論文内容の要旨

論文題目　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

氏名　　中臺一博

本論文では, 将来, ロボットが人間と共生する上で重要なインタフェースであるロボット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に注目し, 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し, その工学的な実現, ヒューマン・ロボットインタラクションへの応用を通じて, 有効性を明らかにする.

ロボットを対象とした知覚機能の研究のうち, 聴覚は, 人間とのソーシャルインタラクションで最も重要な機能の一つであるにもかかわらず, 視覚研究と比較し, あまり盛んではない. また, 実環境・実時間でロボット聴覚を実現するため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, 本論文では, まず, 新たにロボット聴覚研究を定義し, ロボット聴覚機能の課題を体系的に整理した. 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, 知覚を向上するアクティブパーセプションがロボット聴覚の向上にも本質的であると捉え, アクティブな動作を利用し, 聴覚情景分析を向上させる枠組みとしてアクティブオーディションを提案した. さらに, これを実現するための様々な課題の中から, (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という 4 つの課題に取り組み, 図 1 に示すロボット聴覚システムを構築した. 構築したシステムは, 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」から構成されており, 複数の音源が存在し, 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, 視聴覚統合を利用して, これらを定位・分離・認識することが可能である.

動作時のノイズキャンセル部は, 課題 (1) に対応し, 外装によって, ロボットに音響的な身体性を構築し, 内部音抑制を行う. 具体的には, まず, 音響的に隔離されたロボットの頭内部に一組, 外装の耳位置に一組の, 計 4 本のマイクロホンを設置する. 次に, 外装の音響測定結果をテンプレートとして利用し, ヒューリステックルールにより, 動作時に最も問題となるバーストノイズのみをキャンセルするフィルタを構築した. これにより, 信号処理的なノイズキャンセル手法に見受けられる位相情報の歪みを伴わない新しいノイズキャンセルを実現し, ロボット動作時に生じるノイズのため, 聞くために一旦停止しなければならいという *"stop-perceive-act"* 原理を緩和した.

実時間複数人物追跡部は, 課題 (2),(4) に対応する. 「音源定位」, 「顔認識・定位」, 「話者同定」, 「ステレオビジョン」, 「モータ制御」, 「アソシエーション」, 「アテンション制御」, 「ビューワ」 の 8 モジュールから構成され, マイクロホンとカメラから得られる視聴覚情報を統合し, 複数人物の定位・追跡が可能である.

「音源定位」では, 未知環境でも 2 本のマイクロホンで音源定位を可能にする聴覚エピポーラ幾何を提案し

た. 両耳聴の研究では, 頭部伝達関数 (Head-Related Transfer Function, HRTF) から導出される両耳間位相差 (IPD) や両耳間強度差 (IID) を用いて音源定位を行うことが一般的である. HRTF は, 通常, 無響室で, 各方向からのインパルス応答測定によって取得する頭部形状の音響特性を表す伝達関数である. しかし, 環境が変わる毎に再測定が必要であり, 離散的な関数であるため連続的な定位が難しいため, ロボットへの搭載には適していない. 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, 測定が不要である. このため, 高速に連続的な定位が可能であり, ロボットに搭載し, 音源追跡を可能にした. さらに, より一般的な環境でのロバストな動作を目指し, IPD, IID, 調波構造といった複数の聴覚的な手がかりを Dempster-Shafer 理論を用いて統合するモデルを提案した. 部分的に歪んだ音響信号の定位や 4 音源の同時定位を通じて, 聴覚エピポーラ幾何の有効性, 聴覚情報統合の有効性を示した.

「音源定位」以外にも, 「顔認識・定位」, 「ステレオビジョン」といったモジュールからは位置や名前情報が抽出される. 「アソシエーション」では, これらの情報の時間の流れを考慮し, その種類ごとにストリームを形成する. 視聴覚統合は, ストリームベースのシンボリックな統合手法であり, 同じ人物に由来する複数のストリームを一つに束ねたアソシエーションストリームの生成により行われる. 提案した統合法は理論的に最適性を保障するわけではないが, 人物追跡や視聴覚情報の曖昧性の相互に解消できることを示し, 実環境で十分有効であることを示した. また, 様々な情報を階層的に統合し, スケーラビリティの高い実装が容易な統合手法であることを示した.

音源分離部は課題 (3) に対応する. 日常, 耳にする音は複数の音源からの音が混じった混合音であることから, 音源分離は一般的音理解で重要な機能である. 本論文では, 音声認識の前処理として実時間・実環境で使用することができる分離能力を目指し, 特定方向の音響信号を抽出する ADPF を提案した. ADPF は, 音源定位情報を入力とし, 周波数領域でのサブバンドセレクションにより, 高速でマイクロホンの数以上の分離能力を有する. ロボット正面の音源定位精度は周辺部に対して高いという聴覚中心窩ともいうべき現象を示すことから, ADPF は, 正面方向では狭く, 周辺部では広くなるような通過幅制御を行う. 反響のある部屋で, 3 話者同時発話に対して 9 dB 程度のノイズ除去率を示した. また, ADPF の通過幅やロボット方向のアクティブな制御が音源分離を向上させることから, 音源分離におけるアクティブオーディションの有効性を示した.

システムの応用として, “自然な” ヒューマンロボットインタラクションの実現についても扱った. ロボットに備わったマイクロホンを用い, 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であっても積極的な動作を行って, フレンドリな音声によるインタラクションを行うことが “自然な” インタラクションであると定義し, 同時発話の孤立単語認識, およびパーソナリティを導入した選択的注意制御によるインタラクションを行った.

一般に特定音声を完全に分離することは難しく, 信号の歪みやノイズの混入が避けられない. そこで分離音声の孤立単語認識では, 音源方向, 話者ごとに音響モデルから得られる複数の認識結果と顔認識から得られた名前情報に対し, 確率ベースの統合を行う手法を提案した. これにより, 同時 3 話者発話の音声認識が可能であることを示し, 音声認識におけるアクティブオーディション, 視聴覚統合の有効性を示した.

パーソナリティを導入した注意制御では, 心理学で用いられるインターパーソナル理論を用い, friendly, dominant, hostile といったパーソナリティを受付やコンパニオンといったケースに適用した. 人間とのインタラクションは, 非言語でパッシブなインタラクションであっても, 話者方向を向くことによって, フレンドリなインタラクションや, 人々の興味を喚起させるという意味で重要であることを示した.

本論文には, 3 点の意義がある. 1 点目として, ロボット聴覚という研究の確立である. 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, 定位・分離・認識や, カクテルパーティ効果として知られるような選択的注意は一般的であるが, 従来の聴覚処理では定位・分離・認識に対して広範囲な研究が行われてきたに

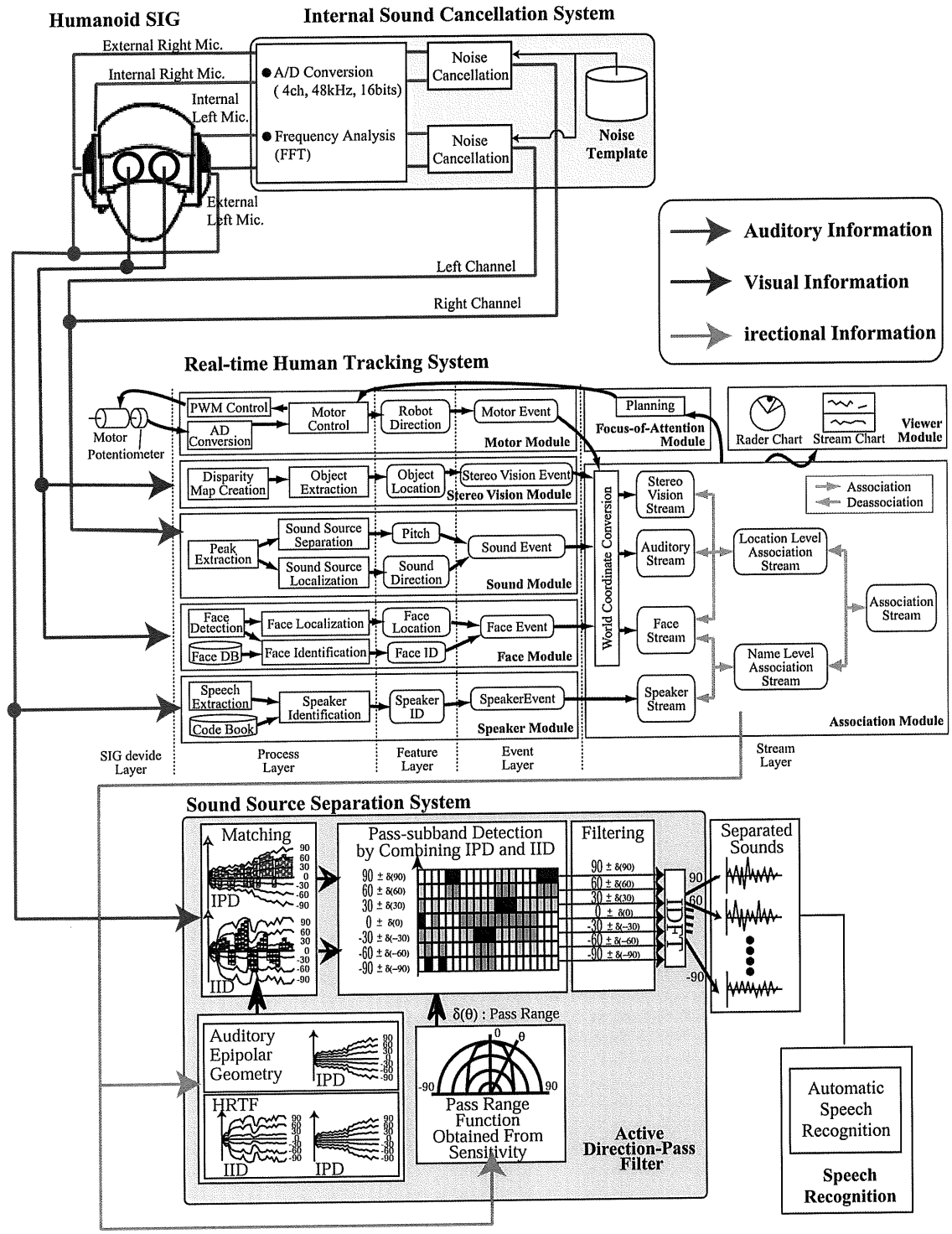

図 1: ロボット聴覚システムの構成図

もかかわらず, 音源やマイクの動作を前提とした研究は, 明示的に行われていなかった. そこで, 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新たな研究分野として定義し, その課題を明確にした.

2 点目は, 応用的な観点として, ロボット聴覚システムを工学的に実装し, 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では, 混合音や自身が発生するノイズに対する考慮が不十分であり, 実環境で聴覚による自然なヒューマン・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.

3 点目は, アクティブな動作を伴った聴覚処理のモデル化とその評価である. 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, その評価は難しかった. 本論文では, 動作を伴う音源定位・分離・認識のモデル化を行い, 2 本のマイクロホンを備えたロボットを用いた評価を通じて, アクティブな動作がロボット聴覚向上に本質的であることを示した.

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, 動作が可能な対象であれば, ロボット以外の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.