] は行列 H の各成分は, $y_i^3 y_j$ としたが, H の選び方については, 様々な議論があり, 数多くの研究が報告されている.

しかし, ICA もビームフォーミングと同様計算コストが高く, マイクロホンと音源の数が等しいことを前提にしているため, 音源数が動的に変化する場合への適用は難しい. また, 背景雑音が動的に変化したり, マイクロホンアレイや音源が移動する場合のチューニングも難しく, ロボットへの搭載は難しい.

ロボットへ適用したアプローチ: 実際, ロボットへの適用についてもあまり行われていない. 多くのロボットは単一音源を仮定した処理を行っている [47, 26, 52]. 例えば, 早稲田大学の *Hadaly* [70] は, 頭部に設置された 2 本のマイクロホン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ても, 話者発見と定位が可能である. 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, 同時発話という状況を仮定していない. このため, 音源分離能力は備えていない.

もちろん, このような技術をロボットに適用し, 音源分離機能を備えたロボットも研究されている. 例えば, *Jijo-2* [15, 13] は, 8 チャネルのマイクロホンアレーを搭載し, 一般的なオフィスで音源の定位・分離を行い, さらに分離音の認識を行って, コマンドによる命令を認識することができる. しかし, 予め, マイクロホンアレーのパラメータの動的な制御や室内音響について, 時間を要する測定が必要であることから, 実環境への適用は難し

い. マイクロホンアレーを用いたロボットによる話者トラッキングについては西浦らの報告 [91] もあるが, やはり, 現時点ではシミュレーション環境での定位に留まっている.

東京大学の *SmartHead* [10] 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.

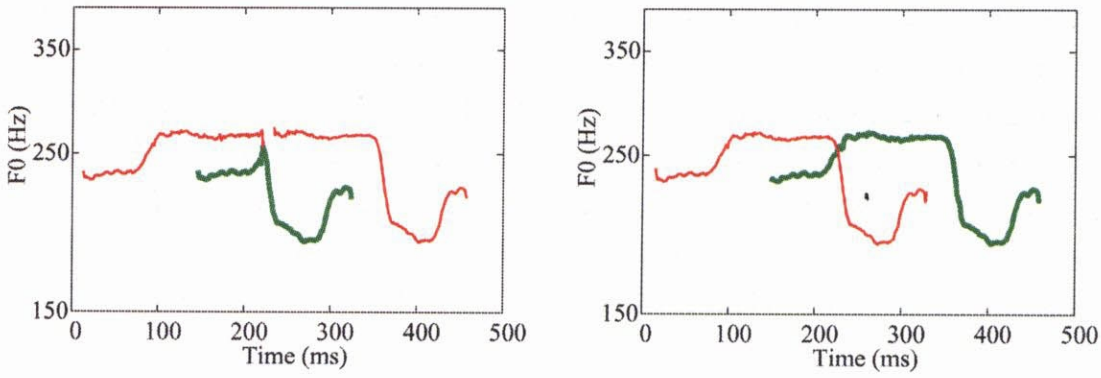
つまり, 現時点では, 残響や環境の動的変化など実環境での処理の難しさから, ロボットの音源分離機能は実用には至っていない.

■ロボットに適した音源分離 これに対して, 本研究では, アクティブ方向通過型フィルタ (*Active Direction-Pass Filter*, ADPF) を提案している. これは, 音源方向情報を利用して音源を分離抽出することができる方向通過型フィルタ (*Direction-Pass Filter*, DPF) を, アクティブオーディションに基づき実時間で動作するように, 拡張したものである. また, 本研究で, 聴覚中心窩 (*Auditory Fovea*) と命名した現象を利用して, 方向通過型フィルタの通過帯域をアクティブに制御すること, および, 対象物の方向を向くというアクティブな動作を行うことにより, 音源分離の向上を図ることができるようになっている. ここでいう, 聴覚中心窩とは, 2 本のマイクロホンによる音源定位, および方向通過型フィルタによる分離に見られ, 正面方向の感度や分離結果が最もよく, 正面から離れるに従い悪くなるという霊長類の視覚に見られる中心窩 (*Fovea*) に似た現象を指している.

2.5.4 センサ情報の統合

実環境処理では, どのような場合でも信頼できるセンサ情報を得ることは難しく, すべてのセンサはエラーを含んでいることを前提に扱う必要がある. 従って, 単一の情報 (モダリティ) の抽出精度を向上させるというアプローチで実環境処理を向上させることは難しく, 複数の情報を統合することで処理のロバスト性を確保することが必要である. これは, 複数の聴覚情報を統合して聴覚処理を向上させるという意味だけでなく, 他のセンサ情報と統合してロボットの知覚をロバストにするという意味でも重要である. 以下では, 両者について説明する.

■聴覚処理におけるセンサフュージョン 聴覚処理を向上させるために, 複数の聴覚情報を統合することは本質的である. 例として, 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における音響ストリーム分離問題を挙げてみよう. ここでは, Bi-HBSS [86] を用いて, 各時刻ごとに調



a) Ambiguity in stream separation by one microphone

b) Disambiguation in stream separation by a pair of microphones

図 2.6: 2 話者同時発話のストリーム分離問題における曖昧性 (各線は分離ストリームを示す)

波構造を抽出し, 基本周波数の時間的連続性を基にグルーピングすることによって, 混合音から, 調波構造をもつ音響ストリーム (以下, 調波構造ストリームと呼ぶ) を分離するものとする.

図 2.6 は, 同じ話者の「あいうえお」という発話を 150 ms ずらして録音したモノラル音を分離させた結果を示している. 図 2.6a) では, 一方が「あいうーーえお」, もう一方が「あいえお」と誤分離されている. つまり, 2 つの音から構成される混合音の分離を 1 本のマイクロホンで行うと, 分離結果に 2 本の音響ストリームの周波数が交差しているケース, および, 実際には交差せず周波数が近づいて離れるケースという, 2 つの可能性が存在するという曖昧性が含まれることを示している. この曖昧性は, 調波構造という単一の情報だけでは解決できず, 他の情報を用いる必要があることは明らかである.

例えば, 2 つの音が別々の方向から到達する前提の下では, 2 つのマイクロホンを用いれば, 方向情報が利用できるので, 調波構造ストリーム分離での曖昧性は解消できる. 図 2.6b) は, HRTF を畳み込むことによって, 150 ms ずらした同じ発話が左右 30 ° から演奏したように聞こえるように生成したステレオ音に対し, Bi-HBSS で分離した場合の結果を示す. マイクロホン 1 本の時には, 調波構造のグルーピングが間違ってしまうのに対して, マイクロホン 2 本を使用した分離では, 方向情報が利用できるので, 調波構造だけを手がかりとした分離では解決できない曖昧性が解消されている. つまり, Bi-HBSS は倍音構造だけではなく, HRTF よって得られた音源方向を併せて利用することによって, 音響ストリーム分離の曖昧性を解消している.

このような聴覚情報の統合手法については, ノンパラメトリックカルマンフィルタを用いた手法が提案されている [3]. 安部らは, この手法に基づき, 周波数軸方向の統合については, 観測値群からその発生系の確率密度を推定することができる Parzen 推定 [36] を利

用した投票法を用い [2], 時間軸方向の統合については, ストリームのパラメータが既知のダイナミクスに従って緩やかに変動するという仮定の下でノンパラメトリックカルマンフィルタを用いることにより混合音のストリーム分離問題を定式化している [3]. さらに, 彼らはノンパラメトリックカルマンフィルタを用いた統合手法を汎用的な情報統合手法として提案しており, 聴覚処理に限らず, 画像処理など様々な分野に応用しその有効性を示している [66].

■視聴覚処理の統合 ロボットが "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ためには, 聴覚のみでは限界がある. 人間でも定位誤差は, 数度から十数度といわれており [20, 29], 決して聴覚の精度が高いわけではなく, 状況の把握に関しては他の感覚情報を利用している度合いが強い. 実際に人間が視聴覚を統合して処理を行っていることを示す例としてマガーク効果 (*McGurk Effect*) [75] が挙げられる. これは, 人間の聴覚に見られる錯覚の一種であり, 「ば」と発音しながら, 「が」と発音した場合の映像を見ると「だ」もしくは「が」など「ば」とは違う音として認識してしまうというものである. 聴覚による認識が視覚認識の助けを借りる場合があることを示すよい例である.

ロボットにおいても, 状況に即した行動を自律的に行うためには, 様々なセンサ情報を統合して, 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 (Cocktail Party Effect)[30] 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が求められる. また, 複数の顔を識別する能力も必要である. 従って, 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に最低限必要な機能といえる [25].

例えば, 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, 上述の Bi-HBSS を拡張している [81]. 具体的には, 音響ストリームを抽出する際に, HRTF によって得られる音源方向だけではなく, 画像処理による正確な方向情報を利用して音源分離の精度を向上させた. HRTF を畳み込んで作成したステレオ音に対して, 3 話者同時発話の分離を行い, 分離の正確さを音声認識システムを用いて示している.

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 [1]. 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, 音源定位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 [126]. 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 [43], 池田らの相互情報量を計算する際に時間窓長を適応させる視聴覚統合の研究 [48] でも視聴覚統合によるロバスト性の向上が示されている.

ロボットでは, このようなマルチモーダルな統合によってロバスト性を向上させる研究は, センサフュージョンの分野で盛んに行われている. ロボットにおける視聴覚統合という点では, 先にあげた理由から, 超音波センサを除いて, 聴覚はあまり統合の対象とされていない. しかし, 音声認識の分野では, 音声認識とリップリーディングを組み合わせて音声認識を向上させる研究として, 視聴覚統合が盛んに行われており, その重要性は明らかである.

以上より, 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, ロバストなシステムを構築するための重要な鍵であるといえる.

■ストリームベースの視聴覚統合 そこで, 本研究では, 時間の流れを考慮したストリームを用いた視聴覚統合を行う. この手法では, 視覚・聴覚情報はすべて, ストリームを用いた表現として扱うことにより, 視覚・聴覚情報を透過的に扱うことができる.

安藤らの提案しているノンパラメトリックかルマンフィルタなど, 数学的・情報理論的アプローチも情報統合を行う有効な手段であるが, モダリティの追加性, スケールアップ時の様々な例外処理やノイズを扱いやすくするため, 本研究ではシンボリックなストリーム表現を用いたシンボリックな統合を行う. 詳細は 6 章で述べる.

2.6 ロボット聴覚の応用 – より自然なインタラクションの実現

これまで述べてきたような課題を克服し, ロボット聴覚を構築することによって, より "自然な" ヒューマンロボットインタラクションが期待できる. ここで, "自然な" とは, 本研究では, 以下の 4 項目を実現するものと捉えるものとする.

1. ロボットに備わったマイクロホンを用いた音声によるインタラクション
2. 複数の人物 (音源) が同時に存在する場合のインタラクション
3. 音源やマイクロホンの位置が動的に変化する場合のインタラクション
4. 積極的な動作によるフレンドリなインタラクション

また, 本論文のタイトルに用いられている“自然な”という言葉も同義とする. 具体的には, ロボット聴覚をソーシャルインタラクションに応用した例として, 同時発話の音声認識によるインタラクション, およびパーソナリティの導入したロボットによるインタラクションについて検討する.

一般に複数の音が混在している環境下で音声認識を行うことは難しい. そのため, 本研究では, ロボット聴覚の混合音を扱う音源分離処理を音声認識のフロントエンド処理に利用する. これにより, 音声認識の精度を向上させ, より自然なインタラクションを行うことを可能とした. これについては, 8 章で, その手法および応用例を紹介する.

自然なソーシャルインタラクションを実現する上で, 複数のストリームが存在する場合, どのストリームに注意をむけるべきか, という選択的注意も, 重要な課題である. エンターテイメントロボットは, 購入後すぐに飽きてしまうという声もしばしば聞かれるが, これは, 単純な動作しかできないというだけでなく, ロボットが明確なパーソナリティを持っていないことによるところも大きいと考えられる. そこで, 人間では, 選択的注意は, 人それぞれのパーソナリティによって異なるという事実に習い, ロボットにパーソナリティを導入した選択的注意制御機構を実装し, より自然なヒューマンロボットインタラクションを目指す試みを 9 章で紹介する.

2.7 まとめ

本章では, ロボット聴覚に対して, 2 本のマイクロホンの必要性について述べ, 実現可能な機能について解説をした.

ロボットで実環境ロボット聴覚を実現するためには, 体を動かして聴くというアクティブオーディションが鍵となることを述べ, これを実現するには, 以下の課題を考慮することが必要であることを論じた.

1. ロボット自身の発生する音の抑制機構, ロボットが動きながら音を聞く機構
2. 頭部伝達関数を使用しない音源定位・音源分離など, 特定の環境に依存しない音の知覚
3. 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)
4. 画像処理や他の処理との情報統合

最初の課題については, 外装の内部と外部にマイクロホンを備えることにより, 動作時のノイズをキャンセルするアプローチを述べた. この詳細は 4 章で述べる. 2 番目の課題については, HRTF を用いずに音源定位を行う手法を述べた. これは, 4, 5 章で詳細に述べる. また, 3 番目の課題については, アクティブ方向通過型フィルタを提案し, 7 章を中心に議論を行う. 最後の課題については, 視聴覚情報を統合してロバストに人物を追跡す

るシステムを6章で述べる.

さらに, ロボット聴覚の応用として, より自然なヒューマンロボットインタラクションを目指すために, 分離音の音声認識について, 8章で述べる. また, パーソナリティを導入して, 人間をより豊かなインタラクションに誘うようなヒューマンロボットインタフェースを構築する試みを9章で述べ, これらの有効性を示す.

次章では, ロボット聴覚の課題を考慮したロボット聴覚システムの全体像を述べる.

第3章

ロボット聴覚システム

本章では, 前章で検討した, 課題を考慮して, アクティブオーディションを利用したロボット聴覚システムの概要, および, システムの有効性を確かめるため, 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ロボット *SIG* について述べる.

3.1 ロボット聴覚システム

第2章で検討した課題を考慮して, アクティブオーディションに基づいたロボット聴覚システムの概要を図3.1に示す. このシステムは, 音源が複数ありかつ動作している場合でも, ヒューマノイドロボット (*SIG*) のカメラ, マイク入力から, 動作時のノイズをキャンセルし, ロボット自身のアクティブな動作, 視聴覚の統合による効果を用い, これらを定位・分離することが可能である. また, 分離音の音声認識機能もシステムのアプリケーションとして付加できるように構成されている. 本システムは大きく3つのサブシステム「動作時のノイズキャンセル」, 「視聴覚を統合した実時間複数人物追跡」, 「アクティブ方向通過型フィルタ (Active Direction-Pass Filter, ADPF) による音源分離」からなる. 以下にそれぞれを説明する.

3.1.1 動作時のノイズキャンセル

動作時のノイズキャンセルは, 4本のマイクロホンを用いて行っている. 基本的には, 外装の頭内部に設置したマイクロホンと外装の耳の位置に設置したマイクロホンの差を用いることによって, 動作時に問題となるバーストノイズに注目し, これをキャンセルする手法を用いている. バーストノイズのキャンセルは, 音源定位で大きな効果を及ぼすが, 音源分離では, バーストノイズキャンセルにより, 音響情報を失う可能性があるため, ノイズキャンセルを行わず, 左右のマイクロホンで収音したデータをそのまま用いている. 音源

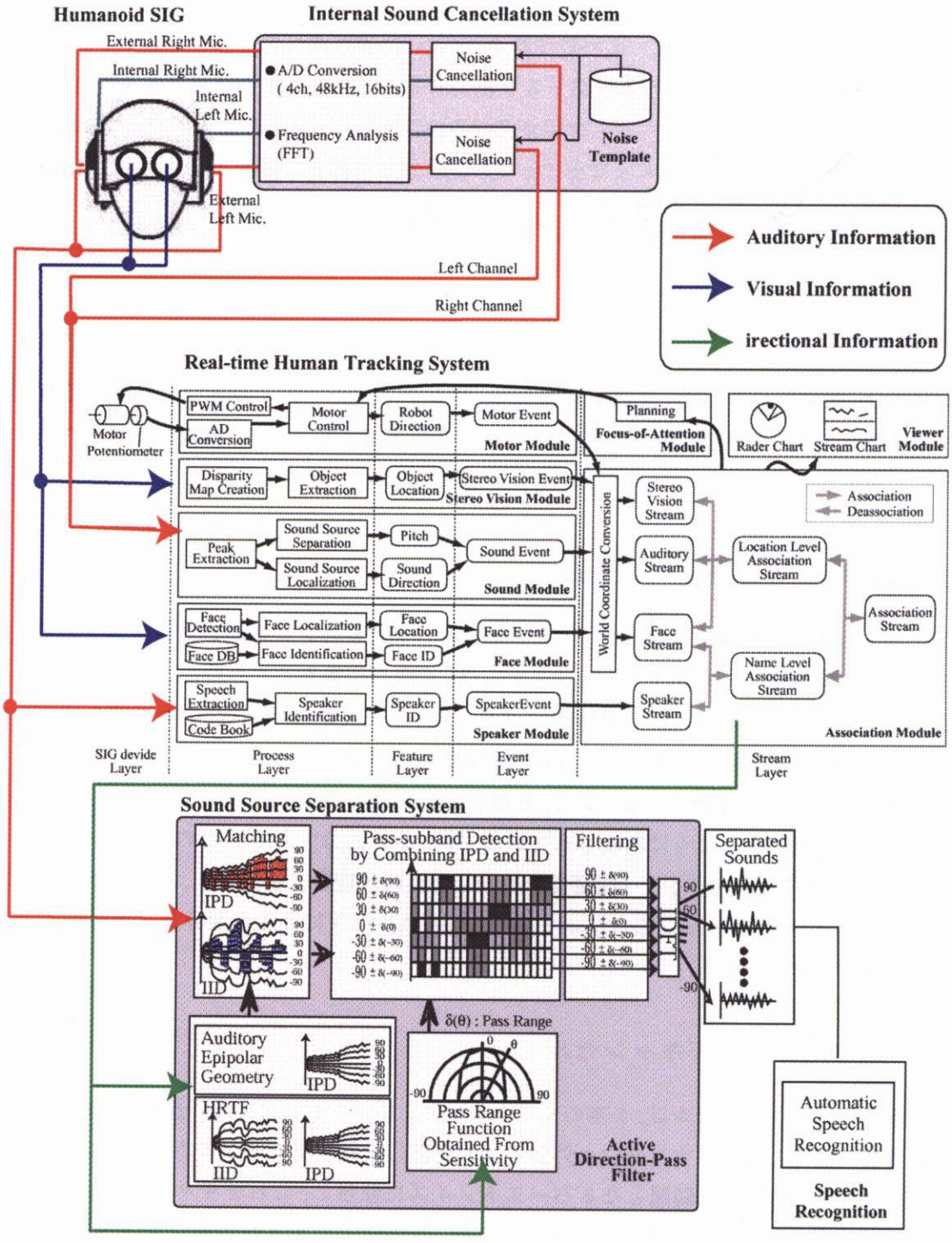


図 3.1: ロボット聴覚システムの構成図

分離部では, 音源の方向情報を利用できるので, その際にノイズがキャンセルされ, 信号対ノイズ (S/N) 比の高い分離音を提供することができる.

3.1.2 視聴覚を統合した実時間複数人物追跡

実時間複数人物追跡は, *SIG* のカメラやマイクから得られるセンサ情報を統合して, 複数人物の位置を把握し, 追跡する. 本論文では, このシステムで認識される音源方向情報を音源分離部への入力としている. 具体的には, 音源定位, 顔認識・定位, ステレオビジョン, モータ制御といったモジュールが位置, 名前情報などの特徴を抽出し, 時間情報を付加して, 特徴の種類に拠らない透過的な表現であるイベントに変換する. これらはストリーム生成部で, イベント種類ごとに時間方向に接続され, ストリームが形成される. 複数ストリームが同じ人物に由来するものと判断されると, これらを一つに束ねアソシエーションストリームを生成することにより, 視聴覚の統合を行う. この際, 音源分離部には音情報を含むストリームの方向情報が送出される. さらに, ストリームの状態に応じて注意対象に *SIG* の体を向け, 対象音源方向を向くというアクティブな動作も可能である. 注意制御については, ロボットのパーソナリティを考慮した制御も可能なように実装されている. ロボットのパーソナリティについては, ロボット聴覚システムの応用として 9 章で述べる. また, システムは, 分散処理と Kalman フィルタによる予測により実時間動作が可能である.

3.1.3 アクティブ方向通過型フィルタによる音源分離

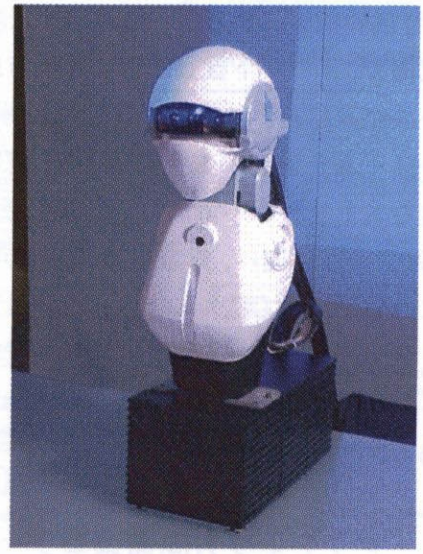
音源分離には, 特定の方向の音響信号を抽出することができるアクティブ方向通過型フィルタを用いる. フィルタの入力は, 左右チャネルのスペクトル, スペクトルから計算される両耳間位相差 (IPD) と両耳間強度差 (IID), および, 実時間人物追跡システムから得られる音源方向情報の 4 つである. フィルタの出力は, 入力方向に対する分離音響信号である. アクティブ方向通過型フィルタでは, 入力の音源定位情報から, 周波数領域で, 周波数, 音源方向毎に IPD, IID に関する仮説 $H(f,\theta)$ を生成する. 仮説の生成には, ロボットの伝達関数を利用している. 仮説と入力 $I(f)$ を照合し, $|H(f,\theta) - I(f)| \leq \delta(f,\theta)$ を満たすサブバンドを抽出し, これを音響信号に逆変換し音源の分離を行う. ここで, $\delta(f,\theta)$ は, 照合の是非を決定する閾値であり, アクティブ方向通過型フィルタの通過幅にあたる. この通過幅は, 聴覚中心窩に基づいて正面方向では狭く, 周辺部では広くなるように定めた. 聴覚中心窩とは, ロボットの正面の音源定位精度が周辺部に対して高いことから, 我々が視覚の中心窩になぞらえて呼んでいる現象である. この通過帯域制御により, 方向通過型フィルタでは難しかった, 実環境での高速な音源抽出を可能にしている. また, 分離音に

対する音声認識も可能になっている．音声認識は，複数の音響モデルを使用した方法に基づいている．話者や方向ごとにトレーニングされた音響モデルの結果を統合し，精度の高い分離音の孤立単語認識による音声認識を行う．

3.2 ヒューマノイド SIG

ロボット聴覚を実現するためのテストベッドとして，図 3.2 に示すヒューマノイドロボット *SIG* (以後，単に *SIG*) を独自に開発し，利用している．

a) body (upper torso)



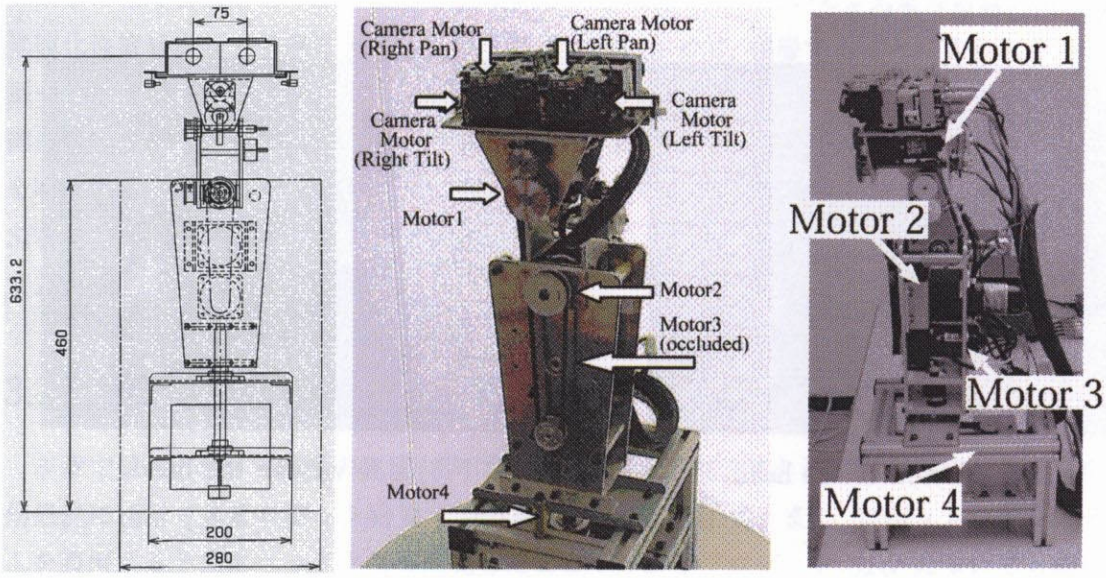
b) body and pedestal

図 3.2: Humanoid *SIG*

図 3.3 に示されるように，*SIG* は，4 自由度を有した上半身ロボットとして設計されている．各アクチュエータには，DC モータを使用しており配置は以下の通りである．

- 頭の左右振り用のモータ (Motor 1)
  稼動範囲は ±30°
- 頭の縦振り用のモータ (Motor 2)
  稼動範囲は ±30°
- 首の左右振り用のモータ (Motor 3)
  稼動範囲は ±15°
- 胴体の回転用のモータ (Motor 4)
  稼動範囲は ±45°

Motor 1 と Motor 3 は, 動作方向が同じであり, 組み合わせて使用すると歌舞伎の勧進帳のように頭部のみを横にスライドするような動きが可能である. また, 各モータには, ポテンショメータが取り付けられており, これによって, 速度・位置制御が可能である. また, これらのモータは, 日立の SH を用いたボード AP-SH2F-0A*1を介して, コンピュータからシリアル通信で制御できるようになっている. なお, 本研究で, 実際にアクティブに制御を行うモータは, ロボットの水平方向の回転用のモータ (Motor4) のみである.



a) mechanical design　　b) front view　　c) side view

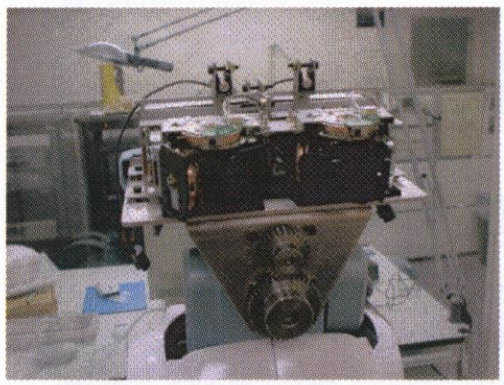
図 3.3: *SIG* の機構図

視覚センサには, 図 3.4 に示されるように頭部に, 2 個の CCD カメラ (Sony EVI-G20) を搭載している. このカメラは, オートフォーカス機能が付いており, シリアル経由で, 独立にパン, 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, カメラ間のベースラインは 7.8cm である. ただし, 本研究では, カメラの制御は基本的に行わず, 常にカメラが, *SIG* の正面方向を向くように固定して使用している. カメラの画像出力は NTSC であり, 一般的なコンピュータ用のキャプチャカードで取り込みができるようになっている. また, ステレオ画像については, 左右のカメラの同期を取るため, Sony Quad Switcher YS-Q440 を通して, コンピュータに取り込むようになっている.

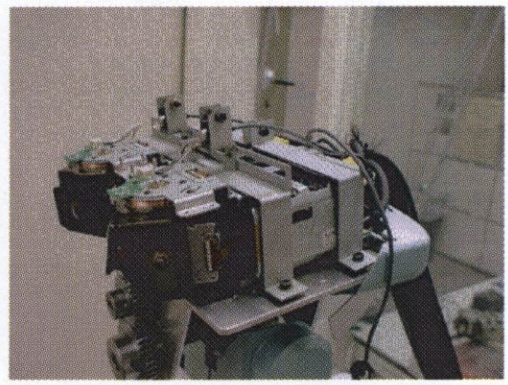
聴覚センサには 図 3.5 に示すように, 2 組, 計 4 本の無指向性のマイクロホン (Sony ECM-77S) を用いている. 一組のマイクロホンは, 図 3.5b) に示されるように, 耳の位置

*1 http://www.apnet.co.jp

に設置されており, *SIG* 外部からの音を集めるためのものである. このため, 内部からの音をできるだけ拾わないように外装内部とは隔離されている. 外部マイクロホンのマイク間距離は直線で 18.0 cm であり, カメラのベースラインと平行になるように設置されている. また, マイクロホン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. 図 3.5a) に示される外装内部にあるもう一組のマイクは, 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. 主に, *SIG* のモータノイズなど内部音の収音用に使用している.

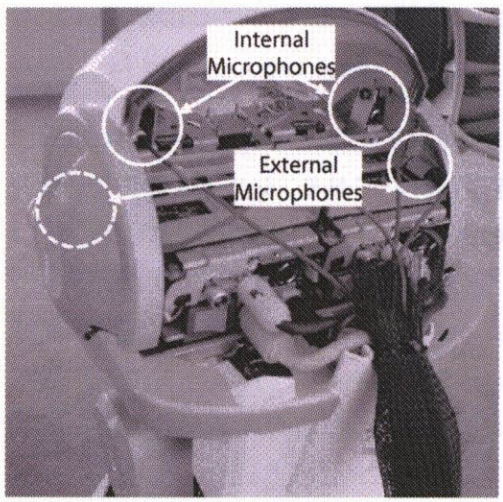


a) front view of the head

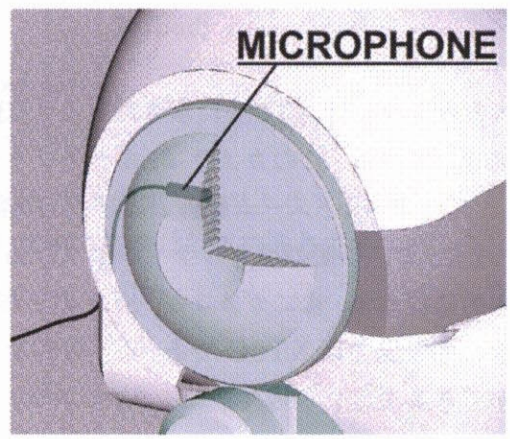


b) side view of the head

図 3.4: *SIG* のカメラ



a) internal and external microphones



b) external microphone and the earlobe

図 3.5: *SIG* のマイクロホン

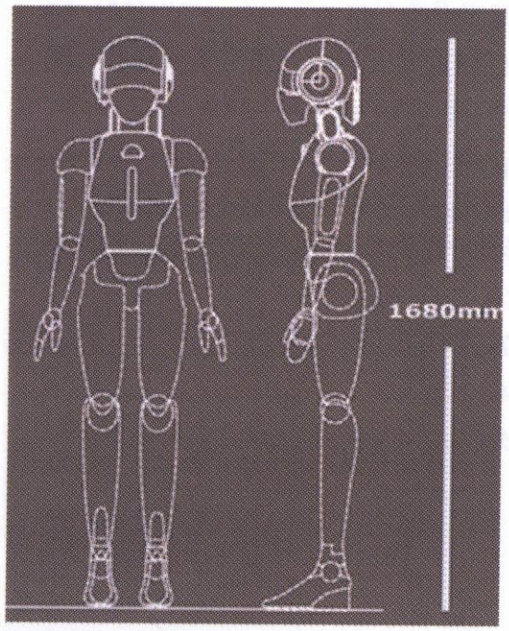
図 3.2 に示す外装は, *SIG* 内部の音が外部に漏れるのを防ぎ, 音響信号処理の複雑性

を緩和するための, 音響的な身体性の獲得や内部機構を, 外的な要因から保護するだけではなく, インダストリアルデザインを考慮し, 審美性の追求も行っている. そのため, 作成に際しては, 図 3.6a) に示すような 8 等身のフルボディをデザインするところから始まり, 最終的には CAD ソフトを用いて, 図 3.6b) のようなイメージを作成した. CAD ソフトは, 3 次元的な表現が簡単であるだけでなく, データが電子化されているために, Rapid Prototyping にも有効である. *SIG* では, 高速に, かつ正確にプロトタイプを作ることができる方法として近年注目されている光造形法を用いている. 光造形とは, 液状の光硬化性樹脂をレーザ等の光ビームで一層ずつ硬化させて, 積層することにより成形用の型や切削工具等を用いずにプラスチックの 3 次元立体物を精度良く作成する手法である. 図 3.7a) は光造形法によって作成された *SIG* 外装の各パーツである. このパーツから型を取ることによって, 図 3.7b) に示すようなプロトタイプを簡単に作成することができる. *SIG* の場合では, 設計終了後, 研磨や塗装を含めて 3 週間ほどで, 外装を作成することができた[*2].

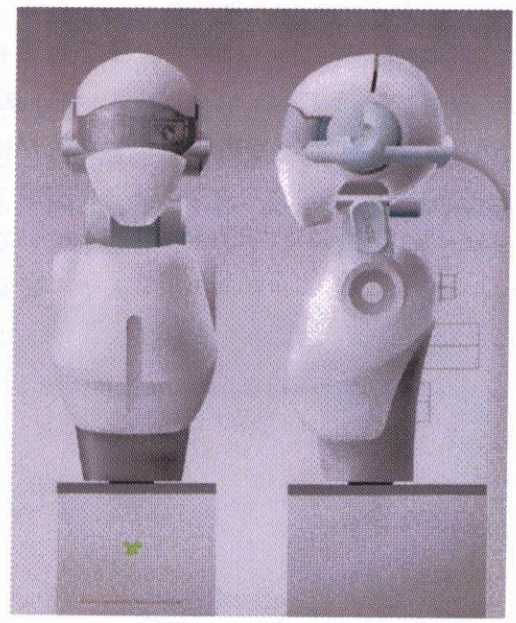
3.3 まとめ

本章では, ロボット聴覚の実現に向けた課題を考慮して, アクティブオーディションを効果的に利用できるロボット聴覚システムの全体像を示した. また, ロボット聴覚システムを適用するテストベッドとして開発したヒューマノイドロボット *SIG* についても紹介し, その諸元, 製作過程を説明した. 次章以降では, ロボット聴覚システムを構成する各技術について詳細に解説を行う.

[*2] 協力:株式会社インクス (http://www.incs.co.jp)

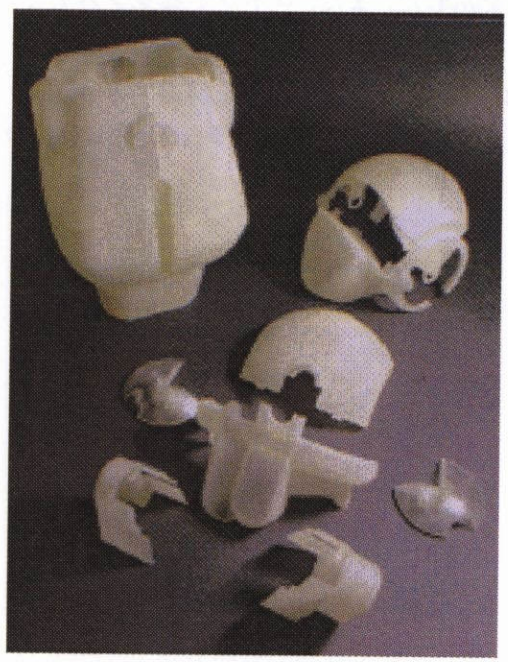


a) full body image of *SIG*

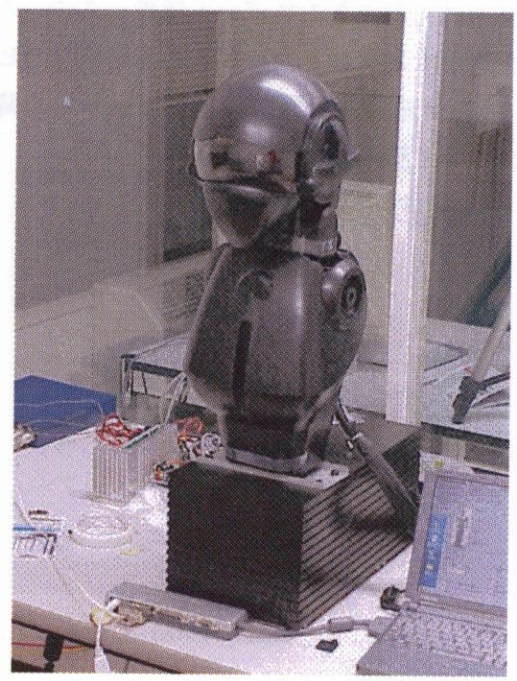


b) final image of the cover

図 3.6: *SIG* の外装のデザイン



a) cover components of *SIG*



b) prototype of *SIG*

図 3.7: 光造形によるラピッドプロトタイピング