知 能 と 複 雑 系　126－7
（２００１．１０．１２）

アクティブな視聴覚統合を用いた実時間人物追跡ヒューマノイド SIG

中 臺 一 博[†]　日 台 健 一[†]
奥 乃　博[†,††]　北 野 宏 明[†,†††]

本稿では, ロボットを対象にロバストな知覚機構を実現するため, 聴覚処理の面から, アクティブな動作, および, 視聴覚情報統合の有効性を示す. 一般に, 実環境では, 反響, 雑音が存在し, かつ動的に変化するため, 高精度な音源定位や音源分離を行うことは難しい. 我々は, 音源定位に関しては, アクティブな動作と視聴覚処理を統合し, 実時間で複数人物 (音源) の定位・追跡が可能なシステムを開発している. そこで, システムから得られる方向情報を利用して, アクティブ方向通過型フィルタによる音源分離のプロトタイプを実装した. 結果として, 一般的な部屋において, 音源に正対するようなアクティブな動きにより 2 話者の分離や, 移動音源の抽出に対してアクティブな動作と視聴覚情報統合の有効性を示した.

Active Audio-Visual Integration in Real-Time Human Tracking Humanoid SIG

Kazuhiro Nakadai,[†] Ken-ichi Hidai,[†] Hiroshi G. Okuno[†,††]
and Hiroaki Kitano[†,†††]

This paper describes improvement of auditory processing by active motion and audio-visual integration. Generally, environmental noises and reverberation affect sound source localization and separation in the real world badly. Our real-time human tracking system for humanoid robots attained robust sound source licalization in the real world by active audio-visual integration. Then, we propose a new sound source separation method by active direction pass filter. Our experiments proves that active audio-visual integration is essential to robust perception for extraction of tracking sound source.

1. はじめに

我々は, ロボットを対象に, ロバストな知覚処理の実現に向け研究を行ってきた.

これまでに, 動きながら音源定位を行う能力を持つアクティブ聴覚 (Active Audition) と複数顔認識システムを実時間で統合し, 上半身のヒューマノイド SIG 上にロバストで精度の高い人物追跡を実現した[8]. 人物追跡システムでは, 常に音源に正対するよう向きを変えるなどアクティブな動作と聴覚, 視覚といったセンサ情報を統合することにより, 聴覚だけでは難しい精度の高い音源方向を得ることができる. また一般に視野が狭い、オクルージョンが存在するといった視覚情報の欠点を補うことも可能である.

しかし, 相手の方向を向くだけでなく, 人間が行うような傾聴や注視を実現しようとすると, よりアクティブで主体的な行動が必要である. また, 傾聴や注視において、注意が移るきっかけは各々の感覚情報だけでなく、環境から得られる複数の感覚情報に基づいたマルチモーダル情報の影響が大きい.

従って, 実際に, 傾聴や注視を実現するためには, 能動的な行動とマルチモーダルな情報統合が必須である. さらに, 話者同定, 音声認識, 物体認識による名前や意味レベルの情報を 取得することも必要となる. このような高次の情報は, 傾聴や注視のためにも必要であるが, 低次の情報と階層的に統合することにより知覚処理における曖昧性の解消を期待することができる点でも有効であろう.

しかしながら, 聴覚処理では 一般に混合音が入力となるため, 音声認識や話者同定により, 高次の情報を得るためには, それらのフロントエンドとして精度のよい音源分離が求められる.

音源分離問題は一般に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. そのため, 従来, 音楽を対象とした音源分離システム[5] 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[9]

† 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
†† 京都大学大学院情報学研究科
Kyoto University
††† ソニーコンピュータサイエンス研究所株式会社
Sony Computer Science Laboratories, Inc.

など心理学的知見から得られたヒューリスティックを利用した音源分離が取り組まれてきた．しかし，これらは，シミュレーション環境で実験を行っており，かつ実時間処理は困難であった．

また，近年では，残響まで考慮した独立成分分析 (ICA) による音源分離[4] なども行われている．しかし，これらはロボット自身の動作により発生する雑音や環境の変化が考慮されていないため，そのままロボットへ適用することは難しい．

一方，聴覚機能を備えたロボットの研究としては，MIT AI Lab の *Kismet*[1] や，早稲田大学の *ROBITA*[6] などが挙げられる．これらは，対話などソーシャルインタラクションに焦点を当てており，音源分離は行われていない．そのため，音声認識は各話者の口元に取り付けられたマイクを利用しなければならないという制約がある．

そこで本稿では，ロボットにおける音源分離機能を実現することを目標に，ロボットに適用可能なアクティブ方向通過型フィルタを使用した音源分離法を提案する．

以下，2 章では，本稿で使用したヒューマノイドについて述べ，3 章では，実時間人物追跡システム，および方向通過型フィルタによる音源分離について説明する．4 章でシステムの実験と評価を行い，5 章でまとめる．

2. ヒューマノイド *SIG*

図 1 Humanoid *SIG*

研究のテストベッドとして，Fig. 1 の上半身ヒューマノイド *SIG* を使用している． FRP 製の外装は，音響的にロボットの内外を区別できるよう設計されており，カメラには，一組の CCD カメラ (Sony EVI-G20) を，マイクには，計 4 本の無指向性マイク (Sony ECM-77S) を使用している．一対は，外界からの音響信号を収音するよう *SIG* の耳の位置に，もう一対はモータによって発生する内部ノイズをキャンセルするよう外装内部に配置されている．また，*SIG* は，4 自由度を有し，各モータには，ポテンショメータによって位置，速度の制御が可能な DC モータを用いている．

3. 実時間複数人物追跡システム

システムの構成を Fig. 2 に示す．システムは，“音源分離・定位”，“顔抽出・認識”，“ステレオビジョン”，“アソシエーション”，“アテンション制御”，“モータ制御”，“ビューワ” の大きく 7 つのモジュールから構成されている．実装上は，7 つのモジュールを LAN で接続された 4 台の Pentium III ベースの Linux ノードに分散させている，各ノードはギガビット，ファストイーサの 2 つのインタフェースを備え，モジュール間で非同期に発生する通信は，頻繁でかつ大きいため，ギガビットイーサを用い，同期信号などの，通信はファストイーサを使用している．結果として，ノード間の同期は 100 マイクロ秒以下の精度で、またレイテンシは 500ms でリアルタイム処理を実現している．

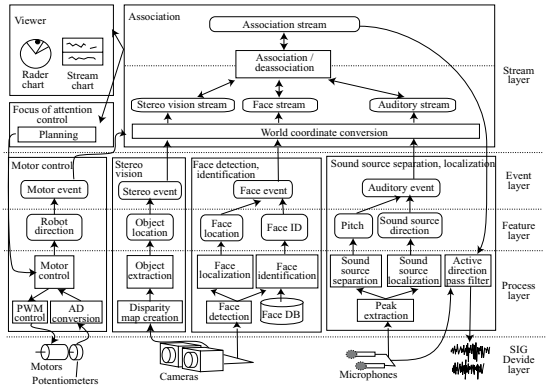

図 2 システムのモジュールとその階層構造

モジュール内のサブモジュールや情報は，5 つの階層に分けられている．*SIG* デバイス層は，*SIG* が備えているカメラ，マイク，モータシステムなどのセンサデバイスを指す．これらのセンサから得られたローレベルデータがプロセス層へ入力され，位置，名前情報といった特徴として特徴層に出力される．各特徴は，抽出されたタイミングでモジュール単位にまとめられ，イベント層に出力される．イベント発生のタイミングは，非同期であり，各モジュールごとに異なる．ストリーム層では，イベントを各種類ごとに時間方向に接続し，ストリームを形成する．さらに，ストリーム間の距離に応じて，複数のストリームを束ねて，アソシエーションストリームを生成する．

3.1 音源分離・定位モジュール

音源分離・定位モジュールでは，入力信号は，異なる方向からの混合音を仮定し，以下のような流れで処理を行っている．

ピーク抽出と音源分離: まず，入力音の周波数解析 (FFT) により得られるスペクトルに対し，ノイズの大きい低周波域とパワーの小さい高周波域を除去するため，バンドパスフィルタを適用する．次に，部屋の騒音

を計測することによって自動的に得られる閾値以上のパワーを持った ローカルピークを抽出する．同時刻に抽出されたローカルピークは互いの周波数が整数倍の関係と見なせるものをクラスタリングすることにより, 音源分離を実現している．

音源定位:

一般環境での音源定位は、部屋の反響などのために極めて難しい。そこで, 聴覚処理においても, (1) 音の倍音構造の利用, (2) 両耳間位相差 (IPD) を用いた聴覚エピポーラ幾何による定位, (3) 両耳間強度差 (IID) を用いた定位, (4) Dempster-Shafer 理論を用いた 2, 3 の結果の統合, によって, 音源定位のロバスト性を向上させている．

音の倍音構造の利用に関しては, 一つの倍音成分から音源方向を推定するのではなく, 複数の倍音成分を併せて使用することにより, 音源方向を推定している．

IPD, IID による定位に関しては, 一般に, 両耳聴における音源定位には, 頭部伝達関数 (HRTF) から求められる IPD, IID が使用される．しかし, HRTF は頭部の形状や環境に大きく依存し, 環境が変わる都度, 計測が必要であり, 実環境アプリケーションには不向きである．そこで, HRTF に依らない IPD を利用した音源定位法として, ステレオ視におけるエピポーラ幾何の概念を聴覚に拡張した聴覚エピポーラ幾何[7] を用いている．具体的には, 聴覚エピポーラ幾何より音源方向の仮説を 5° 毎に生成し, 入力と照合し, 方向毎に確信度を算出している．また, IID に関しては, 全倍音成分の IID の総和を計算し, 左, 右, 正面の 3 方向に対する確信度を算出している．

このようにして, IPD および IID から複数の音源方向の候補が確信度付で算出される．そこで, IPD から得られた音源方向を支持する確信度 (B_{P}), IPD から得られた音源方向を支持する確信度 (B_{I}) を Eq. (1) で示される Dempster-Shafer 理論によって統合し, IPD と IID の両方から音源方向を支持する新しい確信度 ($B_{\mathrm{P+I}}$) を生成する．

$$B_{\mathrm{P+I}}(\theta) = B_{\mathrm{P}}(\theta)B_{\mathrm{I}}(\theta) + \big(1 - B_{\mathrm{P}}(\theta)\big)B_{\mathrm{I}}(\theta) + B_{\mathrm{P}}(\theta)\big(1 - B_{\mathrm{I}}(\theta)\big) \quad (1)$$

最終的に音源分離・定位モジュールは, 分離した音ごとに, 音高情報, 確信度付き音源方向 (確信度の高い順に上位 20 位まで) および観測時刻からなる聴覚イベントを生成する．

3.2 顔抽出・認識モジュール

顔抽出・認識モジュールでは, 肌色検出と相関演算に基づくパターンマッチングの組合せによる顔の抽出[2], オンライン LDA[3] による顔の認識, 顔の大きさを仮定した定位を行うことによって, 高速で 顔の位置, 大きさ, 明るさに ロバストな顔抽出を実現している．

各顔毎に, 上位 5 つの確信度付きの顔 ID(名前) と位置 (距離, 方位角, 仰角) からなる顔イベントを生成する．

3.3 ステレオビジョンモジュール

ステレオビジョンモジュールは, ステレオ視による視差画像から人物らしい物体を抽出し, その正確な 3 次元位置を得る．具体的には, 左右のカメラの視差画像の生成, 視差画像からの物体抽出, 物体定位, ステレオイベント生成の順に処理が行われる．

視差画像は, 局所領域のマッチングによる対応点探索に基づいて生成される．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[11] を用いている．また, 事前にアフィン変換を用いた補正を施している．視差画像からの物体抽出は, 人体は縦長であることを利用して, 細かいノイズに左右されない人体およびそれに類する形状・大きさを持った物体の抽出を実現している．つまり, 2 次元の視差画像に対し, 視差値の縦軸方向のメディアンを横軸に沿って求めていくことによって, 視差画像を 1 次元化し, その 1 次元視差画像に対して視差の近い領域を分割することで, 物体の抽出を行う．抽出した物体はエピポーラ幾何により定位を行い, 最終的に, 距離, 方位角, 物体幅および, 観測時刻からなるステレオイベントを生成する．

3.4 アソシエーションモジュール

アソシエーションモジュールは, *SIG* がロバストに周り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．

ストリーム生成: まず, 各モジュールで発生する イベントの位置情報はイベントが観測された時刻にロボットから見た座標系 (*SIG* 座標系) における情報であるため, モータイベントを利用して絶対座標変換を行う．

各イベントは, 以下に該当するアルゴリズムによってストリームに接続され, 接続可能なストリームが存在しない場合, そのようなイベントから新しいストリームが生成される．

- **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られた値である.

- **顔, ステレオイベント:** 共通の ID をもち, 40 cm の範囲内で最も近い既存の顔, ステレオストリームに接続される. この値は, 秒速 4m 以上で人間が移動しないことを前提にして定めている.

また, ストリームは, 500ms 以上イベントが接続されない場合消滅する.

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物体 (人物) の連続的な動きを把握できるようになるだけでなく, ピッチ抽出, 顔認識などの一時的なミスによる曖昧性をストリーム全体の情報を利用して解消できるという利点がある.

アソシエーション: 複数のストリームが同一の人物に対するストリームであると判断された場合, これらのストリームはアソシエーションされ, より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を形成する. また, アソシエーションストリームを形成するストリームが消滅した場合, もしくは, 同一人物に由来するストリームであると判断されなくなった場合, 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, 複数のストリームに分割される. アソシエーションを行うことにより, 一時的なオクルージョンを聴覚情報を利用して解消したり, またその逆に音源方向として精度の高い視覚情報を利用できるようになるなど, 互いに情報を補い合うことで, 処理のロバスト性を高めることができる.

3.5 アテンション制御モジュール

注意を向けているストリーム方向に向くように 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. なお, 注意制御は下記の優先順位で行われる：

- アソシエーションストリーム
- 視覚ストリーム
- 聴覚ストリーム

4. アクティブ方向通過型フィルタ

方向通過型フィルタは, 基本的には特定の方向の IPD と同じ IPD をもったサブバンドを選択することによって特定の方向の音響信号を抽出するフィルタである[10].

しかし, 従来使用していた方向通過型フィルタは, 以下のような制約が存在した.

- HRTF を用いているため, 新しい家具を入れてしまったり, 湿度が変わったりするなど部屋の環境の動的な変化に対応できない.
- シミュレーション環境における動作確認のみである.
- 方向による感度の違い, および, 自らが動くこと (アクティブな動作) を考慮していないため, 感度のよい正面以外の音源定位の精度が悪い.
- 移動する物体を追跡する際は, HRTF が離散的な関数であるため, 補間が必要になる.

ここでは, HRTF に依らない聴覚エピポーラ幾何に基づいたアクティブ方向通過型フィルタによる手法を提案する. 詳細なアルゴリズムは以下の通りである.

(1) 注目しているストリームの方向をアソシエーションモジュールより取得する.
(2) 得られる方向は, 絶対座標系での方向であるため, 処理のレイテンシを考慮しつつ, 現在の SIG 座標系における方位角 θ を算出する.
(3) 方位角 θ の IPD $\triangle\varphi$ を聴覚エピポーラ幾何を適用し, 各サブバンド毎に計算する.
(4) 入力からピークを抽出し, IPD $\triangle\varphi'$ を計算する.
(5) 計算によって得られた IPD が $|\triangle\varphi' - \triangle\varphi| \geq \delta(\theta)$ を満たしているようなサブバンドを集める. ここで δ は θ で決定される関数であり, 測定によって求められる. 一般には, SIG の正面方向は感度がよいため δ は小さくなり, 側面方向に行くに従い, 感度が悪化するため δ は大きくなる.
(6) 集めたサブバンドからなる波形を再構築する.

このように, アクティブな SIG の動作に対応し, かつ方向によってアクティブに感度を 調整することで, アクティブな方向通過型フィルタを実現している. さらに, アソシエーションモジュールからのストリーム情報のフィードバックによって, 移動する物体, および自分が移動する場合でも分離が行えるようになった.

また, 聴覚ストリームがステレオなどの視覚ストリームとアソシエーションしている場合には, 聴覚ストリームの方向情報として, より精度の高い視覚処理による方向情報が得られるため, 精度の高い分離が可能である. さらに, ストリームは時間の流れを持っているため, ストリームの一部に調波構造を持たない部分が存在しても前後の情報によって, 分離を行うことが可能である.

5. 実験と評価

アクティブ方向通過型フィルタの効果を調べるため, 以下の 3 種類の実験を行った. 実験環境は, 約 10 平

方メートルの部屋で行い, ロボットとスピーカ間距離は 50cm とし, スピーカの方向は, ロボット正面方向を 0 ° としている. また, 音響信号には, “音声認識システム”[12] に付属する 毎日新聞記事の読上げデータを使用した.

評価指標として, Eq. 2 による分離前と分離後の SNR の変化, および ASR による単語認識率の変化を使用した. なお, $s(n)$, $s_o(n)$, $s_s(n)$ は, それぞれ, スピーカから出力される原波形信号, ロボットのマイクで収音された観測波形信号, 分離波形信号を指し, β は原信号と観測信号の減衰比を示す. また, ASR には, 日本語ディクテーションソフトウェア Julius を使用した.

$$SNR = 10log_{10}\frac{\sum_n (s(n) - \beta s_o(n))^2}{\sum_n (s(n) - \beta s_s(n))^2} \qquad (2)$$

実験 1　音源方向による感度の違いを調べ, アクティブ方向通過型フィルタの $\delta(\theta)$ を求めるため, 各モジュールで, $0° - 90°$ の定位の誤差を調べた. 結果を図 3 に示す.

次に, $0°$, $30°$, $60°$, $90°$ とスピーカ位置を 4 段階に変化させ, スピーカから出力される音声を抽出することを試みた. 音声抽出では, スピーカ方向は既知であるものとし, それぞれの場合に対し, 方向通過フィルタの通過範囲を $\pm 5° - \pm 90°$ として, 方向通過フィルタを使用しない場合に対する分離した音声の ASR による単語認識率の変化を調べた. 結果を図 4 に示す.

実験 2　一方のスピーカを $0°$ の方向に, もう一方のスピーカを $30°$, $60°$, $90°$ と 3 段階に変化させ, 同時に音声が出力されている状態で, 正面方向のスピーカからの音声の分離、抽出を試み, S/N 比を比較した. 結果を図 5 に示す.

実験 3　2 台のスピーカを正面に及び $\pm 60°$ に配置し, 正面の音源が移動する場合の音源分離、抽出を試みた. 視覚情報を使用する場合と使用しない場合の比較結果を図 6 に示す.

図 3 より, 聴覚による定位情報は視覚によるものより誤差が大きいことがわかる. また, 正面方向から $30°$ 付近までは $\pm 5°$ 以内それ以後, 悪化することがわかる.

図 4 の結果は, やはり正面方向は感度が高いことを示している. 例えば, δ を $20°$ とすると, 側面からの音声に対し, 正面を向いた場合と, そうでない場合で最大 50% 近い音声認識率の差が見られる. また, 正面から $30°$ 以上離れると, 方向通過型フィルタの δ を大きくとっても性能が上がらないことがわかる. これ

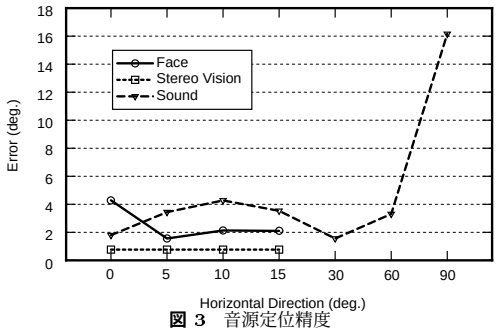

図 3　音源定位精度

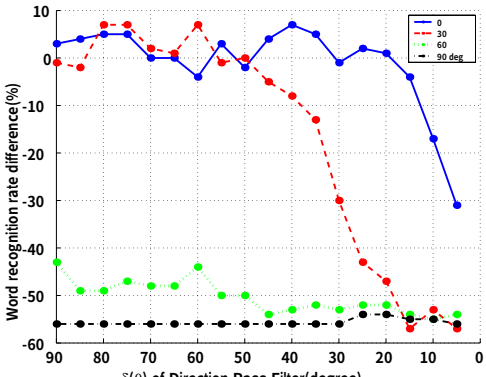

図 4　音声認識に見る音源方向別定位精度

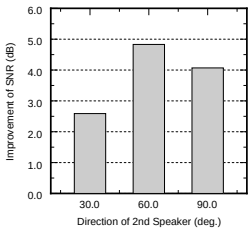

図 5　正面話者抽出 (2 話者)

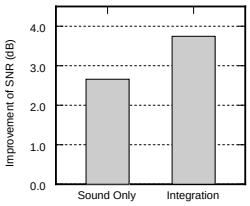

図 6　移動話者の抽出 (2 話者)

は, 外装によってマイクが前方に指向性を有しているためであろうと考えられる. これより, 音源と正対することにより, 感度が向上し, S/N 比の高い信号が取得できることがわかる. また $00°, 30°$ で, 分離を行わない場合よりも $5 - 10\%$ 認識率が上昇しているのは, 方向通過型フィルタにより周囲の雑音が除かれ, S/N 比が向上したためと考えられる.

図 5 の結果より, 互いの話者が $30°$ と近い場合は $3dB$ 程度, 遠くなると $4 - 5dB$ 程度の分離効果が見られることがわかる. 音声認識では, 良好な結果は得られなかった. さらに, 精度の高い分離が必要であろう.

図 6 より、視覚情報の使用によって, 若干の効果が見られる. 効果がそれほど大きくないのは, カメラの視野 ($40°$ 程度) 内でスピーカを移動させていたため, 移動量が少なかったことが挙げれれる. 断続的な聴覚

ストリームの場合に, 今回は手動でストリームのつなぎ併せを行ったが, 視覚情報ではそれを解消できることを考えれば, S/N 比だけでなく視覚情報は有効である.

6. 結　　論

本稿では, アクティブな動作に対応し, かつアクティブに感度を制御するアクティブ方向通過型フィルタを構築し, その評価を行った. 結果として, アクティブな動作, および視聴覚の情報統合による上位モジュールからのストリーム情報のフィードバックによって, アクティブ方向通過型フィルタの精度を向上できることを示した.

しかし, 音声認識や話者認識などのフロントエンドとして実用に耐えうる音源分離を行うためには, 部屋の反響や雑音の考慮, 雑音によって元の信号が復元できない場合などのミッシングデータの扱いを考慮する必要があろう.

よりロバストな知覚処理のためには, 位置情報のみを統合の対象として利用するのではなく, 話者同定, 音声認識, 物体認識などによる名前や意味レベルでの情報を, 位置情報と統合するような, 複数メディア間での階層的な情報統合の枠組みが必要であろう. 実際に, このような階層的な情報統合は, 一般に人間の知覚が階層性を有していることを考えれば, 妥当であろうし, 人間とのスムーズなソーシャルインタラクションが可能なレベルの知覚処理を実現するためには, 複数のメディア間で階層的な情報統合を行うことが必然であろう.

参 考 文 献

1) C. Breazeal and B. Scassellati. A context-dependent attention system for a social robot. In *Proceedints of the Sixteenth International Joint Conference on Atificial Intelligence (IJCAI-99)*, pages 1146–1151, 1999.
2) K.Hidai, H.Mizoguchi, K.Hiraoka, M.Tanaka, T. Shigehara, and T. Mishima. Robust face detection against brightness fluctuation and size variation. In *Proceedings of IEEE/RAS International Conference on Intelligent Robots and Systems (IROS-2000)*, pages 1397–1384. IEEE, 2000.
3) K.Hiraoka, S.Yoshizawa, K.Hidai, M.Hamahira, H. Mizoguchi, and T. Mishima. Convergence analysis of online linear discriminant analysis. In *Proceedings of IEEE/INNS/ENNS International Joint Conference on Neural Networks*, pages III–387–391. IEEE, 2000.
4) M. Z. Ikram and D. R. Morgan. A multiresolution approach to blind separation of speech signals in a reverberant environment. In *Proceedings of 2001 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2001)*, pages 2757–2760. IEEE, 2001.
5) K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka. Application of bayesian probability network to music scene analysis. In *Working Notes of the IJCAI-95 Computational Auditory Scene Analysis Workshop*, pages 52–59. AAAI, 1995.
6) 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 *Proceedings of 6th European Conference on Speech Communication Technology (EUROSPEECH-99)*, pages 1723–1726. ESCA, 1999.
7) K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano. Active audition for humanoid. In *Proceedings of 17th National Conference on Artificial Intelligence (AAAI-2000)*, pages 832–839. AAAI, 2000.
8) K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano. Real-time auditory and visual multiple-object tracking for robots. In *Proceedints of the Seventeenth International Joint Conference on Atificial Intelligence (IJCAI-01). submitted*, 2001.
9) T. Nakatani and H. G. Okuno. Harmonic sound stream segregation using localization and its application to speech stream segregation. *Speech Communication*, 27(3-4):209–222, 1999.
10) H.G. Okuno, K. Nakadai, T. Lourens, and H. Kitano. Separating three simultaneous speeches with two microphones by integrating auditory and visual processing. In *Proceedings of European Conference on Speech Processing(Eurospeech 2001)*. ESCA, 2001.
11) 岡田 慧, 加賀美 聡, 稲葉 雅幸, and 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**, 18(6):138 – 143, 2000.
12) 鹿野 清宏, 伊藤 克亘, 河原 達也, 武田 一哉, and 山本 幹雄. **音声認識システム**. オーム社, 2001.