第 5 章

聴覚情報の統合による音源定位と追跡

4 章では, アクティブな動作によって, 発生するモータノイズを, モータ制御信号とヒューリスティクスを利用してキャンセルし, 聴覚エピポーラ幾何により, 動作時の連続的な純音の音源定位を可能にした. 本章では, より一般的な環境でのロバストな動作を目指し, 複数の聴覚的な手がかりを統合し, 実時間・実環境での調波構造を持った複数の音源を扱う音源定位・追跡システムを実現する.

5.1 複数の聴覚情報の統合した音源定位・追跡システム

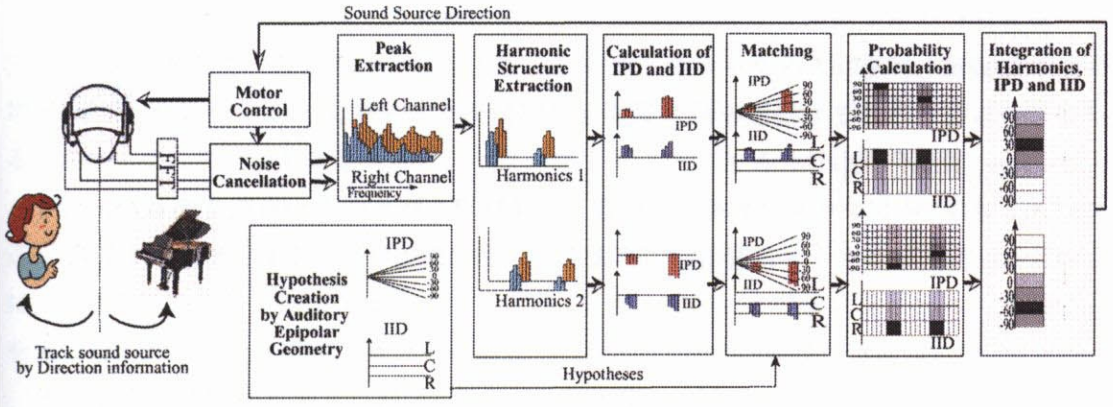


図 5.1: 音源定位・追跡システムの構成図

4 章では, 実環境アプリケーションには, HRTF を用いず, 連続的に音源定位を行うことができる方法が望ましく, そのような音源定位法の一つとして, 聴覚エピポーラ幾何を提

案した. しかし, 定位に使用する特徴は IPD のみで, また, 単一周波数を扱っていたため調波構造を有した音が扱えず, 音響環境による影響を受けやすいなど汎用性や精度の面で乏しかった. そこで, Dempster-Shafer 理論により, 倍音成分すべてについて, IPD, IID という聴覚情報を統合するように拡張することで定位のロバスト性向上を目指す.

音源定位・追跡システムにおける処理の概略を, 図 5.1 に示す. 入力信号には, 異なる方向からの混合音を仮定しており, 4 章で説明したシステムと同様にサンプリング, FFT による周波数解析, ノイズキャンセルを行い, 左右のチャネル毎のスペクトルを生成する. スペクトル情報は, ピーク抽出部に送られ, 左右のスペクトルからローカルピークが抽出される. 抽出したピークに対し, 調波構造抽出部で調波構造を抽出する. 調波構造は, 基本周波数が重なっていない限り, 純音を含め, 複数の調波構造を抽出することができるように設計されている. 各調波構造を抽出した後それぞれの倍音成分に対し, IPD, IID を計算する. 次に, 聴覚エピポーラ幾何を用いて, 5° おきに音源方向ごとの IPD, IID 仮説を生成し, 入力とのマッチングを行う. マッチング時の距離に応じて, IPD および IID の確信度が計算され, これらの確信度を音源方向ごとに Dempster-Shafer 理論に基づき統合する. 最終的に, 最も確信度の高い方向情報を出力とする. この方向情報は, モータ制御部に送られ, 必要に応じて音源を追跡できるようになっている.

それでは, 以下に各処理部の詳細を説明する.

5.2 ピークの抽出

ピークの抽出はスペクトラルサブトラクション法 (*Spectral Subtraction, SS*) [21] に基づいた方法によって行う. この方法は, スペクトルからピークを抽出し, これをスペクトルから減算し, 残差スペクトルを生成する. そして, その残差スペクトルからピークが見つからなくなるまで処理を繰り返す残差駆動型アーキテクチャ (*Residue Driven Architecture, RDA*)[83] である.

この方法では, インクリメンタルに正確なピーク抽出を行うことが可能であるが, 一般にこのような方法をロボットに適用して正確なピークの抽出を行う場合, 次のような問題が発生する.

一つ目は, 周波数解析の問題である.

周波数解析には従来から様々な手法が試みられている。 高速性を考慮して高速フーリエ変換 (FFT) を用い、 短時間 FFT を行う手法が一般的である [83] が、周波数解析や聴覚フィルタに関しては様々な研究が行われている。例えば、人間の聴覚特性に注目した場合、周波数軸の対数スケール変換が FFT と同等の計算速度で得られる Wavlet 変換を用いている研究も多い [2]. 特に Wavlet-Mellin 変換は聴覚の周波数解析との対応がよいことから、聴覚の計算理論を構築する上で注目を集めている [54, 124]。また、音楽を対象と

した場合には、音階の周波数値に対応するフィルタを用いて、周波数軸が対数スケールとなるようなフィルタバンクも用いられている [127]。Brown や Cooke は、聴覚の生理学的な知見に基づき、蝸牛における基底膜の応答特性を模したガンマトーンフィルタを用いた周波数解析を行っている [27]。また、より基底膜応答に適しているガンマチャープフィルタも注目されている [53] 音の持つ調波構造に着目した場合、調波構造を直接、取り出すような櫛型フィルタを用いたアプローチもある。特に、安藤らの一連の研究は、新たなサブバンド特徴量として時間領域の零点に注目し、従来、外乱要因であった調波間干渉を積極的に利用した新しい調波信号分析用のフィルタバンクを提案しているという点で興味深い [89, 99]。本研究では、調波構造を有しない音に対応させ、かつ、ロボットへの適応を前提にしているため、実時間性の点で有利である短時間 FFT を用いるものとした。しかし, FFT は高速である反面, 人間の聴覚は周波数軸がほぼ対数軸であることを考慮すると, 周波数軸が線形であるため, 低周波数域で十分な周波数分解能を得ることができず, 正確なピーク抽出が難しい. もちろん, 両耳聴処理における, 左右のチャネルの対応や音楽のピッチ抽出問題を考えると精度の高いピーク抽出は本質的である. FFT を用いて低周波数域で, 高い解像度を得るためには, 時間分解能を犠牲にする必要があるが, FFT のポイント数が大きくなるため, 処理時間的にも不利である. 高速に低周波数域で十分な周波数分解能を得るような手法が必要である.

二点目は, サブトラクション処理の速度の問題である. ロボットへ適用するためには, 実時間処理が求められる. しかし, 正確なサブトラクションを行うためには, 正確な基本周波数の取得, 位相情報の計算, 倍音成分のパワーの抽出など, 最適値を求めるために, 計算コストの高い処理が必要である. 実際, 残差駆動型アーキテクチャにより正確な調波構造を抽出できる Bi-HBSS では, 倍音成分のパワー合計を最大にするような基本周波数を推定した上で, 残差のパワーが最小になるような位相を求めるという計算を, 時間 (波形) 領域で行う必要があるため, 数秒のデータに対して数時間 (Sun SS5) の処理時間が必要である. 処理速度の点からも高速なスペクトラルサブトラクション法が必要である.

また, Bi-HBSS に限って言えば, 調波構造に基づいて処理を行うため, 純音など調波構造を持たないピークを抽出することはできないことも問題である. このため, Bi-HBSS は, 音声の無声子音を抽出することができず, 最後に抽出した調波ストリームに残差成分を足し合わせるという方法で無声子音成分への対応を行っている.

そこで, 以下では, 正確で高速なピーク抽出法を導入する.

5.2.1 スペクトラルサブトラクションを用いたピーク抽出

これらの問題を解決するためのピーク抽出法を提案する. このピーク抽出法は, スペクトラルサブトラクションに基づいているが, 高速で調波構造を有しないピークの抽出も可

能である.

ピークの抽出では, 低周波数域での周波数分解能不足を補うために, ピークサブバンドと両隣のサブバンドから正確なピークの周波数値やパワー値を近似する方法を用いる. ただし, このピーク周波数近似では, 次のような仮定をおくことにより, 比較的簡単な計算で, 周波数分解能を補完し, 正確なピーク近似を行っている.

- 抽出したピークの近傍に他のピークがない.
- 入力信号が正弦波である.

これらの仮定は, 一般に, 音声など非定常な信号であっても, 十数 ms 以下の短い時間では正弦波を仮定できること [50], 一般的な状況では, 音楽など特殊な状況を除き, 周波数成分が重畳するような状態は, 長期間にわたって続かず, 10Hz から 20Hz 程度のサブバンドであれば, 同一時刻のある短い瞬間においては, 2 つの信号の周波数軸上での重なり具合が最小となる [11] といった知見を考慮している. 本研究では, 48 kHz で 4,096 点の FFT を行っており, これらの仮定は現実的であるといえる.

スペクトラルサブトラクションでは, この近似法により得られたピークの倍音成分も同時に抽出することにより, 調波構造の抽出を行い, 抽出倍音を含めたピークを入力スペクトルからサブトラクションしている.

以下では, ピーク周波数の近似, 調波構造の抽出についての説明と, これらを用いた高速で正確なピーク抽出の具体的な説明を行う.

5.2.2 ピーク周波数近似関数の導出

入力信号が正弦波の場合, 入力波形は式 (5.1) として表すことができる. また, 窓関数としてハニング窓を使用した場合, 窓関数は, 式 (5.3) となる. ここで, A_0 は入力信号 $f(t)$ の振幅, ω_0 は周波数, ϕ_0 は位相項, $w(t)$ は窓関数, T は窓長を表す.

$$f(t) = A_0 \cos(\omega_0 t + \phi_0) \tag{5.1}$$

$$w(t) = \frac{1}{2} w_r(t) \left[1 + \cos \frac{2\pi(t - T/2)}{T} \right], \tag{5.2}$$

$$w_r(t) = \begin{cases} 1 & 0 < t < T \\ 0 & t \leq 0, t \geq T \end{cases}$$

入力信号 $f(t)$ に対して, フーリエ変換 $F(\omega)$ を求めると式 (5.3), (5.4) として表すことができる.

$$F(\omega) = \int_{-\infty}^{\infty} f(t)\, w(t)\, e^{-j\omega t} dt \tag{5.3}$$

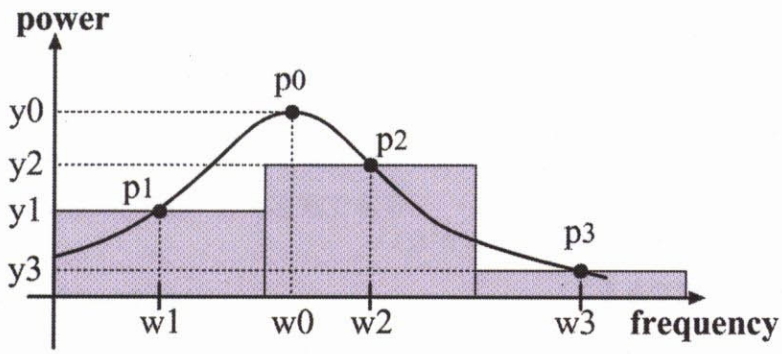


図 5.2: フーリエ変換を用いて算出したスペクトル上のピーク例

$$= G(\omega - \omega_0) + G(\omega + \omega_0)$$

$$G(\omega) = \frac{2\pi^2 A_0 e^{-j\left(\frac{T}{2}\omega - \phi_0\right)}}{T^2\omega\left(-\omega^2 + \frac{4\pi^2}{T^2}\right)} \sin\frac{T}{2}\omega \tag{5.4}$$

式 (5.3) において, 右辺の $G(\omega - \omega_0)$ と $G(\omega + \omega_0)$ は $|\omega_0| > \frac{4\pi}{T}$ であれば, お互いに及ぼしあう影響が 25 dB 以下になるため, $|\omega_0| \leq \frac{4\pi}{T}$ の時は, $F(\omega) = G(\omega)$ とみなすことができる. 従って, $|\omega_0| \leq \frac{4\pi}{T}$ の時, 式 (5.5) が成り立つ.

$$F(\omega) = \frac{2\pi^2 A_0 e^{-j\left(\frac{T}{2}(\omega-\omega_0) - \phi_0\right)}}{T^2(\omega - \omega_0)\left(-(\omega - \omega_0)^2 + \frac{4\pi^2}{T^2}\right)} \sin\frac{T}{2}(\omega - \omega_0) \tag{5.5}$$

ここで, 図 5.2 のようなピーク p_2 が検出されたとする. このピークの両隣のサブバンドを p_1, p_3 とする. p_1, p_2, p_3 の周波数値, パワー値を式 (5.5) に代入して, 真のピーク p_0 の周波数値を表す ω_0 について解くと, 式 (5.6), (5.7) を得ることができる.

$$\omega_0 = \omega_2 - \frac{2\pi(-|y_2| + 2|y_3|)}{T(|y_2| + |y_3|)} \tag{5.6}$$

$$= \omega_2 + \frac{2\pi(2|y_1| - |y_2|)}{T(|y_1| + |y_2|)} \tag{5.7}$$

同様にして, ω_0 における位相 ϕ_0 と, パワー値 y_0 は, 式 (5.8), (5.9) となる.

$$\phi_0 = Arg(y_2) + \frac{T}{2}(\omega_2 - \omega_0) \tag{5.8}$$

$$y_0 = \frac{T^2 d\omega\left(-d\omega^2 + \frac{4\pi^2}{T^2}\right)|y_2|}{2\pi^2 \sin\frac{T}{2}d\omega}, \tag{5.9}$$

$$d\omega = \omega_2 - \omega_0$$

式 (5.6), (5.7), (5.8), (5.9) を用いれば,

$$p_0 = F_{approx}(p_1, p_2, p_3) \tag{5.10}$$

として,

- 抽出したピークの近傍に他のピークがない
- 入力信号が正弦波である

という仮定の下で, 正確なピークの周波数値, およびパワー値を得る関数 F_{approx} を定義することができる.

5.2.3 調波構造の抽出

ピーク p_0 が抽出された後, そのピークが調波構造を有しているかどうかを調べる方法を示す. まず, p_0 の周波数値 ω_0 より低い周波数からなる調波構造の仮説を生成する. 各調波構造仮説は以下のように表すことができる. ここで, f_{min} はピークを抽出する最小周波数値で 90 Hz としている.

$$Hl_j = [\frac{1}{j}\omega_0, \frac{2}{j}\omega_0, \cdots, \frac{k}{j}\omega_0, \cdots, \omega_0], \quad j = 1, 2, \cdots, \lfloor \omega_0 / f_{min} \rfloor$$

次に, 入力スペクトル S 上で, 各仮説 Hl_j の倍音成分 $\frac{k}{j}\omega_0$ に近接する周波数にピークがあるかどうかをチェックする. 式 (5.11) の条件が満たされるピーク ω_l が見つかった場合, $\frac{k}{j}\omega_0$ に近接するピークとみなす.

$$\frac{\left|\frac{k}{j}\omega_0 - \omega_l\right|}{\frac{k}{j}\omega_0} \leq 0.06 \tag{5.11}$$

ここで, 0.06 という値は, 二つの周波数成分の周波数値のずれが整数倍から 6%以内であれば, 一つの音として聴こえるという心理学的実験から得られた知見 [127] を考慮して定めた値である. 生成された仮説の中で, 以下の条件を満たす仮説を最も信頼性の高い仮説とみなし, その最大公約数となる周波数値 (通常は最小周波数値) を基本周波数 $F0$ とする.

1. 仮説の全周波数に対する見つかったピークの数の率が一定以上であること (実験的に 80% を用いている).
2. 条件 1 を満たすすべての仮説の中でピーク数が最大であること.

ω_0 よりも高い周波数を持っている倍音の検索は, $F0$ を基準として, 周波数の低い方から行う. 検索は, 連続して倍音成分が見つけられなかった場合に終了する. このようにして,

得られた調波構造を H とすれば,

$$[F0, H] = Harmonics(p0, S) \tag{5.12}$$

として, 調波構造抽出を定義することができる. この方法は, 基本周波数成分が存在しない場合 (*missing fundamental*) にも有効である.

5.2.4 高速なピーク抽出

ピーク周波数の近似, および, 調波構造の抽出を踏まえて, ピーク抽出の具体的な処理を示す. まず, FFT を用いて得られる, 時刻 t のスペクトル S を以下のように定義する. ただし, f_k は k 番目のサブバンドの周波数値, p_k は f_k のパワー値, $NFFT$ は FFT のポイント数を示す.

$$S = \left[S_d(0), \cdots, S_d(k), \cdots S_d\left(\left\lfloor \frac{NFFT}{2} \right\rfloor\right)\right]$$
$$S_d(k) = [f_k, p_k] \quad \left(k \in N | 0 < k < \left\lfloor \frac{NFFT}{2} \right\rfloor\right)$$

次に, S から以下のような処理を繰り返すことによって, 調波構造を有したピークの抽出を行う.

1. S から, パワーが閾値 (部屋の暗騒音の音圧レベルに, 感度パラメータ 10 dB を加えた値) 以上で, かつ, 90 Hz から 3 kHz の周波数を持ったサブバンドのうち, 最大パワーを持つサブバンドを $S_d(pmax)$ として抽出する. ここで, 感度パラメータは実験的に求めた値であり, 部屋の暗騒音は, 事前にシステムにより計測されている. 周波数に帯域制限をかけているのは, 抽出されたサブバンドのうち低周波ノイズとパワーの小さい高周波域をカットするためである.

$$S_d(pmax) = [f_{pmax}, p_{pmax}]$$

2. 式 (5.10) を用いて, 抽出したサブバンドの周波数値, パワー値とともにその両隣のサブバンド値を利用して, 近似ピーク P を推定する.

$$P = F_{approx}(S_d(pmax - 1), S_d(pmax), S_d(pmax + 1)) \tag{5.13}$$

3. スペクトル S に, 抽出したピーク P と高調波関係にあるピークが含まれる場合, これらを式 (5.12) を用いて H として抽出する. P が調波構造を持たない場合には, H には P のみが含まれる.

$$[F0, H] = Harmonics(P, S) \tag{5.14}$$

4. 式 (5.15) に示すように, 入力スペクトル S から, 抽出した調波構造 H をサブトラクトし, 残差 R を求める.

$$R = S - H \tag{5.15}$$

5. 残差 R を次のループにおける入力スペクトル S とし, 残差のパワーが十分小さくなるまで処理を繰り返す.

結果として, 時刻 t におけるピークの集合は, H の集合として抽出される. 抽出した一つの調波構造 H は一つの音とみなすことができる. また, H は調波構造を有しない単独のピーク抽出が可能であるため純音のような音源に対応することができる. さらに, 最適値を求めるようなコストの高い計算が必要なく, かつ周波数領域で計算が行われるため, 高速にピークを抽出することが可能である.

5.3 IPD と IID に関する仮説生成と照合

抽出した調波構造音に対し, IPD と IID を用いて音源方向 (方位角) を計算する. ここで便宜上, H に含まれる周波数値は昇べきの順に並んでいるものとする.

まず, 左右のチャネルのうちパワーの大きい方のチャネルで抽出された H を基準として, その各倍音ごとに IPD $\Delta\varphi_s$ を計算する. $\Delta\varphi_s$ は, H に含まれる各倍音の周波数値に対応するサブバンドを左右のチャネルから選択し, 式 (4.1) に基づいて, その位相差を求めることによって得られる. ただし, IPD による定位は, 回り込みが発生するような高い周波数域では有効性が薄れてしまう. *SIG* では, 左右のマイクロホン間距離が約 18 cm であることから, 計算上, 1.2 ~ 1.5 kHz の間にその閾値が存在するため, 本稿では, IPD は, 1.5 kHz 以下の倍音についてのみ扱うものとする. 次に, 聴覚エピポーラ幾何の式 (4.8) を変形し,

$$\Delta\varphi = \frac{2\pi f}{v} \times r\,(\theta + \sin\theta) \tag{5.16}$$

とすると, IPD が音源方向 θ と周波数 f の関数となる. これを利用して, ±90° (*SIG* 正面の方向を 0° とし, 左手が正, 右手が負の値とする) の範囲で 5° おきに, $\Delta\varphi_s$ に対応する IPD 仮説 $\Delta\varphi_h$ を生成する. 式 (5.17) に定義された距離関数により, 入力音と各仮説間の距離 ($d(\theta)$) を計算する. ここで, n_{th} は周波数が 1.5 kHz 以下である倍音数である.

$$d(\theta) = \frac{1}{n_{th}} \sum_{i=0}^{n_{th}-1} \frac{\left(\Delta\varphi_h(\theta, H(i)) - \Delta\varphi_s(i)\right)^2}{H(i)} \tag{5.17}$$

IID に関しては, IPD と同様に入力音の各倍音の左右チャネル間パワー差から求める. 但し, IID については, 仮説推論ではなく, 式 (5.18) で示される判別関数を用いて, 音源が *SIG* の正面方向か左側か右側かのみを判定するものとする. つまり, 周波数が f である倍

音の IID を $\Delta I_s(f)$ とした場合, S_I が正なら音源は SIG の左方向に存在し, 0 に近ければ正面方向, 負であれば右方向に存在するものとする.

$$S_I = \sum_{i=n_{th}}^{n-1} \Delta I_s(H(i)) \tag{5.18}$$

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

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

IPD, IID から得られる音源方向を支持する値から, これらを Dempster-Shafer 理論によって統合しロバストな音源方向を推定するために, それぞれの値を確信度に変換する. Dempster-Shafer 理論は, Bayes の確率理論では, うまく表すことができないような人間の主観にかかわる確信度などを表すことができる枠組みを提供している [33, 55]. Dempster-Shafer の確率理論によれば, Dempster の結合規則によって, 独立な証拠から推論された基本確率を統合することができる.

A_{1i}, A_{2j} $(i, j = 0, 1, 2 \cdots)$ なる焦点要素 (focal element) につき, 独立した証拠から得られた基本確率 m_1 および m_2 があるとしよう. この際に, 統合された確実性尺度 $m(A_k)$ は式 (5.19) で表すことができる.

$$m(A_k) = \frac{\displaystyle\sum_{A_{1i} \cap A_{2j} = A_k} m_1(A_{1i})\, m_2(A_{2j})}{1 - \displaystyle\sum_{A_{1i} \cap A_{2j} = \phi} m_1(A_{1i})\, m_2(A_{2j})} \quad (A_k \neq \phi) \tag{5.19}$$

本研究では式 (5.19) の A_1 を特定の方向 θ に対する音源定位とし, A_1 のみを考える. また, $m_1(A_{11})$ と $m_2(A_{21})$ は, それぞれ IPD, IID に基づく θ の定位を支持する確信度であって, それぞれ $B_{\mathrm{IPD}}(\theta)$, $B_{\mathrm{IID}}(\theta)$ であると考える. 即ち, IPD, IID から得られる音源定位の評価値を, その音源方向を支持する「信用」ととらえると, 式 (5.19) に示す Dempster の結合規則の各値は,

$$\begin{cases} m_1(A_{11}) &= B_{\mathrm{IPD}}(\theta), \\ m_2(A_{21}) &= B_{\mathrm{IID}}(\theta), \\ m_1(\{A_{11}, A_{12}\}) &= 1 - B_{\mathrm{IPD}}(\theta), \\ m_2(\{A_{21}, A_{22}\}) &= 1 - B_{\mathrm{IID}}(\theta) \end{cases} \tag{5.20}$$

となる. 従って, 式 (5.19) より, 統合された θ を支持する定位の確信度は,

$$B_{\mathrm{IPD+IID}}(\theta) = m(A_1)$$

表 5.1: IID の確信度 $B_{\mathrm{IID}}(\theta)$ の定義

	θ	$90° \sim 35°$	$30° \sim -30°$	$-35° \sim -90°$
S_I	+	0.35	0.5	0.65
	-	0.65	0.5	0.35

$$= 1 - (1 - B_{\mathrm{IID}}(\theta))(1 - B_{\mathrm{IPD}}(\theta)) \tag{5.21}$$

となる.

そこで, 次に, $B_{\mathrm{IPD}}(\theta)$ と $B_{\mathrm{IID}}(\theta)$ の定義を行う. IPD の確信度 $B_{\mathrm{IPD}}(\theta)$ は, 式 (5.17) によって得られた距離に対し, 式 (5.22) で定義される確率密度関数を適用することによって得る.

$$B_{\mathrm{IPD}}(\theta) = \int_{-\infty}^{\frac{d(\theta)-m}{\sqrt{\frac{s}{n}}}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \tag{5.22}$$

ここで, m と s は, それぞれ, $d(\theta)$ の平均と分散であり, n は d の個数である. IID の確信度 $B_{\mathrm{IID}}(\theta)$ は, 式 (5.18) の S_I の値に応じて, 表 5.1 のように定義した.

これらによって得られた値を, それぞれ IPD と IID から音源方向を支持する確信度と捉え, 式 (5.21) で示される Dempster-Shafer 理論によって, IID と IPD の確信度を統合し, IPD と IID の両方から音源方向を支持する新しい確信度 $B_{\mathrm{IPD+IID}}(\theta)$ を各 θ ごとに生成する.

5.5 音源の定位と追跡

音源定位システムは, 統合によって得られた $B_{\mathrm{IPD+IID}}(\theta)$ のうち, 最大の確信度を有する θ とその時の基本周波数 ($F0$) を出力する. また, モータ制御に, 音源定位情報を送出することにより, 特定の音源を追跡することが可能である.

ただし, 本章の音源追跡は, MATLAB によるシミュレーション環境で行う. 再現性を考慮するという意味で, 予め, 録音したベンチマークデータに対して, 下記のアルゴリズムを利用して追跡を行っている. ベンチマークデータは同じ音響信号に対して *SIG* の向きを 5° おきに, ±90° の範囲で変えて計測したデータで, *SIG* の向きごとにラベリングされている. ベンチマークの詳細については, 評価実験で述べる. 音源の追跡は, 以下のアルゴリズムを用いて行う.

1. *SIG* の初期位置に対応したベンチマークデータを入力とし, 提案した音源定位法に基づき音源定位を行う.

2. *SIG* は, 推定された音源方向に基づき, 音源方向を向く. 向く角度は, 実際に推定した方向と現在の方向の差の半分とする. つまり, 現在の方向と推定方向の差が 60° であれば, その半分の 30° だけ音源方向に向くものとする. ただし, 現在の角度と推定された音源方向の角度差が 10° 度以下の場合は, 完全に推定方向を向くものとする.
3. *SIG* が向いた方向に対応するベンチマークデータを入力とし, 音源定位を行う. この処理を以下の条件のいずれかが満たされるまで繰り返す.
   追跡成功: *SIG* の方向と推定方向の差が 0° になる.
   タイムアウト: 繰り返しが 10 回以上になる.
   定位範囲外: 推定音源方向 ±90° を超える.

実環境における実時間追跡については, 6 章で述べる.

5.6 音源定位・追跡システムの評価

構築した音源定位・追跡システムを評価するために,

1. ピーク抽出
2. 音源定位と追跡

の 2 点について評価を行った.

5.6.1 ピーク抽出の評価

ピーク抽出の評価は, Windows NT (PentiumIII 550Mz dual) で MATLAB を用いて以下の 3 つの実験を通じてその評価を行った.

実験 5-1: 正弦波のピーク抽出の精度
実験 5-2: 音声のピーク抽出精度
実験 5-3: Bi-HBSS に対する計算量の評価

最初の実験では, ピーク抽出の基本性能を検証し, 2 つめの実験では, ピーク抽出法の音声への適応可能性を検証する. また, 最後の実験では, 計算量の評価を Bi-HBSS と比較することにより, 高速な動作が可能であることを示す.

■**実験 5-1: ピーク抽出の精度** 提案したピーク抽出法を正弦波のピーク抽出を用いて評価する. 評価方法としては, ピークサブバンドとその両側のサブバンドの周波数とパワー値から, それらの点を通る 2 次関数を導出し, その頂点を求めることで, 擬似的にピーク

値を算出する 2 次近似法と比較するという手法をとった．想定したケースは以下の 2 通りである．

A: 正弦波を入力として用いた．周波数値は，5500 Hz から 6500 Hz までスイープさせた．

B: 2 つの正弦波を混合した信号を入力した．一方は 6000 Hz に固定し，もう一方は 5500 Hz から 6500 Hz までスイープさせた．

A の場合のピーク周波数値の抽出結果を図 5.3 に，パワー値の抽出結果を図 5.4 に示す．また，**B** の場合のピーク周波数値の結果を図 5.5, 5.7 に，パワー値の抽出結果を図 5.6, 5.8 に示す．

検証した 2 つの方法の計算量はほぼ同等であったが，精度的には，以下のような結果となった．

図 5.3 より，正弦波に関して，スペクトラルサブトラクション法は 2 次近似法に比べて，正確な抽出が可能であることがわかる．図 5.4 からはパワー値においても，入力信号のパワー (256) を正確に抽出できていることがわかる．

図 5.5 から，混合した正弦波のうち 6000Hz に固定された音源を，正確に抽出していることがわかる．これは 2 次近似法でも同様である．また，図 5.6 から，どちらの手法もピーク周波数値が十分離れていれば，ピークのパワー (256) を正確に抽出していることがわかる．しかし，周波数が近接している場合は，2 次近似法の方が安定していることがわかる．

図 5.7 より，周波数をスイープさせた音源に対しては，2 次近似法と比べ正確である．これは，周波数値の微妙な変化を捕らえられるということを示しており，これにより，FFT の周波数分解能を補うことができることがわかる．しかし，6000 Hz の場合を見てもわかるように，この方法では周波数が重なっている場合のピークの抽出は難しい，図 5.8 より，ピークパワー (205) の抽出についても，同様の傾向が見られる．

■実験 5-2: 音声のピーク抽出精度 Bi-HBSS で利用されたベンチマークである「あきち」と「おもむき」の混合 [93] からピークの抽出を試みた．なお，Bi-HBSS の制約から，実験は，12kHz, 16bit のサンプリングレートで行った．このベンチマークのスペクトログラムを図 5.9 に示す．

Bi-HBSS により最終的に残った残差スペクトログラムを図 5.10 に示す．また，本研究で用いたピーク抽出法による残差スペクトログラムを図 5.11 に示す．各図では明るい部分が高いパワーを示し，暗い部分がパワーが低いことを示している．

図 5.10 は，図 5.11 に比べ，明るい部分が多い．これは，最終的にピークの抽出ができなかった部分であり，音声では，調波構造を持たない子音にあたる．図 5.11 では，調波構造を持たないピークも純音として抽出されているため，スペクトログラムが全体的に暗く

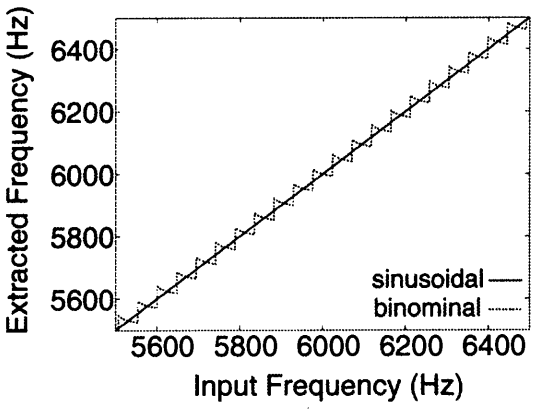


図 5.3: **A** から抽出された正弦波のピーク周波数値

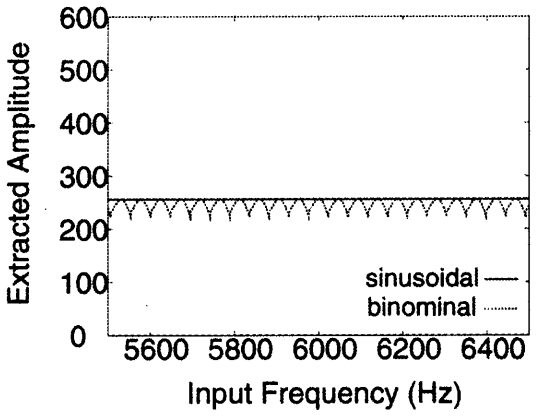


図 5.4: **A** から抽出された正弦波のピークパワー値

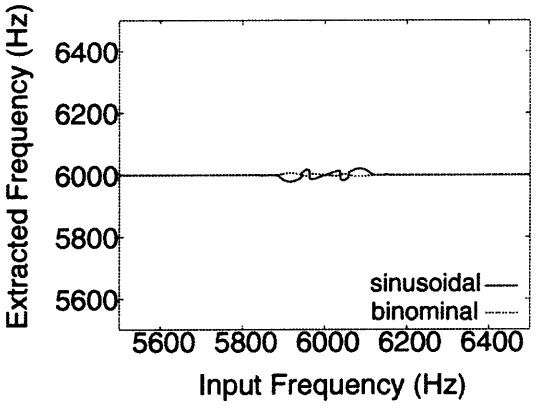


図 5.5: **B** から抽出された 6000 Hz の正弦波の周波数値

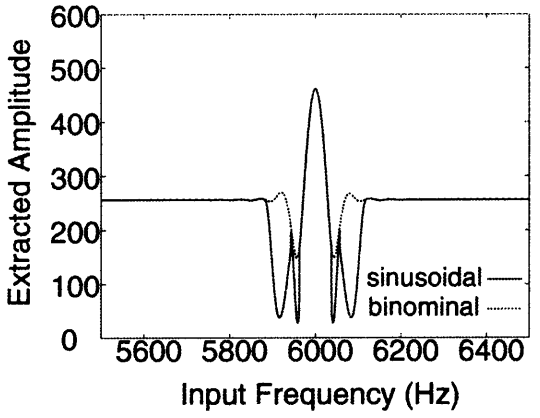


図 5.6: **B** から抽出された 6000 Hz の正弦波のパワー値

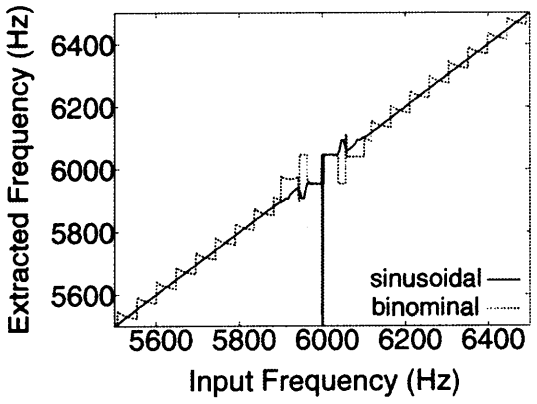


図 5.7: **B** から抽出されたスイープ信号のピーク周波数値

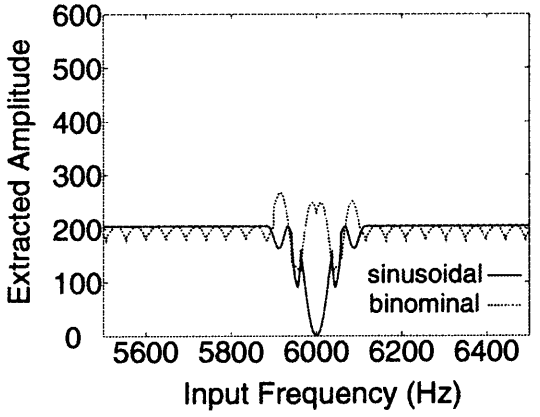


図 5.8: **B** から抽出されたスイープ信号のピークパワー値

なっており, ピークが的確に抽出されていることがわかる. これは, 抽出したピークから再合成した音を聞いても明らかである.

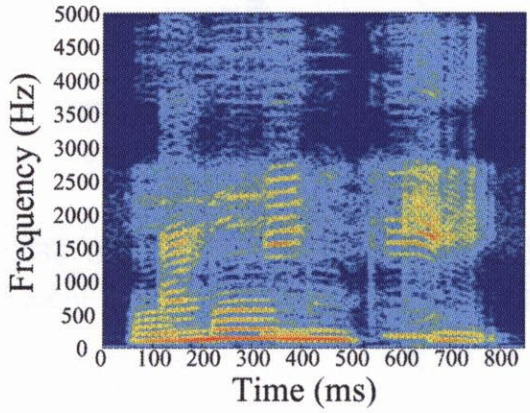


図 5.9: 「あきち」と「おもむき」を混合したスペクトログラム

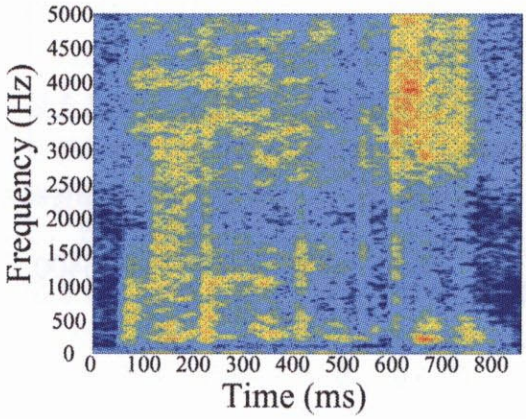


図 5.10: Bi-HBSS による残差スペクトログラム

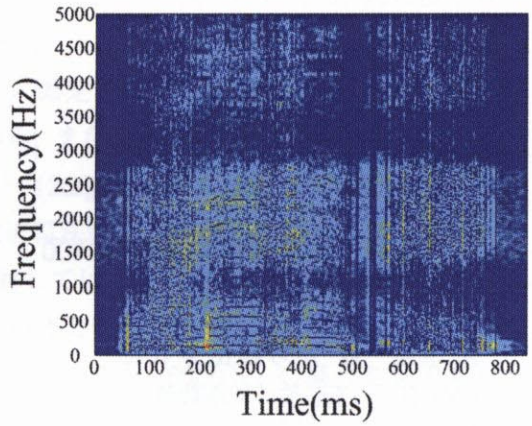


図 5.11: 提案手法による残差スペクトログラム

■実験 5-3: Bi-HBSS に対する計算量の評価 ピーク抽出の際に, 計算コストが同じになる部分のループの数をカウントすることで, 同じ問題に対する Bi-HBSS との計算量の比較を行った. 双方のプラットホームが異なるため, 計算時間の比較は無意味である. 表 5.2 にその結果を示す. ループの回数が減っていること, 抽出ピークも提案した手法のほうが多いことがわかる. 一つのピークを抽出するために必要なループ数を比べると, 劇的に計算量が減っていることがわかる.

表 5.2: ピーク抽出に伴う計算量の比較

number type	wave form subtraction	spectral subtraction
number of extracted peaks	1,398	2,951
number of loops	514,377	5,200
number of average loops per peak	367.9	1.8

5.6.2 音源定位・追跡実験

音源の定位, 追跡を行うため, 以下の 3 つの実験を行った.

実験 5-4: 周波数, 音源方向に対する定位精度・倍音統合の効果
実験 5-5: 音源数に対する定位精度
実験 5-6: IPD, IID 統合, 音源追跡の効果

■実験 5-4: 周波数, 音源方向に対する定位精度・倍音統合の効果 周波数に対する音源定位の精度を調べるため, 100 Hz から, 2000 Hz まで, 100Hz おきに純音に対する定位実験を行った. それぞれの周波数に対して, 音源の方向についても 0° から 90° まで 10° おきに測定を行った. さらに, 倍音成分の統合の効果を調べるため, 100 Hz の調波構造を有する音 (2000Hz までの倍音成分を持つ) に対しても, 音源の方向についても 0° から 90° まで 10° おきに測定を行った. 使用した部屋は 4 章で説明した残響時間 0.2 – 0.3 秒程度の部屋である. また, 音源距離は 1 m とした. 純音の定位結果を図 5.12 に示す. また, 調波構造音の定位結果を図 5.13 に示す.

図 5.12 より, 周波数が低い場合の定位精度は高いことがわかる. このことは, 式 (5.17) で距離を倍音番号 $H(i)$ で割り, 低い周波数ほど IPD による定位の信頼性を上げるようにしたことが妥当であることを裏付けている. また, 高い周波数域では, IID による判断は, 左右および正面の方向のみを判断するようになっていることから, 定位の精度が落ちてしまうことは, 避けられない. 音源方向が, 正面から離れるにつれ, 全体的な定位精度は悪くなっている. マイクロホン自体は無指向性であるが, 外装の耳部の形状によって, 前方に指向性を有していること, 音源からマイクロホンへの距離の差が大きくなり, 微妙な変化が

抽出しにくくなっていることが原因と考えられる.

純音を定位するという点においては, それほど性能が高いというわけではないといえる. これは, 人間においても純音の定位は比較的難しいという知見を裏づけするものである.

これに対し, 図 5.13 では, 各倍音での結果を統合することにより, 音源方向が正面から離れた場合でも, 比較的正確に定位を行うことができていることがわかる. つまり, 情報統合により定位のロバスト性を向上させることができたことを示している. 正面付近と, 正面から離れた角度では, やはり正面での定位精度が高くなっている. これは, 前述の理由が調波構造を有する音についても当てはまることを示している.

純音定位では, 周波数の高い部分では, 定位精度が悪いという現象が見られた. 従って, もともと定位精度の悪い高い周波数を基本周波数とするような調波構造音には, この手法でも対応は難しいと考えられる. 純音の定位結果から推定すると, 500 Hz 以下の基本周波数を持った調波構造音には, この手法は有効であると考えられる.

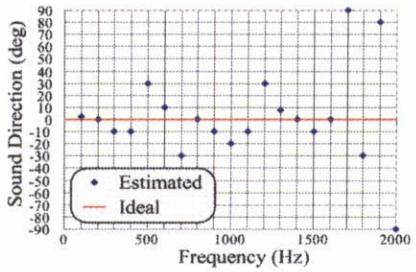
複数の音源が存在する場合については, 互いの周波数成分が影響を受け, 情報が歪んでしまい, 定位精度が悪くなることが考えられる. このような場合については, 最大音源数の検討と合わせ, 実験 5-5 で扱うものとする.

■実験 5-5: 音源数に対する定位精度 音源数, 2, 3, 4 の場合について, 音源間の角度を変えながら, 音源定位の実験を行った. 音源には, スピーカを使用し, 基本周波数成分, 倍音成分がなるべく重ならないような周波数として, 100, 111, 146, 234 Hz を選択した. 音源数 2 の場合では, 正面に 234 Hz, 左方向に 100 Hz を, 音源数 3 では, 正面に 100 Hz, 左右に同じだけ角度が離れた位置にそれぞれ 111, 146 Hz を配置した. 音源数 4 の場合は, 音源間の角度を同じに保ち, 左右が対象になるように, 左から順番に 111, 100, 146, 234 Hz となるように音源を配置した. また, どの場合も, スピーカ間の角度は, すべて等角度であり, 10 度おきに計測した, 前方向のみを対象としたため, 音源数 4 の場合は, 最大音源角度が 60 度となっている.

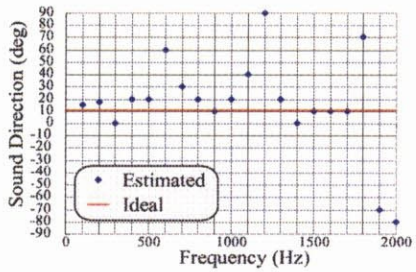
音源数 2, 3, 4 の場合の結果を図 5.14, 5.15, 5.16 に示す. 音源数が 2 の場合は良好な定位結果が得られていることがわかる. ただし, 左方向の音源 (100 Hz) は, 正面から離れるにつれ定位精度が悪くなっている. これは, 一般に 2 本のマイクロホンによる定位についてあてはまる現象であり, 詳細については, 7 章に述べる.

音源数が 3 の場合も何とか定位はうまく行えていることがわかる. ただし, 音源数が 2 の場合と比較し, 定位の精度は悪くなっている. これは, 互いの倍音成分が近いために, 位相差や強度差に影響を及ぼすことが原因である.

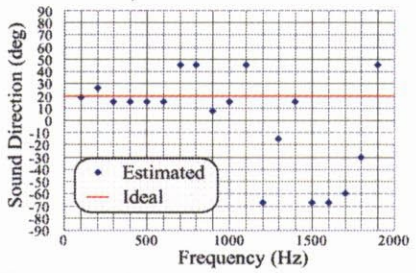
音源数 4 の場合は, 誤差は大きいが, 定位が可能であることがわかる. ただし, 4 音源となると, 10 Hz 程度の周波数分解能で, 倍音の周波数成分が重ならないように基本周波数を設定することは難しい. この実験でも, 定位のアルゴリズムより, 互いの音の影響による歪



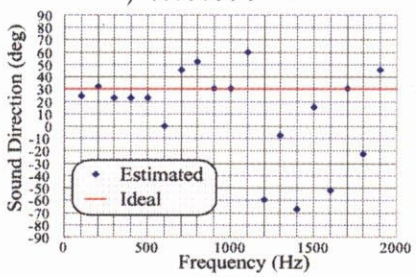
a) 音源方向 0°



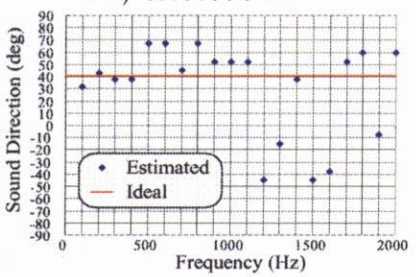
b) 音源方向 10°



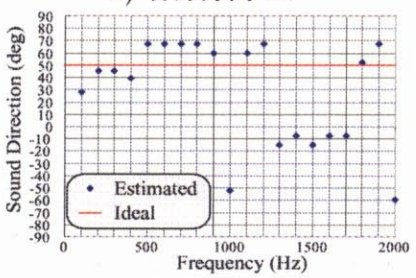
c) 音源方向 20°



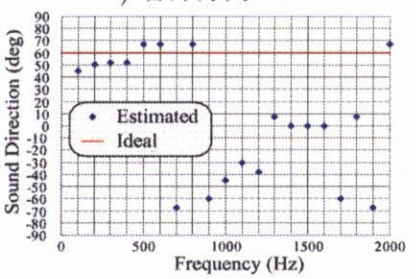
d) 音源方向 30°



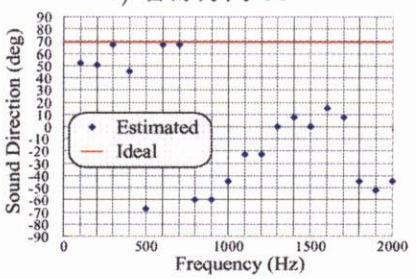
e) 音源方向 40°



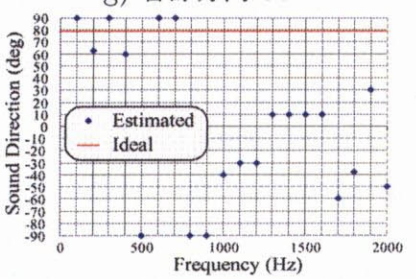
f) 音源方向 50°



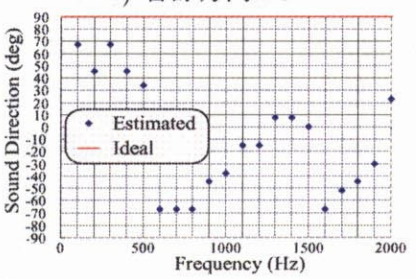
g) 音源方向 60°



h) 音源方向 70°



i) 音源方向 80°



j) 音源方向 90°

図 5.12: 音源方向, 周波数に対する定位精度

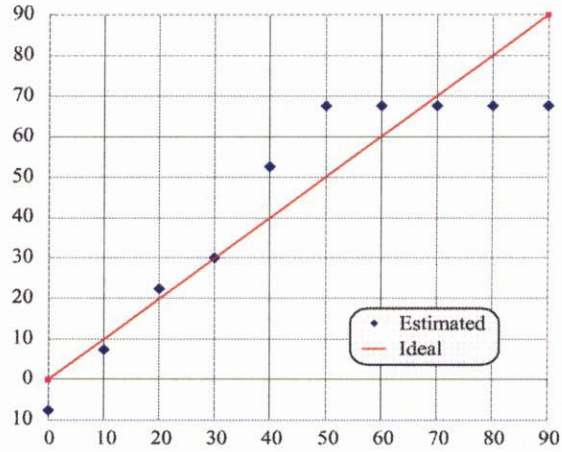


図 5.13: 音源方向に対する 100 Hz の調波構造音の定位精度

みにより定位精度が悪くなっていると言える. 5 音源以上については, 設備の関係で実験を行っていない. 2 音源の結果では, 音源間の間隔が 15 度で, 誤差が 10 度程度であることから, 近接した音源の定位分解能は 30 度程度と推定される. 従って, 原理上最大 7 音源を定位する能力があると考えられる. しかし, 実際の環境では, 周波数成分の重なり, 横方向での定位精度の悪さを考えると, 最大同時音源数は 3 から 4 音源であろう.

■実験 5-6: IPD, IID 統合, 音源追跡の効果 追跡で使用したベンチマークは以下のようにして録音した. 収録環境は, スピーカ (B&W Noutilus 805) による 2 音源を用い, 部屋は, 4 章の測定でも用いた約 10 平方メートルの音響的には, 若干 *dead* な小部屋である.

スピーカ A は, *SIG* の右 45° に置かれており, 基本周波数 234 Hz の調波音を発する. また, スピーカ B は, 基本周波数 100 Hz, 150 Hz, 200 Hz のいずれかの調波音を発する. つまり, スピーカ A と B から発せられる音は, 周波数成分の重なりがないようになっている. スピーカ A のみ, スピーカ B のみ, スピーカ A, B の両方 から音を発した場合の計 7 種類の場合についてベンチマークを作成した. 7 種類の状況に対して, *SIG* の向きを 5° おきに, ±90° の範囲で変えて *SIG* のマイクロホンで録音を行った. 各ベンチマークデータは, 録音した *SIG* の向きごとにラベリングされている.

このようなベンチマークに対して, *SIG* の初期方向は, 図 5.17 に示されるように 0°, ±45°, ±90° のいずれかとして, 説明したアルゴリズムによる音源追跡を行った.

ここでの音源の定位では, 提案した IPD と IID を統合する方法の他に, 比較のために, IPD のみを利用した場合, IID のみを利用した場合を検証した. また, 聴覚エピポーラ幾何を使用せずに, HRTF を利用して, IPD のみを利用した場合, IID のみを利用した場合, 両者を統合した場合について検証した. HRTF を利用する場合は, IID についても仮説の

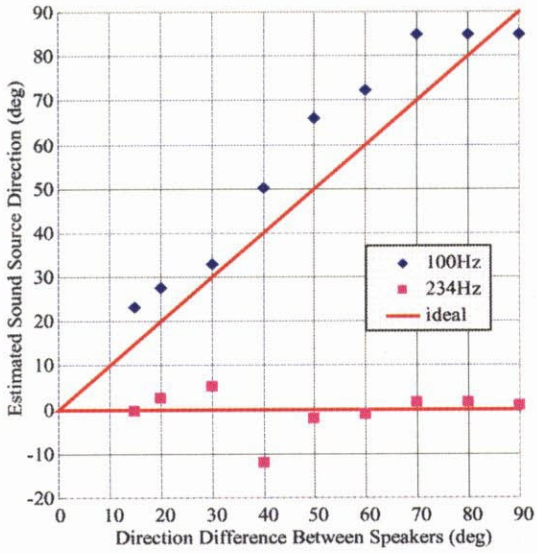


図 5.14: 同時発音数 2 の音源定位

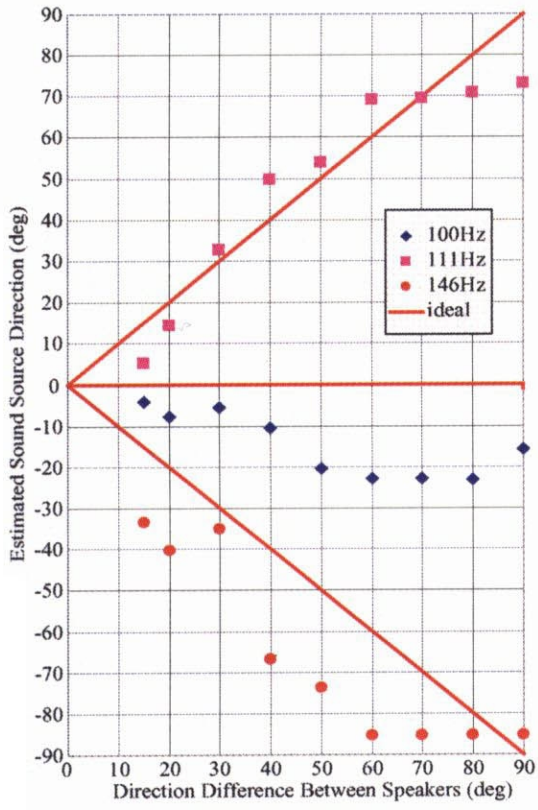


図 5.15: 同時発音数 3 の音源定位

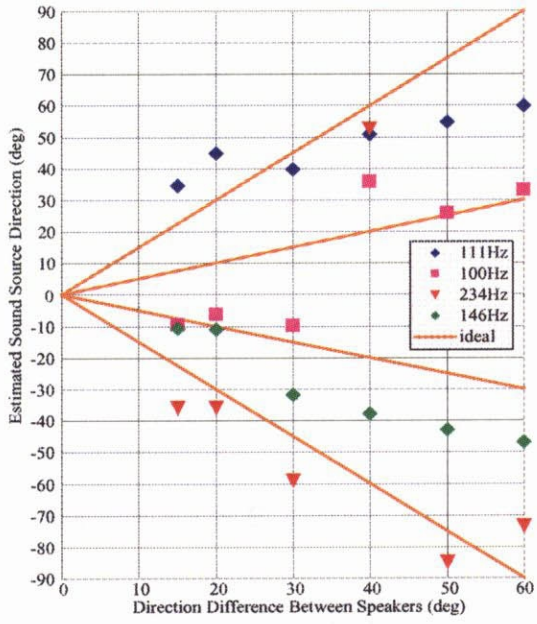


図 5.16: 同時発音数 4 の音源定位

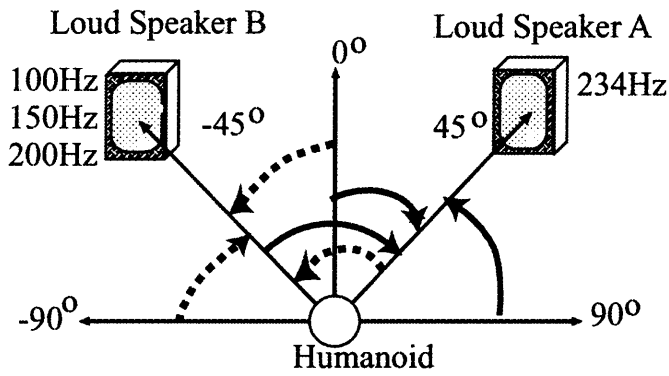


図 5.17: 音源定位・追跡実験

生成が可能であるため, IID 仮説を $I_h(\theta, f)$ とした場合, 下記の式により距離を算出し, 式 (5.22) と同じ式により, 確信度 B_{IID} を求めるものとした.

$$d(\theta) = \frac{1}{n - n_{th}} \sum_{n_{th}}^{n-1} (I_h(\theta, f) - I_s(f))^2 f) \tag{5.23}$$

このように条件を変えながら, 全 202 回の音源追跡実験を行った. このうち, 表 5.3 に示すように, 音源追跡に成功したのは 200 回であった. 失敗した 2 回は, 想定した ±90° の範囲を超えてしまった場合であった.

詳細な結果を, 表 5.4 に示す. 表の各値は, 推定された音源方向と実際の音源方向との誤差を示している. ただし, 聴覚エピポーラ幾何を利用した場合の IID のみによる定位は, 既に説明した通り, 音源の左右の判断しかできないため, 左右の判断が正しかった場合を正答として, その正答率を示している.

表 5.3: 音源追跡結果 1

Total Trials	202
Tracking Success	200
Tracking Error(divergence)	2
Tracking Error(timeout)	0

結果から, 以下のようなことがわかる.

- IID と IPD の統合により, 音源追跡が向上する. この傾向は, 音源と *SIG* の方向の差が大きいほど顕著である.

表 5.4: 音源追跡結果 2

		Initial Error			Fianl Error		
		IPD	IID	Both	IPD	IID	Both
0	Epipolar	1	100%	1	0	35	0
	HRTF	0	12	1	0	6	0
45	Epipolar	19	100%	19	3	30	3
	HRTF	-10	5	5	-3	7	-2
-45	Epipolar	-11	100%	-11	0	-5	0
	HRTF	5	-8	5	0	5	0
90	Epipolar	-10	100%	-10	3	15	3
	HRTF	-30	0	0	-5	-1	-5
-90	Epipolar	50	100%	3	5	-5	5
	HRTF	53	53	3	60	47	0

- 聴覚エピポーラ幾何による定位は, 追跡前の初期状態では, HRTF よりも精度が悪い. しかし, 音源方向を向くというアクティブな動作により, これを解決することができている.
- IPD もしくは IID の一方のみを使用した場合の音源定位も音源方向を向くというアクティブな動作により, 適切な動作が可能である.
- アクティブな動作による定位向上は, HRTF を利用する場合についてもあてはまる.

5.7 まとめ

IPD, IID, 調波構造といった複数の音響情報を統合して, 音源定位・追跡を行うシステムを示した.

正確に調波構造を抽出するためのピーク抽出に関しては, スペクトルサブトラクション法に基づく手法を示した. この手法は, 簡単な制約を有するものの, FFT の周波数解像度の不足を補い, 十分高速な動作が可能である. また, 調波構造を有しないピークの抽出も可能であり, その有効性を同様に 2 本のマイクロホンにより調波音を抽出できる Bi-HBSS と比較することにより示した.

音源の定位や追跡に関しては, 調波構造が重ならない場合に, 一般的な部屋の環境で, 同時に 2 つの音を定位することを可能にした. また, 倍音構造を利用し, IID, IPD を統合す

ることによって, 定位のロバスト性が向上することを実際にロボットに実装して示した. つまり, *SIG* の頭部や胴体, 部屋の音響効果によって歪んでしまった IPD, IID を複数の音響情報を統合することによって克服できることを示した. また, 最大同時音源数については, 各音源の角度差が 30 度以上という条件で, 3 から 4 音源であることを示した. これは, 2 章で説明した信号処理的な音源定位法のような, マイクロホン数以下の音源数しか扱うことができないという制約がないことを示している. また, 遅延和型アレーのように多くのマイクロホンを用いなくても, 感度の高い音源定位を実現できることを示している. さらに, 音源追跡実験を通じて, マイクロホンと音源の相対関係が変化する場合でも, 定位が継続できるというロボットに適した利点があることも示した. ただし, 現時点では水平方向, かつ前方向の定位に限られるため, 上下方向への対応や前後問題の解決は, 今後の課題である.

第6章

視聴覚統合による実時間人物追跡

5章では, 複数の聴覚情報を統合することにより, ロバストな音源定位・追跡を行うことができるシステムを構築し, その有効性を示した. しかし, 実環境では多くの人間がそうであるように, 聴覚情報だけで周囲の状況を把握することは難しい. そこで, 本章では, この音源定位・追跡システムを顔認識・定位を行う視覚処理と統合して, 聴覚の曖昧性を解消するだけではなく, 視覚の視野の狭さ, オクルージョンも解決できるロバストな実時間複数人物追跡システムをヒューマノイド *SIG* 上に実現する. さらに, 実時間複数人物追跡システムをステレオビジョン, 話者同定など他のセンサ情報とも統合するように拡張し, 情報統合によるロバスト性の向上について検討する.

6.1 実時間人物追跡システム

構築した実時間人物追跡システムの構成を図 6.1 に示す. システム内のモジュール群や情報は, *SIG* デバイス層, プロセス層, 特徴層, イベント層, ストリーム層の 5 つに分けられており, 図 6.1 の上にいくほど, 高次の情報や処理を扱うような階層的な枠組みを備えている. *SIG* デバイス層は, *SIG* が備えているカメラ, マイクロホン, 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, 観測時刻を付与されたイベントという形で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. ストリーム層には, 名前レベルと位置レベルのサブレイヤを有する. イベントの時間方向のシーケンスとして形成されるストリームは, その種類に応じて, 位置レベルと名前レベルのどちらか, もしくは, 両方のサブレイヤに属する.

実装上は, システムは, 「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」というセンサ情報を抽出するモジュールと, 「アソシエーショ

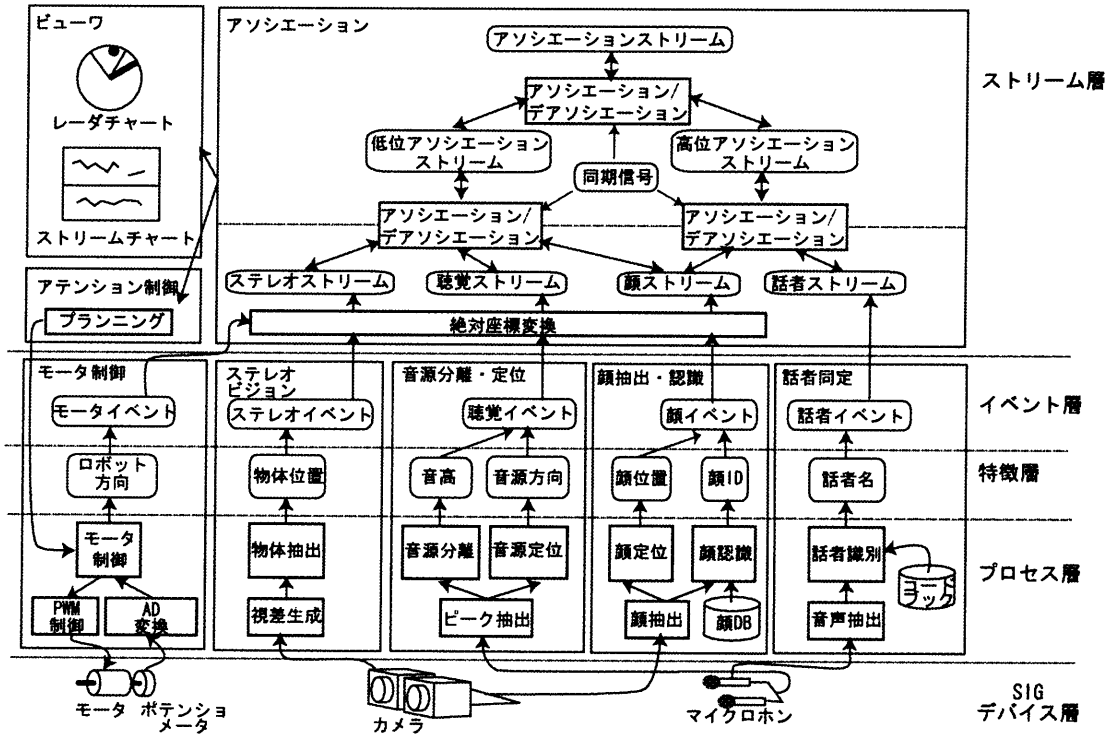


図 6.1: 実時間人物追跡システム構成図

ン」,「アテンション制御」,「ビューワ」 といったモジュールの大きく 8 つのモジュールから構成されている. 各モジュールは複数のサブモジュールから構成され, モジュールの内部およびモジュール間では様々なレベルの情報の通信が非同期に発生する. システムを実時間で動作させるために, これら 8 つのモジュールは Gigabit Ethernet で接続された 5 台の Pentium III 1GHz Linux ノードに分散されている. モジュールの配置については,「音源定位」,「顔認識・定位」,「話者同定」,「ステレオビジョン」といったセンサ情報抽出モジュールは CPU パワーを必要とするため, それぞれ別のノードに配置し, その他のモジュールは, 残りのノードに配置している. また, ノード間の情報通信は, TCP/IP によって行われるが, ノード間で正確な同期をとるために, 各ノードは NTP (Network Time Protocol) によって同期を行った上で, 各モジュールの起動時に再同期を行い, モジュール間の誤差が 100 μs 以内になるようにしている.

音源定位モジュールは, マイクロホンの入力信号に対して, 特徴抽出を行い, 観測時刻とともに, 位置, 音高 (ピッチ) 情報などを含んだ音イベントを生成し, これをアソシエーションモジュールへ送出する. 顔認識・定位モジュールは, カメラの入力信号から, 観測時刻とともに, 位置, 人名情報を含んだ顔イベントを生成する. ステレオビジョンモジュールは, 高速な視差マップ生成法を用いて, 人物のように縦に長い物体を抽出・定位し, 人物位

置情報を含んだステレオイベントを生成する．横を向くなど顔が見えない場合にも人物の位置情報を得ることができるため，システムのロバスト性を向上することができる．話者同定モジュールは，話者を同定し，観測時刻と共に，話者イベントを生成する．ただし，現在の実装では，話者数が 1 名の場合に限って有効である．また，モータ制御モジュールはモータ方向情報をイベントとしてアソシエーションモジュールへ送出すると共に，注意制御モジュールの要求に従い，PWM (Pulse Width Modulation) 信号を生成し，DC モータを駆動する．

アソシエーションモジュールは，センサ情報抽出モジュールから送られたイベントを時間的なつながりを考慮して接続し，各センサ情報ごとにストリームを生成する．ストリーム形成におけるストリームとイベントの接続には，カルマンフィルタ (Kalman filter) を用いて，観測誤差・処理誤差を低減している．また，結びつきの強いストリーム同士を結びつけ，より高次の表現であるアソシエーションストリームを生成する．逆に，アソシエーションストリームを形成するストリームの結びつきが弱くなれば，アソシエーションは解除される．また，アソシエーションの種類は，ストリームのサブレイヤにあわせて位置レベル，名前レベル，位置-名前レベル間の 3 種類のアソシエーションが存在する．

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

ビューワモジュールでは，ストリーム情報をレーダーチャートとストリームチャートの 2 種類のビューワを使って表示することができる．また各センサ情報抽出モジュールに，各モジュールで生成されるイベントを表示できるビューワを用意し，内部状態を把握しやすいインタフェースを提供している．

次節以降では，各モジュールの詳細な処理について説明する．

6.2 センサ情報抽出モジュール – イベントの抽出

上述のモジュールのうち，センサ情報を抽出するモジュールである「音源定位」，「顔認識・定位」，「ステレオビジョン」，「話者同定」について述べる．これらのモジュールは，マイクロホンやカメラといったセンサから，位置や名前情報を抽出し，観測時刻と共にイベントとして，非同期に「アソシエーション」へ送出する．ここで，イベント $\boldsymbol{E}$ を，

$$
\begin{aligned}
\boldsymbol{E} &= (\boldsymbol{e_1}, \boldsymbol{e_2}, \cdots, \boldsymbol{e_k}, \cdots, \boldsymbol{e_n}) \\
\boldsymbol{e_k} &= (\boldsymbol{G}, \boldsymbol{L}) \\
\boldsymbol{G} &= (p, t, r, \theta, \varphi, id) \\
\boldsymbol{L} &= (pitch, volume, size, \cdots)
\end{aligned}
\tag{6.1}
$$

と定義する. n は候補数, k は候補番号を示すインデックスであり, e_k はその存在確率 p が高い順に並んでいる. $\boldsymbol{G}$ はセンサ種類によらない共通のパラメータ, $\boldsymbol{L}$ はセンサに依存するパラメータである. センサ間の情報を扱う場合には, $\boldsymbol{G}$ のみを考慮する事によって, 透過的な表現が可能である. $\boldsymbol{G}$ のパラメータである t は観測時刻, r, θ, φ は極座標系でのイベント発生位置, id は, 名前情報 (顔 ID, 話者 ID) を示している. また p はイベントの存在確率である. これら $\boldsymbol{G}$ に属するパラメータのうち t, p は必ず値が存在する. t, p 以外は, 少なくともどれか一つが値をとるという制約の下, 値をとらないことが可能である. $\boldsymbol{L}$ のパラメータは, センサの種類に依存するが, 聴覚処理では音高や音量, 画像処理ではサイズなどといったパラメータである. イベント $\boldsymbol{e}$ のプロパティは, $\boldsymbol{e_1}$ の値を代表として用いるものとしている.

以降では, 各センサ情報の抽出手法について説明する.

6.2.1 音源定位・追跡モジュール

音源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, 基本的には, 5章で示したように

(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・追跡のロバスト性を向上することができるシステムを用いている. 5章で示した音源定位・追跡システムでは, Dempster-Shafer 理論によって得られる IPD と IID の両方から音源方向を支持する確信度 ($B_{\mathrm{IPD+IID}}$) を生成し, その最大の値を持つ音源方向を出力していた. しかし, 音源定位モジュールでは, 上位のモジュールでのエラー訂正を可能にするため, 各調波構造音ごとに, 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音イベントを生成し, アソシエーションモジュールへ送出するように変更を加えている. 確信度は, P_s は, 5章で定義した $B_{\mathrm{IPD+IID}}$ である. 従って, 音イベント E_s は, 候補数 n が 20 であり, 各候補 e_s は,

$$\begin{aligned} e_s &= (\boldsymbol{G_s}, \boldsymbol{L_s}) \\ \boldsymbol{G_s} &= (P_s, t, \theta) \\ \boldsymbol{L_s} &= (pitch, volume) \end{aligned} \tag{6.2}$$

と表すことができる.

6.2.2 顔認識・定位モジュール

顔認識・定位モジュールは, 複数の顔の発見・同定・定位を行い, その結果を視覚イベントとしてアソシエーションモジュールへ送信する. 顔発見サブモジュールは, 肌色抽出と相関演算に基づくパターンマッチングの組合せにより, 顔の位置, 大きさ, 明るさの動的変化にロバストな抽出を可能にしている. また, 顔が複数存在する場合でも 200 ms 以内にすべての顔領域検出が可能である [44].

個人同定サブモジュールは, 切り出された顔領域画像を判別空間に射影し, 事前登録された顔データとの距離 d を求める. d は各登録顔のクラス中心と抽出顔のマハラノビス距離として表される. この距離 d は, 登録顔数 (L) に依存するので, 式 (6.3) を用いて L に依存しない確信度 P_v に変換する.

$$P_v = \Gamma\left(\frac{1}{2}, \frac{d^2}{2}\right) = \int_{\frac{d^2}{2}}^{\infty} e^{-t} t^{\frac{L}{2}-1} dt \tag{6.3}$$

判別空間の基底となる判別行列は, オンライン LDA によって求める [45]. オンライン LDA では, 通常の LDA と比べ少ない計算で判別行列の更新が可能であり, 実時間に顔を登録することが可能である.

顔の位置同定サブモジュールは, 2 次元の画像平面上の顔位置を, 3 次元の実空間上に変換する. 具体的には, 画像平面上の顔位置を (x, y) , 大きさを $w \times w$, 探索画像の大きさを $X \times Y$ とすると, 3 次元実空間の顔位置は, 次式で与えられる距離 r, 方位角 θ, 仰角 ϕ として得ることができる.

$$r = \frac{C_1}{w},\ \theta = \sin^{-1}\left(\frac{x - \frac{X}{2}}{C_2\ r}\right),\ \phi = \sin^{-1}\left(\frac{\frac{Y}{2} - y}{C_2\ r}\right)$$

ここで C_1, C_2 は, 探索画像サイズ X, Y, カメラの画角, 実際の顔の大きさによって定義される定数である.

最終的に顔認識・定位モジュールは, 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. 従って, 顔イベント E_v は, 候補数 n が 5 であり, 各候補 e_v は,

$$\begin{aligned} e_v &= (\boldsymbol{G_v}, \boldsymbol{L_v}) \\ \boldsymbol{G_s} &= (P_v, t, r, \theta, \varphi, id) \\ \boldsymbol{L_s} &= \emptyset \end{aligned} \tag{6.4}$$

と表すことができる.

6.2.3 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 三次元位置を得る. 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 [92, 58] 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, 二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 一次元化し, その一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.

二次元の視差画像マップは以下の式により定義される.

$$DM_{2D} = \{D(i,j)|i = 1,2,\cdots W, j = 1,2,\cdots H\} \tag{6.5}$$

ここで W と H は, 幅と高さであり, D は視差値を表す.

まず, 縦長の物体を抽出するために, 高さの方向に沿って DM_{2D} のメディアンを抽出する.

$$D_l(i) = Median(D(i,j)) \tag{6.6}$$

一次元の視差マップ DM_{1D} は $D_l(i)$ の集合として作成される.

$$DM_{1D} = \{D_l(i)|i = 1,2,\cdots W\} \tag{6.7}$$

次に, 人間のような縦長のオブジェクトを抽出するために, DM_{1D} において, 視差値の近い領域を切り出す. これにより, 手を横に伸ばした場合などでも胴体部を抽出することができる. オブジェクトの定位はエピポーラ幾何によって行い, 切り出した領域の重心の座標を出力する.

最終的に, 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. ステレオモジュールでは, イベントに含まれる確信度 P_{st} は, 1.0 としている. 従って, ステレオイベント E_{st} は, 候補数 n が 1 であり, 候補 e_{st} は,

$$\begin{aligned} e_{st} &= (\boldsymbol{G_{st}}, \boldsymbol{L_{st}}) \\ \boldsymbol{G_{st}} &= (P_{st}, t, r, \theta) \\ \boldsymbol{L_{st}} &= (width) \end{aligned} \tag{6.8}$$

と表すことができる．なお,$width$ は物体幅である．

6.2.4 話者同定モジュール

話者識別は, ベクトル量子化 (VQ) 歪みに基づいた方法による話者識別が可能な Juno [4] をベースにして, ストリーム処理が可能なように改変したものを使用している．この手法では, コードブックと呼ばれる事前登録した話者データを使用する．コードブックは登録話者数を S, コードブックのクラスタ数を C とした時, $B(i,j)(i=1,...,S, j=1,...,C)$ と表される．話者識別は式 (6.9) で表される距離関数を用いて行う．入力音声を T 個のフレームに分割し, ベクトル列 $v(k)(k=1,...,T)$ に変換した後, フレーム毎に各話者との量子化歪み (ユークリッド距離) をすべてのクラスタに対して計算し, その中で最も小さくなる量子化歪みを出力する．この計算を全フレームにわたって行い, その累積和 $S(i)$ を話者 i との距離としている．

$$S(i) = \sum_k Min_j |v(k) - B(i,j)| (k=1,...,T) \tag{6.9}$$

すべてのクラスからの距離が等しいとき, 各クラスに属する確信度が 50% であるとして, $S(i)$ を 式 (6.10) によって確信度に変換する．

$$P_{sp}(i) = \frac{1}{\bar{S}\sqrt{2\pi}} e^{\frac{-S(i)^2}{2\bar{S}^2}} \tag{6.10}$$

$\bar{S}$ は $S(i)$ の平均である．話者同定モジュールは, 現時点では複数話者には対応していないが, 尤度の高い話者順に確信度を付与してアソシエーションモジュールへ出力する．話者イベント E_{sp} は, 候補数 n が 5 であり, 候補 e_{sp} は,

$$\begin{aligned} e_{sp} &= (\boldsymbol{G_{sp}}, \boldsymbol{L_{sp}}) \\ \boldsymbol{G_{sp}} &= (P_{sp}, t, id) \\ \boldsymbol{L_{sp}} &= \emptyset \end{aligned} \tag{6.11}$$

と表すことができる．

6.3 アソシエーションモジュール – 情報の統合

アソシエーションモジュールは, SIG がロバストに周囲の状況を把握するために, 様々なイベント情報を統合して, ストリーム, およびアソシエーションストリームを生成する．ストリームはイベントを時間方向に接続することによって生成され, アソシエーションストリームは, ストリーム間の状態によって発生するアソシエーションによって生成される高次のストリームである．アソシエーションのプロセスを図 6.2 に示す．

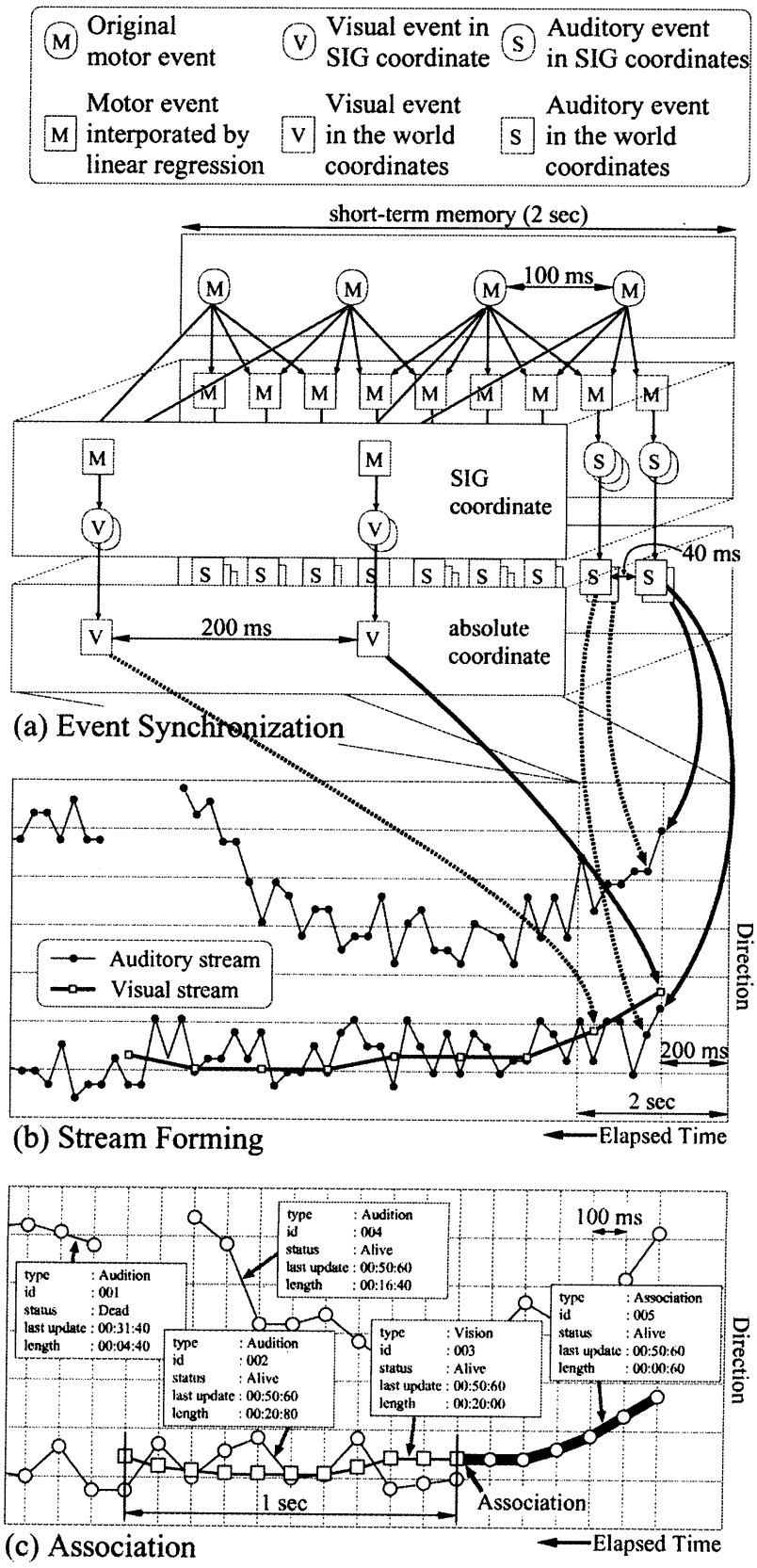


図 6.2: アソシエーションモジュールにおけるストリーム形成

6.3.1 イベントの絶対座標変換

SIG の体の向きはサーボモータのポテンショ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 ±1° である．*SIG* の方向情報を使用して，精度よくイベントの方向情報をワールド座標に変換することができる．

図 6.2a) は各モジュールから送られてくるイベントの座標を絶対座標に変換するために使用する 2 秒間の短期記憶を示している．図 6.2a) において [M], [S], [V] はそれぞれ各モジュールから非同期で送られるモータ，音，顔イベントを示している．音，顔イベントに含まれる位置情報は，イベント観測時刻のロボットから見た座標系 (*SIG* 座標系) における情報である．そこで，モータイベントを利用してこれらのイベントを絶対座標へ変換する．しかし，表 6.1 に示されるように各イベントは遅延時間，到着周期が異なる非同期イベントであるため，各イベントを一旦，2 秒間の短期記憶に格納する．まず，座標変換すべき音や顔イベントの発生時刻に近接する 2 つのモータイベントを一次線形補間することによって，該当イベント発生時刻のモータ方向を推定し，その方向，時刻情報を含んだモータイベント [M] を生成する．次に，[M] のモータ方向を利用して，該当イベントの位置情報を絶対座標系に変換し，変換後の音イベント [S]，顔イベント [V] を生成する．また，以降では，イベント E を絶対座標系に変換したイベントを E' と表記する．

表 6.1: イベント発生周期と遅延

Face Event	150 ms
Sound Event	30 ms
Speaker Event	200 ms
Stereo Vision Event	50 ms
Motor Event	100 ms
Network Latency	10 – 200 ms

6.3.2 ストリームの生成，消滅

ストリームは，図 6.2b) に示すように，短期記憶から座標変換したイベントを取り出し，時間方向に接続することによって生成される．図の横軸は経過時間，縦軸はロボットを中心にした絶対座標での水平角を示している．また，黒点は音イベント，白丸は顔イベントを示しており，これらをつなげた一本の線が一本のストリーム (太線，細線はそれぞれ顔，音ストリーム) に対応している．つまり，ストリームは，時刻 t_k に観測されたイベント

$E'(t_k)$ (G に含まれる t の値が t_k であるイベント) を用いて, 下記の式で定義できる.

$$St(t_k) = (E'(t_k), E'(t_{k-1}), \cdots, E'(t_{k-sl})) \tag{6.12}$$

ここで, sl はストリーム長を表すインデックス値であり, ストリームは, 時刻 t_{k-sl} に生成されたことを示している. つまりストリームは, 同時刻には一つのイベントのみが含まれる分岐を許さないリスト構造となっている.

また, 短期記憶からイベントを取り出す際は, ネットワークや処理のレイテンシを考慮して, 200 ms の遅延をもつようにイベントを取り出す. このため, 図 6.2b) では, 最新時刻から 200 ms の間にはストリームやイベントは存在しない.

取り出されたイベントはカルマンフィルタを用いた方法により, ストリームに接続される. カルマンフィルタはプロセスノイズや観測ノイズを考慮した予測が可能である. 特に曖昧性の大きい音ストリームに対して有効である.

イベント $E'(t_k)$ に含まれる絶対座標系での位置情報 $(r'_k, \theta'_k, \varphi'_k)$ を用いて,

$$\boldsymbol{p}_k = (r'_k, \theta'_k, \varphi'_k)$$

とする. $\boldsymbol{p}_k$ の次元 N はセンサ情報によって異なる. つまり, $\boldsymbol{p}_k$ は, 音イベントでは, r'_k, φ'_k が含まれないため, θ'_k のみの 1 次元ベクトル, 顔イベントでは, 3 次元ベクトルとなる. 次に, $\boldsymbol{p}_{k+1}$ は $\boldsymbol{p}_k$, $\boldsymbol{p}_{k-l}$ を用いて, 下記のような線形近似が可能であると仮定する.

$$\begin{aligned} \boldsymbol{p}_{k+1} &= \boldsymbol{p}_k + \boldsymbol{v}_k \Delta T \\ &= \boldsymbol{p}_k + (\boldsymbol{p}_k - \boldsymbol{p}_{k-l})/l, \end{aligned} \tag{6.13}$$

ここで l は予測に使用する履歴数である. $\boldsymbol{x}_k$ と $\boldsymbol{y}_k$ をそれぞれ, $(\boldsymbol{p}_k, \boldsymbol{p}_{k-1}, \cdots, \boldsymbol{p}_{k-l})$ として表される状態ベクトルと位置ベクトルとして表される観測値とすると, プロセスの状態と観測は, 下記の式によって表される.

$$\begin{aligned} \boldsymbol{x}_{k+1} &= F\boldsymbol{x}_k + G\boldsymbol{w}_k, \\ \boldsymbol{y}_k &= H\boldsymbol{x}_k + \boldsymbol{v}_k, \end{aligned} \tag{6.14}$$

ここで $\boldsymbol{w}_k$ と $\boldsymbol{v}_k$ はそれぞれプロセスノイズと観測ノイズであり, F, G と H は, 下記によって表される.

$$\begin{aligned} F &= \left(\begin{array}{cccc|c} \frac{l+1}{l} I_N & \mathbf{0} & \cdots & \mathbf{0} & -\frac{1}{l} I_N \\ \hline I_N & & & \mathbf{0} & \\ & \ddots & & & \mathbf{0} \\ \mathbf{0} & & & I_N & \end{array}\right), \\ G &= (\, I_N \quad \mathbf{0} \quad \cdots \quad \mathbf{0}\,)^T, \\ H &= (\, I_N \quad \mathbf{0} \quad \cdots \quad \mathbf{0}\,), \end{aligned} \tag{6.15}$$

ここで I_N は, $N \times N$ の単位行列である.

このときカルマンフィルタは下記の式によって定義される.

$$\begin{aligned}\hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k(y_k - H\hat{x}_{k|k-1}),\\ \hat{x}_{k+1|k} &= Fx_{k|k}, \end{aligned} \tag{6.16}$$

$$K_k = \hat{P}_{k|k-1}H^T(I_N + H\hat{P}_{k|k-1}H^T)^{-1}, \tag{6.17}$$

$$\begin{aligned}\hat{P}_{k|k} &= \hat{P}_{k|k-1} - K_kH\hat{P}_{k|k-1},\\ \hat{P}_{k+1|k} &= F\hat{P}_{k|k}F^T + \sigma_w^2/\sigma_v^2GG^T, \end{aligned} \tag{6.18}$$

ここで $\hat{x}$ は $\boldsymbol{x}$ の推定値, K_k はカルマンゲイン, $\hat{P}$ は誤差分散行列, σ_w^2 と σ_v^2 は, $\boldsymbol{w}_k$ と $\boldsymbol{v}_k$ の分散共分散行列である.

現在の位置ベクトルは $\hat{y}_k = H\hat{x}_{k|k}$ によって推定できる. 推定された位置ベクトルを,

$$\hat{y}_k = (\hat{r}_k, \hat{\theta}_k, \hat{\varphi}_k) \tag{6.19}$$

とする. ただし, 前述のように, 音イベントでは, $\hat{r}_k$, $\hat{\varphi}_k$) は存在しない. カルマンフィルタによって予測されたこれらの値と, 短期記憶から取り出された各イベント $E'(t_k)$ の $(r'_k, \theta'_k, \varphi'_k)$ を照合し, そのイベントが下記の条件を満たしている場合, イベントとストリームの接続を行う.

- **音イベント:** 音イベントは, 音高が, 同等もしくは倍音関係にあり, ストリームとの距離

$$d_s = \left|\hat{\theta}_k - \theta_k\right| \tag{6.20}$$

  が ±10° 以内で最も近い既存の音ストリームに接続される. この値は, 聴覚エピポーラ幾何の精度を考慮し定めた値である. また, 倍音関係を許容するのは, 基音の抽出誤りを補完するためである.

- **顔イベント:** 顔イベントは, 共通の顔 ID を持ち, ストリームとの距離

$$\begin{aligned} d_f = \Bigg(&\left(\hat{r}\cos\hat{\theta}_k\cos\hat{\varphi}_k - r\cos\theta_k\cos\varphi_k\right)^2\\ &+ \left(\hat{r}\cos\hat{\theta}_k\sin\hat{\varphi}_k - r\cos\theta_k\sin\varphi_k\right)^2\\ &+ \left(\hat{r}\sin\hat{\theta}_k - r\sin\theta_k\right)^2\Bigg)^{\frac{1}{2}} \end{aligned} \tag{6.21}$$

  が, 40 cm の範囲内で最も近い既存の顔ストリームに接続される. この値は, 4 m/s 以上で人間が移動しないことを前提にして定めている.

- **話者イベント:** 話者イベントは複数話者に対応していないため, 同時刻には高々 1 つのイベントしか発生しない. 従って, 話者ストリームが存在していれば 無条件に話者イベントを接続する.

- **ステレオイベント:** ステレオイベントは, ストリームとの距離

$$d_{st} = \left(\left(\hat{r} \cos \hat{\theta} - r \cos \theta_k \right)^2 + \left(\hat{r} \sin \hat{\theta}_k - r \sin \theta_k \right)^2 \right)^{\frac{1}{2}} \tag{6.22}$$

が 40 cm の範囲内で最も近い既存のステレオストリームに接続される. この値は, 顔イベントと同様の基準である.

この結果, 音ストリーム St_s, 顔ストリーム St_v, 話者ストリーム St_{sp}, ステレオストリーム St_{st} が生成される. すべての既存ストリームに対して探索を行った結果, ストリームと接続できないイベントが存在した場合, そのイベントから新規にストリーム, つまり, St に E が一つだけ含まれるリストによって構成されるストリームが生成される. ただし, 顔イベントに関しては, 既に存在している顔ストリームの顔 ID と同じ顔 ID を持ったストリームは生成せず, 第二候補以降の顔 ID を持ったストリームが生成される. これは, ステレオ, 話者イベントについても同じである. また, 既存ストリームは, 接続するイベントが全く存在しない場合でも, 最大 500 ms 間は存続が可能であるが, 500 ms 以上イベントが接続されない場合, そのストリームは消滅する.

このような時間の流れを考慮したストリーム形成の利点は以下の通りである.

- 音 : 基本周波数の取得失敗が訂正され得る
- 顔 : 名前が異なってもイベント間のユークリッド距離が近い場合は同一ストリームと見なせるため, 名前の誤りが訂正される. またストリーム全体にわたって名前情報をチェックすることにより, ストリームを代表する名前の誤りを訂正できる. 例えば, ストリーム生成時に ID が間違っていても, 時間の経過とともに訂正され得る.
- 話者 : 話者は同一ストリームであれば処理の性質上後からやってくるイベントの方が確信度が高くなる. このため, ストリーム生成時に間違っていた話者名が, 時間の経過に伴い訂正されてより正確になる.
- ステレオ : 物体 (人物) の動きを把握できるようになる (アソシエーション時に有効).

6.3.3 ストリームのアソシエーション

次に, ストリームのアソシエーション判定の前処理として, ストリーム間同期を行う. 具体的には, 図 6.2b) におけるストリームに対し, 100 ms 周期で再サンプリングを行う. サ

ンプリング時刻のストリーム情報は, サンプリング時刻に近接するストリーム内の 2 つのイベント情報を一次線形補間して得ている. サンプリング時刻を t_i とし, $t_k > t_i > t_{k-1}$ という条件を満たす場合,

$$E'(t_i) = \frac{t_k - t_i}{t_k - t_{k-1}} E'(t_{k-1}) + \frac{t_i - t_{k-1}}{t_k - t_{k-1}} E'(t_k) \tag{6.23}$$

として, 再サンプルしたイベント $E'(t_i)$ を計算する. これにより, 再サンプル後のストリーム St' は,

$$St'(t_i) = (E'(t_i), E'(t_{i-1}), \cdots, E'(t_{i-sl'})) \tag{6.24}$$

と時間的に等間隔に並んだイベントのリストとして表すことができる. 図 6.2c) の□, ○は 100 ms 周期で再サンプリングしたストリーム中の顔, 音イベントを示している. この処理により, ストリーム間距離の算出処理を簡潔にしている.

同期後, ストリーム間距離を算出し, 同一の人物に由来すると判断される複数のストリームをアソシエーションし, 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. つまり, アソシエーションストリーム St_A は,

$$St_A = \{\boldsymbol{St} | \boldsymbol{St} \subseteq \{St_s, St_v, St_{sp}, St_{st}\}\} \tag{6.25}$$

として表現される. 図 6.2c) のケースでは音ストリーム, 顔ストリームがアソシエーションされ, アソシエーションストリームが形成されたことを示している.

アソシエーションストリームを形成するストリームの一つが消滅した場合, アソシエーションストリームから, そのストリームは取り除かれる. また, 誤ったアソシエーションであると判断された場合, デアソシエーションにより複数本のストリームに分割される.

実際の, ストリームのアソシエーションメカニズムでは, 階層的な枠組みを導入している. 図 6.3 に示すように, 4 層に分けられた視聴覚情報が, 相互に影響し合うような枠組みとなっている. 本研究では, 位置レベルと名前レベルという上位の 2 層を対象としている.

本研究で扱っているストリームは, 自身を構成するイベントの種類に応じて, 位置レベル, 名前レベルのどちらか, もしくは両方に属するものとする. 参考のため, ストリーム, イベントに含まれる特徴量をまとめたものを表 6.2 に示す. 位置情報は音, 顔, ステレオビジョンストリームに含まれ, 名前情報は話者, 顔ストリームに含まれることがわかる. 顔ストリームに関しては位置, 名前の両方の情報を含んでいるため, 両方に属すものとする. このようにストリームを階層的にカテゴライズした後, 図 6.3 に示すように, 位置レベルストリーム間, 名前レベルストリーム間, および位置と名前ストリームにまたがった階層的な 3 段階のアソシエーションを行っている.

■位置レベルのアソシエーション 位置情報を含むストリームは, 一定時間以上距離が近いストリームが存在した場合, アソシエーションを行う. 具体的な条件は, 1 秒間のうち

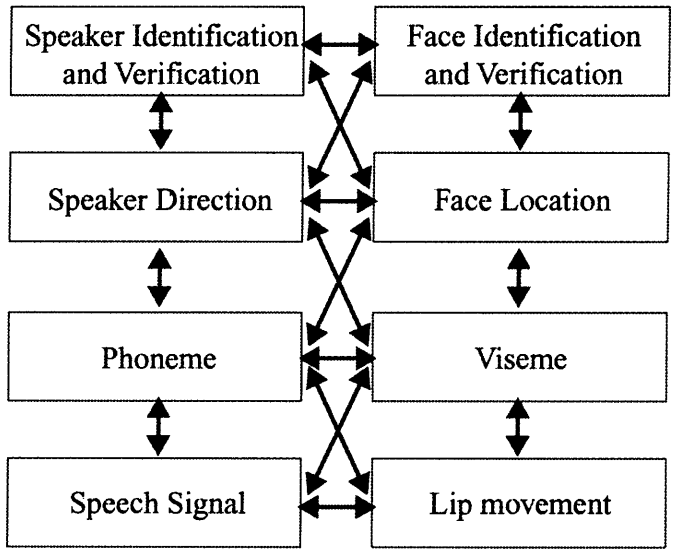


図 6.3: ストリームの階層性

表 6.2: イベント・ストリームと特徴量の関係

Level	Stream	Event	Feature
Location Level	Sound Stream	Sound Event	Sound direction Pitch
	Stereo Stream	Stereo Event	Object location by Stereo
	Motor Stream	Motor Event	Motor direction
	Face Stream	Face Event	Face location
Name Level	Face Stream	Face Event	Face ID
	Speaker Stream	Speaker Event	Speaker ID

500 ミリ秒以上, 二つのストリームが近接していると判断されることであるつまり, 二つの異種のストリームを St_1, St_2 に対して, 下式の $S(St_1, St_2) \geq 5$ となる場合, アソシエーションが行われる.

$$S(St_1, St_2) = \sum_{j=i-10}^{i} s(E_1'(t_j), E_2'(t_j)) \tag{6.26}$$

$$s(E_1'(t_j), E_2'(t_j)) = \begin{cases} 1 & if \quad d(E_1'(t_j), E_2'(t_j)) \leq th \\ 0 & if \quad d(E_1'(t_j), E_2'(t_j)) > th \end{cases}$$

ここで, $E'_1(t), E'_2(t)$ は, St_1, St_2 のサンプリング時刻 t におけるイベント情報である, d は, 2 つのイベント間の距離関数であり, th は近接しているかどうかを判断するための閾値である. これらの値は, ストリームの種類によって異なる.

表 6.3: ストリームに含まれる位置情報

	Sound	Face	Stereo
Distance(r)		○	○
Azimuth(θ)	○	○	○
Elevation(ϕ)		○	

各ストリームは表 6.3 に含まれる位置情報を含んでいるが, 同じ情報であっても抽出方法が異なるため, その精度が異なる. 方位角はすべてのストリームに含まれるため, 距離の算出の際に有効であるが, その精度はステレオストリームに含まれるものが最も高く, ついで顔ストリーム, 音ストリームとなっている. 音ストリームは, 視覚ベースの位置情報と比較すると精度が低いものの, 複数候補を確信度付きで持っているため, 複数の候補が利用できる. また, 距離 (r) の抽出精度は, ステレオストリームの方が顔ストリームよりも高い. このため, d, th は以下のように定義している. なお, 下記では, イベント E の距離 r, 方位角 θ, 仰角 φ を, それぞれ, $r(E)$, $\theta(E)$, $\varphi(E)$ と記述するものとする.

音 - 顔 距離関数 d は,

$$d(E'_s(t_i), E'_v(t_i)) = |\theta(E'_s(t_i)) - \theta(E'_v(t_i))| \tag{6.27}$$

と定義した. また, $th = 10°$ であり, $S(E'_s(t_i), E'_v(t_i))$ は, $d(E'_s(t_i), E'_v(t_i)) \leq 10°$ の時に 1, それ以外では 0 をとるものとしている. つまり, 方位角が $\pm 10°$ 以内に近接する状況が 1 秒間のうち 500 ミリ秒以上観測される場合にアソシエーションが行われる.

顔 - ステレオ 距離関数 d は,

$$\begin{aligned} d(E'_v(t_i), E'_{st}(t_i)) = & \\ & \Big((r(E'_v(t_i)) \cos(\theta(E'_v(t_i))) - r(E'_{st}(t_i)) \cos(\theta(E'_{st}(t_i))))^2 \\ & + (r(E'_v(t_i)) \sin(\theta(E'_v(t_i))) - r(E'_{st}(t_i)) \sin(\theta(E'_{st}(t_i))))^2 \Big)^{\frac{1}{2}} \end{aligned} \tag{6.28}$$

と定義した. また, $th = 10\,cm$ と定義した. つまり, ユークリッド距離で 10cm 以内に近接する状況が 1 秒間のうち 500 ミリ秒以上観測される場合にアソシエーションが行われる.

ステレオ - 音 距離関数 d は,

$$d(E'_{st}(t_i), E'_s(t_i)) = |\theta(E'_{st}(t_i)) - \theta(E'_s(t_i))| \tag{6.29}$$

と定義した. また, $th = 10°$ である. つまり, 方位角が $\pm 10°$ 以内に近接する状況が 1 秒間のうち 500 ミリ秒以上観測される場合にアソシエーションが行われる.

また, アソシエーションストリームに含まれる各ストリームの位置情報が 3 秒間以上に渡り, 水平角の差で 30° 以上になった場合, アソシエーションを解除するものとした. これらの値は実験的に求めた.

アソシエーションストリームの時刻 t における位置情報 p_a は, アソシエーションストリームを構成する音, 顔, ステレオストリームの時刻 t における位置情報 p_s, p_v, p_{st} を用いて, 以下の式により求めることができる.

$$p_a = p_s \cdot B_s + p_v \cdot B_v + p_{st} \cdot B_{st} \tag{6.30}$$

$$B_s = \frac{e_s}{e_s + e_v + e_{st}}$$

$$B_v = \frac{e_v}{e_s + e_v + e_{st}}$$

$$B_{st} = \frac{e_{st}}{e_s + e_v + e_{st}}$$

ここで, e_s, e_v, e_{st} は, 音情報, 顔情報, ステレオ情報の重み付けをするパラメータであり, 0 から 1 までの値をとることができる. 本研究では, (e_s, e_v, e_{st}) は, 本研究では, ステレオストリームが存在すれば, $(0,0,1)$, ステレオストリームが存在せず, 顔ストリームが存在すれば, $(0,1,0)$, 音ストリームしか存在しなければ, $(1,0,0)$ としている. 一時的にはアソシエーションストリームに一本のストリームしか存在しない場合もありうるので, 音ストリームしか存在しない場合も考慮に入れている. 現時点では, このように, 0 か 1 を設定しているので, 最も定位精度の高いストリームの値を統合したストリームの方向情報として採用している. しかし, e_s, e_v, e_{st} を, それぞれ音源定位, 顔認識・定位, ステレオビジョンモジュールにおける定位誤差とすることも可能である. 距離については, ステレオビジョンと顔ストリーム間で同様の処理を行っている. 仰角は, 顔ストリームが存在していれば, 検出可能であるが, その他の場合は, 利用不可となっている.

最も定位精度の高いストリームの値を採用しているので, 統合を行わなくても, 精度の高い視覚情報だけ利用すれば, 十分であるように考えられるが, 実際には, 情報を互いに補い合うことで, そのロバスト性を高めている (disambiguation) . 例えば, 視野外にいる人からの声, 物陰に隠れている人からの声は, 視覚だけでは定位することができない. また, 顔認識・定位モジュールでは壁にかけてある写真を, 人物として抽出してしまうなどの誤

抽出があり, アソシエーションによってこのようなエラーを訂正することができる. さらに, 方位角しかわからない音ストリームであっても, 視覚ベースのストリームとアソシエーションできれば, 正確な方位角が取得できるばかりか, 距離, 仰角, 顔 ID といった情報を取得することができる.

■名前レベルのアソシエーション 顔ストリーム, 話者ストリームは名前情報を含んでいる. アソシエーションは両ストリームの話者名と顔 ID が 1 秒間のうち 500 ミリ秒以上一致した場合に行われる. イベント E の名前情報 id を $id(E)$ と表せば, $S(St_v, St_{sp}) > 5$ の場合, St_v と St_{sp} がアソシエーションする.

$$S(St_v, St_{sp}) = \sum_{j=i-10}^{i} s(E_v(t_j), E_{sp}(t_j)) \tag{6.31}$$

$$s(E_v(t_j), E_{sp}(t_j)) = \begin{cases} 1 & if \quad id(E_v(t_j) = id(E_{sp}(t_j)) \\ 0 & if \quad id(E_v(t_j) \neq id(E_{sp}(t_j)) \end{cases}$$

また, 3 秒間以上に渡り, 名前情報が一致しなかった場合, 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る. 名前レベルのアソシエーションにより, 一方の情報が欠けている場合でも名前情報を継続的に保持することができる.

■位置 – 名前レベル間のアソシエーション 位置 – 名前レベル間についても, 視覚ベースの位置情報と名前情報, および聴覚ベースの位置情報と名前情報についてアソシエーションを行う. 前者については, 顔ストリームが, 位置, 名前情報の両方を含んでいるため顔ストリーム内において, 既に達成されている. これは, 顔ストリームでは, 顔の位置と名前は常に同時に, 顔抽出・認識モジュールで抽出されるためである. 後者については, 現時点では話者ストリームが複数話者に対応していないため, 以下のルールに基づいてアソシエーションを行う.

1. 音ストリームが一つである場合はそのストリームと話者ストリームをアソシエーションする.
2. 音ストリームが複数ある場合は, 話者ストリームの開始時刻と近い方のストリームをアソシエーションする.

また, アソシエーションモジュールでは, 矛盾を防ぐ手段として, 以下のような仮定を用いている.

- アソシエーションされていないストリーム同士は, 種類が異なる場合のみアソシエーション可能である.
- アソシエーションされたストリームは, そのアソシエーションストリームに属しているストリームと同じ種類のストリームを含んでいない任意のストリーム, および,

アソシエーションストリームとアソシエーション可能である.

上述の制約を用いても, 矛盾が発生する場合, 基本的にそのような矛盾のあるアソシエーションは行わない. ただし, そのような状況が, 一定時間以上継続した場合には, アソシエーションの誤り, もしくは, ストリーム生成の誤りが生じている可能性が考えられる. 前者については, 最尤のアソシエーション状態になるようにアソシエーションストリームの再構築を行うことにより対処を行っている. 後者については今後の課題としている.

最終的に, アソシエーションモジュールは, 把握している状況をストリームおよびそのアソシエーション状態として保持し, アテンション制御およびビューワモジュールの問い合わせに応じて, ストリームの情報を送信する.

6.4 注意制御モジュール – 視聴覚サーボ

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

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

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

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

A) アソシエーションストリームの存在は, *SIG* に正対して喋っている人物が現在も存在している, もしくは近い過去に存在していたことを示している. 従って, 一般にそのような人物に対して, 高い優先度でアテンションを向け, 追跡を行うのは妥当である.

B) マイクロホンは無指向であるためカメラのような視野角は存在せず, 広範囲な情報を得ることができる. このため, 顔より音ストリーム優先度を高くすべきである.

6.5 ビューワモジュール – システムの視覚化

多感覚入力のロボットに対して, 様々な情報を実時間で直感的な視覚化を行うことは, ロボットの内部状態を把握する意味から重要である. 実時間人物追跡システムでは, 音, 顔, ステレオ, 話者, モータ, ストリーム情報それぞれに対し, このような要件を満たした視覚化を実現している.

音情報 音情報のビューワを 図 6.4 に示す. ビューワの左側のウインドウにはスペクトルと抽出したピークが表示され, 右側のウインドウには, 縦軸を *SIG* から見た相対的な方位角, 横軸をピッチとして, 音イベントが表示される. 音イベントは円で表現され, 円の直径は音源定位の確信度を表している.

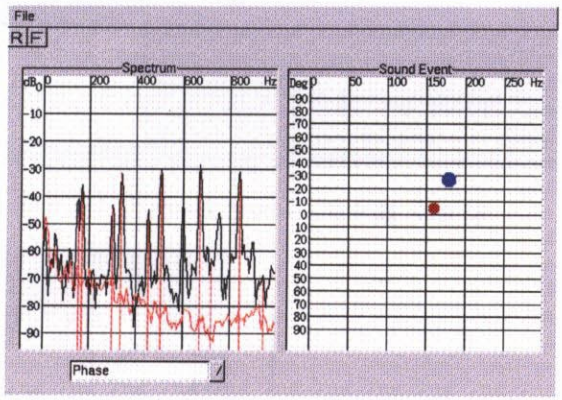


図 6.4: 音モジュールビューワ

顔情報 図 6.5 に示されるように, 抽出した顔は長方形の枠で囲まれ, 確信度付で抽出した顔の名前と位置のリストが表示される. 複数の顔が発見されれば, それぞれの顔の同定と定位の結果が表示される.

ステレオ情報 図 6.6 に示されるように, 左右の画像からのイメージ, および, 視差画像, 視差画像から抽出されたヒストグラムが表示される. ヒストグラムから, 人物の位置が抽出できる.

話者情報 図 6.7 に示されるように, 音声情報から, 話者を同定し, その名前, 登録された話者情報との距離, 確信度を表示する.

モータ情報 図 6.8 に示されるモータ情報のビューワは, ロボットの向きや動作の速度を実時間に三次元表示する事ができる. この三次元ビューワは OpenGL によって実装されている.

ストリーム情報 ストリーム情報については, 図 6.9 に示すように, レーダチャートおよ

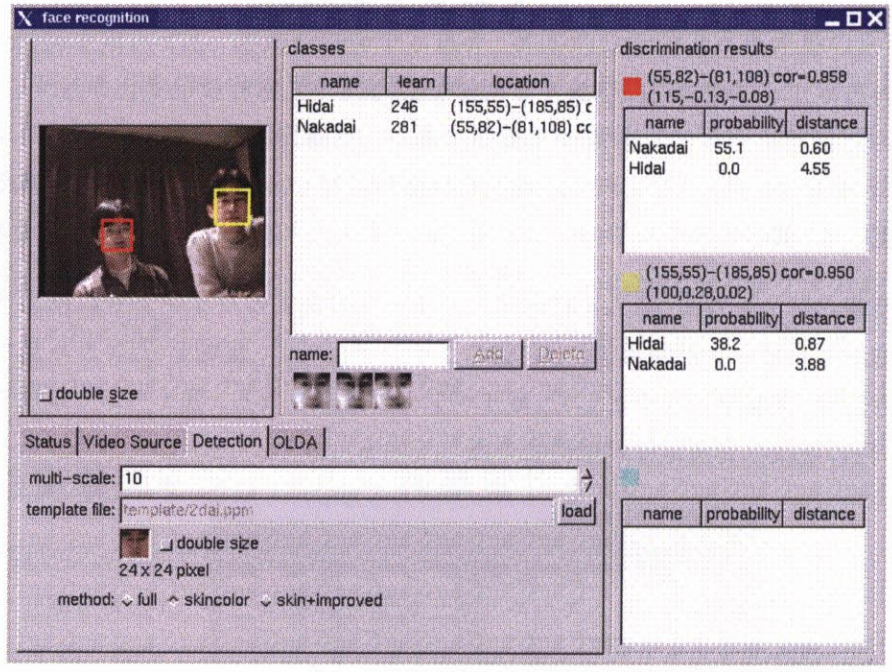


図 6.5: 顔モジュールビューワ

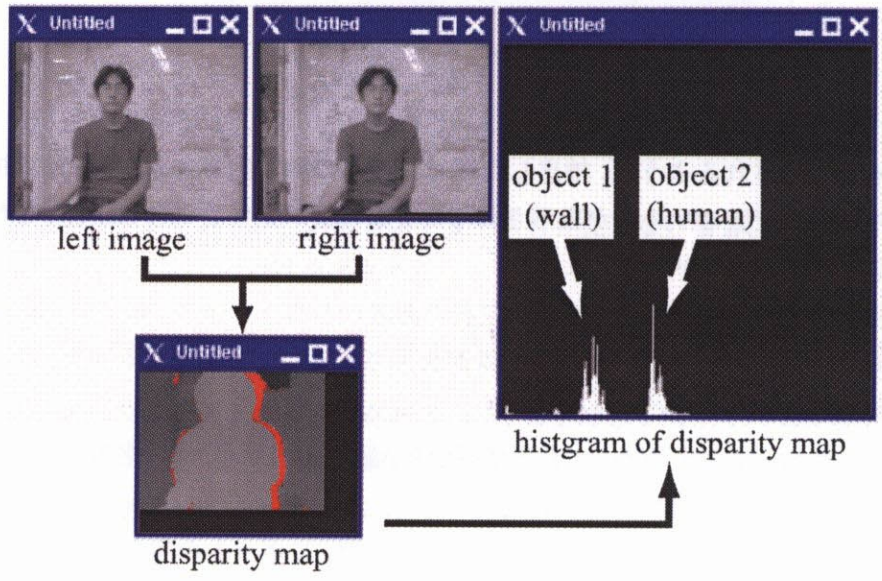


図 6.6: ステレオビジョンモジュールビューワ

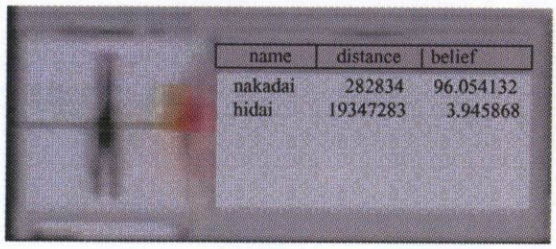


図 6.7: 話者同定モジュールビューワ

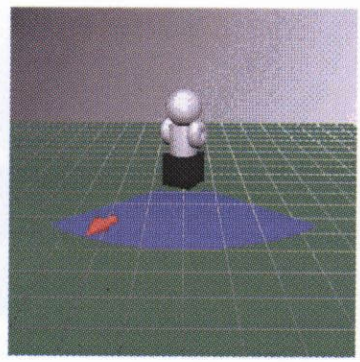


図 6.8: モータモジュールビューワ

びストリームチャートによって視覚化を行っている．レーダーチャートでは，その瞬間におけるストリームの状態を示し，ストリームチャートでは，広く明るい扇型と狭く暗い扇型は，それぞれ，カメラ視野と音源方向を示している．ストリームチャートにおいて，太線はアソシエーションストリームを示し，細線は，音もしくは顔ストリームを示している．

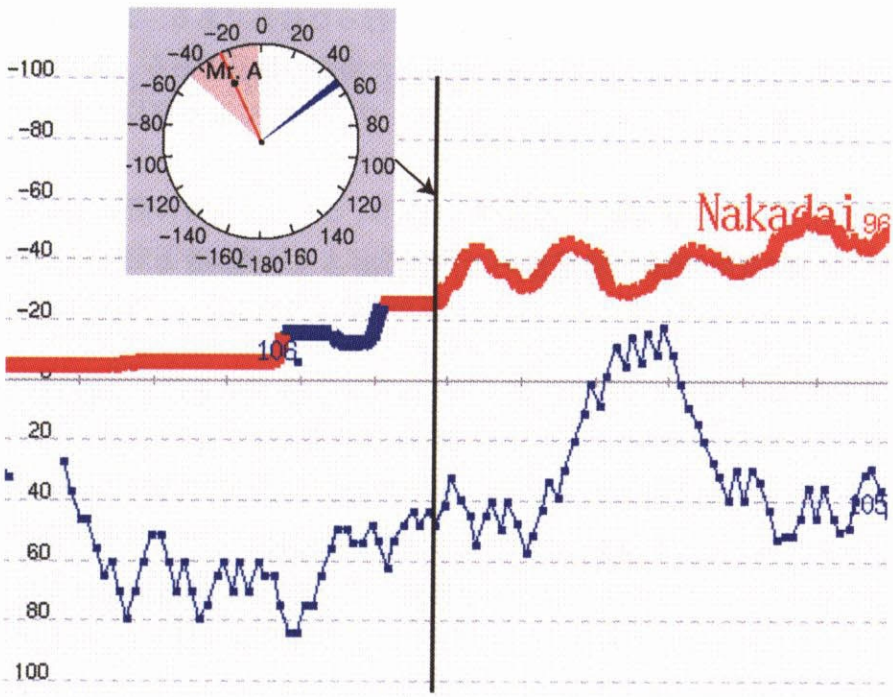


図 6.9: ストリームビューワ

6.6 実時間人物追跡システムの評価

構築したシステムを用いて, センサ情報数の追加に対するロバスト性の向上, 話者数の増加に対するロバスト性の向上を評価するため, 3 つの実験を行った.

実験 6-1: 1 話者による音源定位と顔認識・定位の統合実験
実験 6-2: 2 話者による音源定位と顔認識・定位の統合実験
実験 6-3: 2 話者ですべてのセンサ情報を統合した実験

6.6.1 実験 6-1: 1 話者による音源定位と顔認識・定位の統合

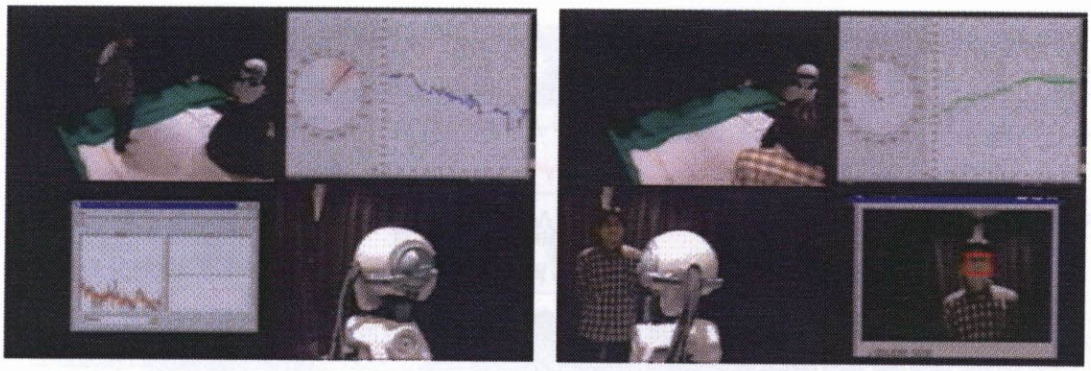
視聴覚の統合の効果を示すため, 音源定位モジュールと顔認識・定位モジュールを利用して, 人物追跡の実験を行った. 実験環境は, 4 章の測定で使用した, $3m \times 3m$ の残響時間が 0.2 ms – 0.3 ms 程度の部屋である. *SIG* と約 1 m 程度のところに話者がおり, 話者は *SIG* に顔を見せて, 呼びかけながら, *SIG* の前を移動する.

図 6.10 a) は, 音源定位モジュールのみを利用して, 追跡を行った結果であり, 図 6.10 b) は, 顔認識・定位モジュールのみを利用して, 追跡を行った結果である. また, 図 6.11 は, 音源定位と顔認識・定位モジュールの双方を利用した場合の結果である. 各図は, 4 分割されており, 左上が俯瞰図, 右上がストリームビューワ, 左下が, 音源定位ビューワ, 右下が顔定位ビューワとなっている.

まず, 図 6.10 a), b) より, 音源定位, 顔認識・定位は, どちらも追跡に適切な精度を持っていることがわかる. また, 両者を比較すると, 顔定位による追跡は音源定位による追跡と比べてスムーズであることがわかる. これは, 音源が視認できるときは, 視覚情報を使うことにより, 定位の精度が高くなることを示している.

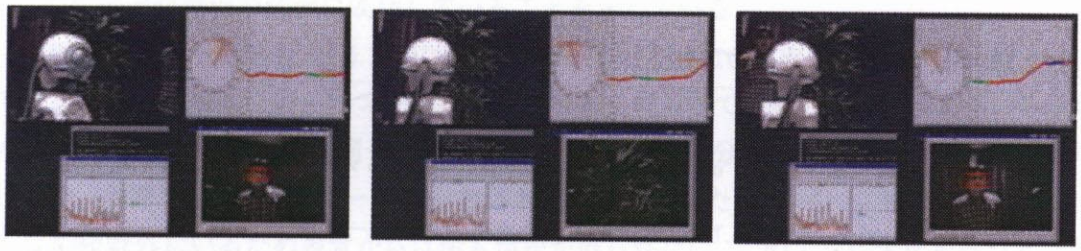
図 6.11 a) では, 顔を見せながら, 発声をしているため顔ストリームと音ストリームがアソシエーションされ, アソシエーションストリームが形成されている. この状態で, 図 6.11 b) のように顔が隠れてしまうと, 顔ストリームが一時的に消えて, 音情報だけのアソシエーションストリームとなる. このような状態が長く続くとアソシエーションは解除される. この場合では, アソシエーションが継続しているため, 顔情報が得られなくても, 音情報によって追跡が続けられる. 図 6.11 c) では, 顔が再び発見され, 音と顔の両方の情報を含んだアソシエーションストリームとなる.

このように, アソシエーションにより, 欠けてしまった顔情報を補うことが可能である. 音が一時的に消えてしまう場合も顔の定位情報によりこれを補うことができるので, お互いに情報を補い合うことができる.



a) 「音源定位」による追跡　　b) 「顔認識・定位」による追跡

図 6.10: 「音源定位」, もしくは「顔認識・定位」の一方を用いた人物追跡



a) アソシエーションストリームが作られる.　b) 顔が隠れる.　c) 顔が再び現れる.

図 6.11: 「音源定位」,「顔認識・定位」を統合した人物追跡

6.6.2 実験 6-2: 2 話者による音源定位と顔認識・定位の統合

本システムの評価を行うため, 図 6.12 に示すシナリオをベンチマークとして使用した. このシナリオでは, 2 話者が約 40 秒に渡って様々なアクションを行う. 具体的な 2 話者のアクションを以下に示す. なお, 図 6.12 の Radar Chart, Stream Chart の方向値は絶対座標系での SIG の水平角を示し, Radar Chart, Robot View の t_n は Stream Chart の対応する時刻を示す.

t_1: A 氏が *SIG* の視野内に入る.

t_2: A 氏が *SIG* に話を始める.

t_3: B 氏が *SIG* の視野外で話を始める.

t_4: A 氏が動き, 物陰に隠れる.

t_5: A 氏が再び, 物陰から現れる. その後, A 氏は話を止め再び物陰に隠れる.

t_7: *SIG* が話をしている B 氏の方向を向く.

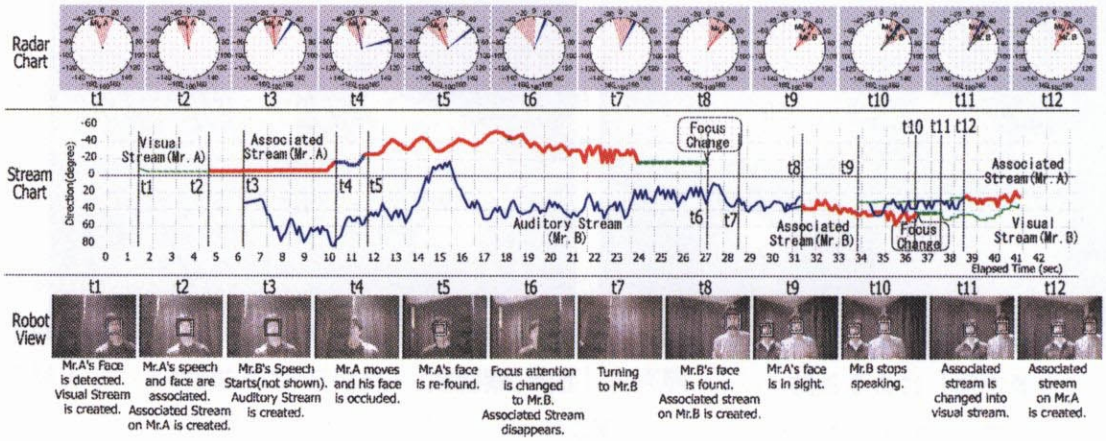


図 6.12: 「音源定位」,「顔認識・定位」を統合した 2 話者追跡における時間シーケンス

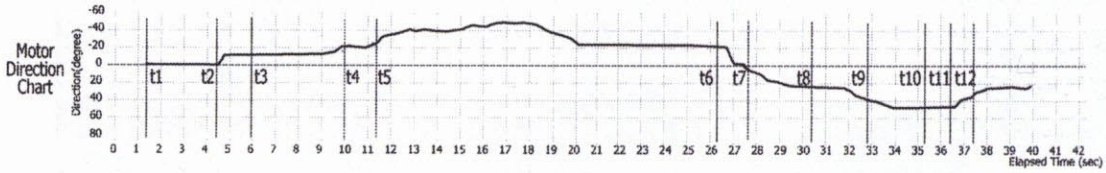


図 6.13: 図 6.16 におけるモータ方向の時間シーケンス

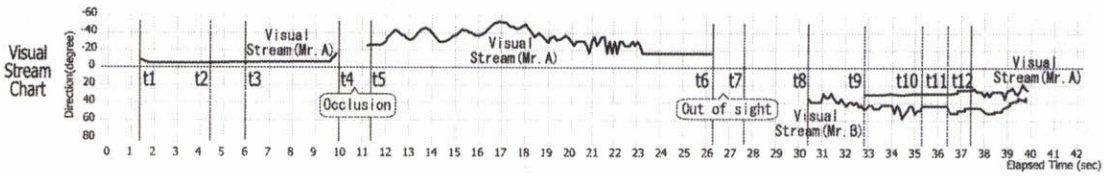


図 6.14: 図 6.16 において「顔認識・定位」のみで追跡を行った場合の時間シーケンス

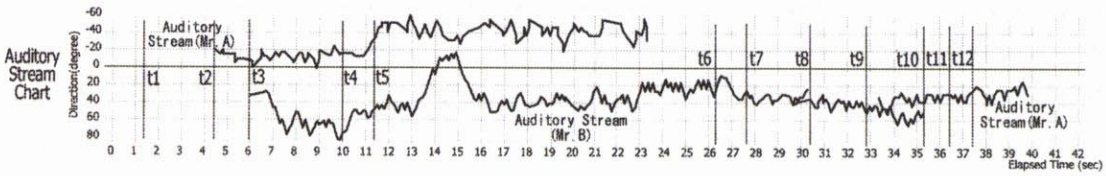


図 6.15: 図 6.16 において「音源定位」のみで追跡を行った場合の時間シーケンス

t_8: SIG が B 氏を視野内に捉える.

t_9: A 氏も話しながら SIG の視野内に入ってくる.

t_{10}: B 氏が話を止める.

図 6.12 より, このシナリオに対してシステムは以下のような特徴をもった動作を行った.

1. 新しいアソシエーションストリームが生成されると優先的に SIG の注意が新アソ

シエーションストリームに向けられる [図 6.16 の t_1 と t_8].

2. オクルージョンにより, アソシエーションストリームの視覚情報が欠如してしまったが, アソシエーションが行われているため, 聴覚情報により追跡が継続できる [図 6.12 の t_4, t_5 間].
3. アソシエーションストリームが消滅したため, アソシエーションストリームの次に優先度の高い音ストリームに注意が向けられる [図 6.12 の t_6, t_{11}].
4. シナリオの 26 s 以降は, 2 話者は同時にカメラの視野に移る程度 (約 20°) まで近づく. この場合でも, うまく話者の追跡が達成されている.

このシナリオにおける *SIG* の体の方向を図 6.13 に表す. 図 6.13 は, 2 話者の場合においても, 注意制御モジュールおよびモータ制御モジュールが, 適切な PWM 信号を生成し, *SIG* の追跡動作の制御に成功していることを示している.

視覚情報による追跡を 図 6.18 に示す. 図 6.18 は, 図 6.16 の実験の際に, 顔認識・定位モジュールが生成した顔イベントの第一候補のみを使用して作成したものである. そのため, モータ動作は図 6.17 と同一である. 図 6.18 の前半部分では, オクルージョンにより t_4 と t_5 の間では, 顔ストリームが分断されてしまう. また t_6 から t_7 までの間は, 人が *SIG* の視野外にいるため視覚からは何も情報が得られない. このようにオクルージョンや視野外といった視覚だけでは解決できない場合でも, アソシエーションによって, 簡単に解決できる事を図 6.16 は示している.

図 6.20 は, 視覚情報による追跡の場合と同様の方法で生成した聴覚による追跡結果を示している. 音源定位モジュールは, t_3 から 23 s 付近まで, および 34 s 付近から t_{10} までの間, 正しく 2 本の音ストリームを分離することができているが, t_8 および t_9 の周辺では, 誤ったストリームが生成されていることがわかる. また, 11 s (t_5) から 17 s までの間は, A 氏の移動および *SIG* の体の回転が同時に起きているため, 図 6.20 の 2 人の話者の定位は, それほど正確ではない. つまり, 話者の移動やモータノイズとその反響音 (エコー) が, 音源定位の品質を悪化させている. このような場合でも, 視覚情報によって聴覚情報の曖昧性が解決できていることを図 6.16 は示している.

6.6.3 実験 6-3: 2 話者ですべてのセンサ情報の統合

本システムの評価には, 図 6.16 に示すシナリオをベンチマークとして使用した. このシナリオでは, A, B の 2 話者が約 20 秒に渡って以下に示す様々なアクションを行う. なお, t_n は 図 6.16 における時刻を示すものとする.

t_1: *SIG* の視野外で A が話を始める. これにより音ストリームが作られ, *SIG* は音の方向へ体を向ける.

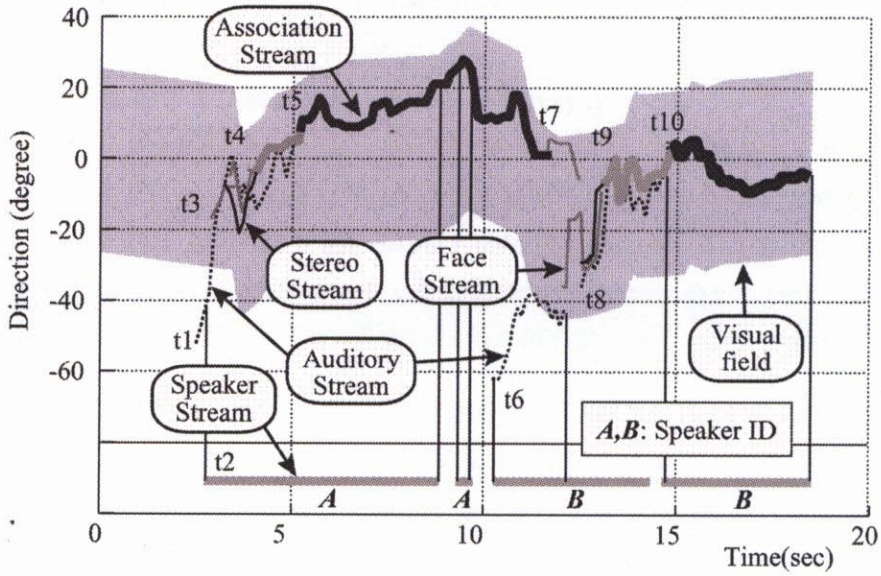


図 6.16: 2 話者における追跡結果

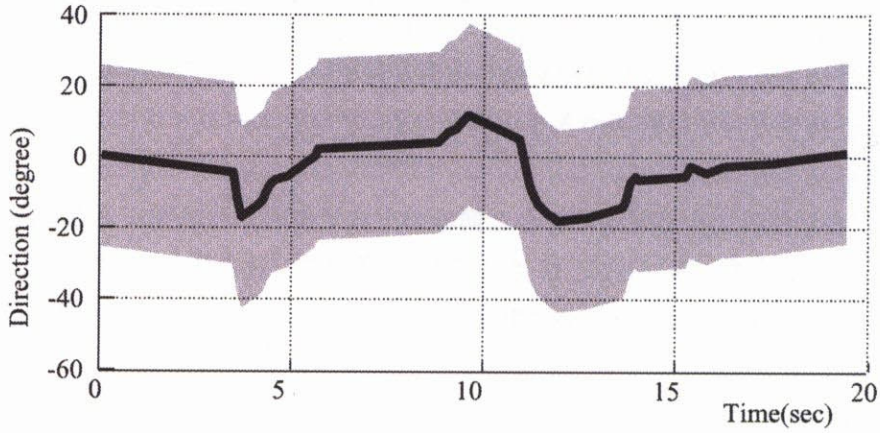


図 6.17: 図 6.16 における *SIG* の視野と正面方向

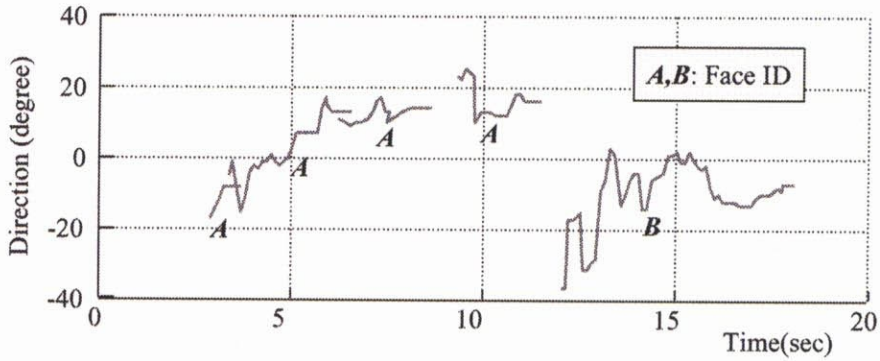


図 6.18: 図 6.16 における顔ストリーム

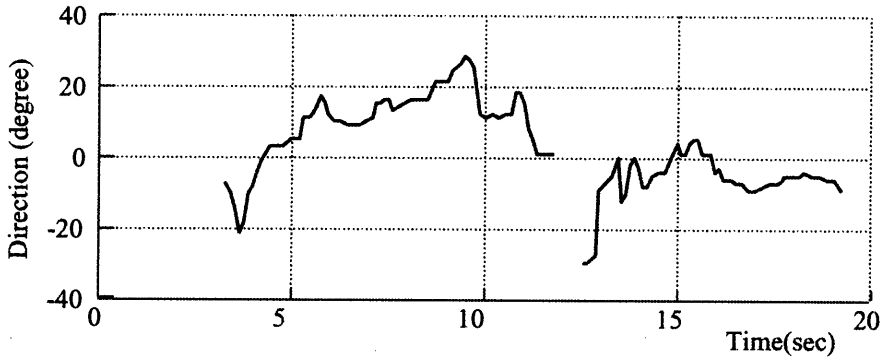


図 6.19: 図 6.16 におけるステレオストリーム

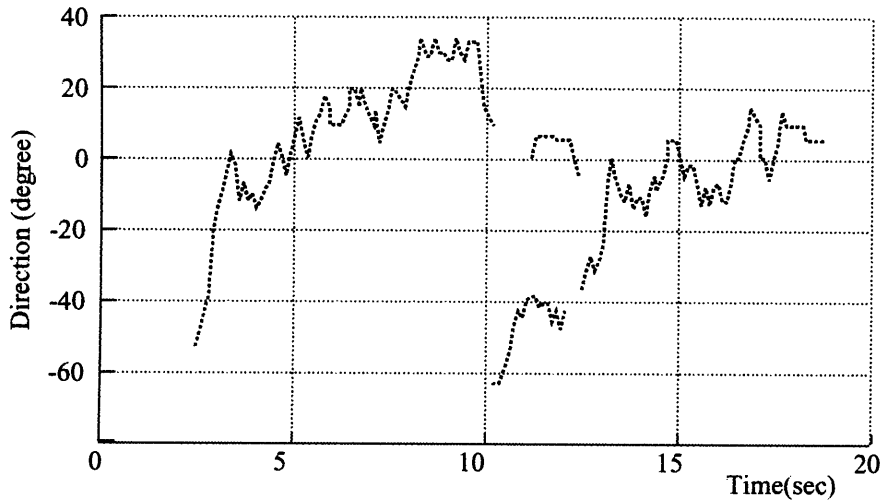


図 6.20: 図 6.16 における音ストリーム

A の声により話者ストリームが生成され, 音ストリームとアソシエーションする.
A も *SIG* の正面に向かって移動し, A 氏が *SIG* の視界に入ったことを契機に視覚的に A 氏が検出され, 顔ストリームとステレオストリームが作られる.
顔およびステレオストリームがアソシエーションする.
A に関するすべてのストリームがアソシエーションする. *SIG* はこのアソシエーションストリームに注意を向け追跡を続ける.
A をトラッキング中に *SIG* の視野外から B が話を始める. 同時に音および話者ストリームが生成され, アソシエーションされる. *SIG* はそのストリームを確かめるために音のする方に向く.
SIG が B の方向に向いたため, A が視野からはずれ, デアソシエーションされる.
B が視界に入ったため顔とステレオストリームが生成される. 同時に B は短時間,

話を中断する.

t_9: 顔とステレオストリームがアソシエーションされる.

t_{10}: B に関するすべてのストリームがアソシエーションされ, *SIG* は B の追跡を続ける.

図 6.16 より, t_5 から t_7, および t_{10} 以降では, 全種類のストリームがアソシエーションされたアソシエーションストリームにより, ロバストな人物の追跡が実現できており, t_6 から t_8 については 2 話者が同時に存在する状況で正確なストリーム分離が達成できている. この際, 複数のストリームが存在する状況下で, 視野外の音ストリームが存在しているため, 視覚情報を用いて, より正確な情報を得るようにアテンションチェンジが行われている (t_6).

このシナリオにおける *SIG* の視野と正面方向を 図 6.17 に表す. アテンション制御モジュールが細かいストリームの動きを吸収して, 滑らかに人物追跡を

可能にするとともに, 話者が見えないような状況においても聴覚情報により, 的確な追跡を達成していることがわかる. 顔ストリームを 図 6.18 に示す. 顔ストリームは比較的正確な位置情報と名前情報が得られるという利点があるが, 横や下を向いてしまったり, ライティングなどの影響で, しばしば抽出に失敗するケースがある. このため 3.5, 6, 9 秒付近ではストリームが分断されてしまっている. このような場合でもアソシエーションによって追跡が継続できることを図 6.16 は示している. 図 6.19 は, ステレオストリームを示している. ステレオストリームは比較的近い物体に関しては非常に精度のよい位置情報を取得できる. しかし, ステレオ情報は, 両眼で捕らえることができる場合のみ有効であるため, 顔ストリームと比較して視野が狭い. 図 6.16 では, 狭い視野もアソシエーションによって, カバーできることを示している. 図 6.20 は音ストリームを示す. 音ストリームはセンサの性質上, 全方位に渡って音情報を検出できる反面, 図 6.20 の 5 ～ 10 秒の間のようにそれほど正確な音源方向の精度を得ることはできない. このような場合でも, アソシエーションによって視覚情報を補うことで, 聴覚情報の曖昧性の解消が達成されている.

また, 本実験では扱わなかったが, 3 話者以上の場合やオクルージョンにより視覚情報が欠如するなど, 一部の情報が欠如した場合でも, アソシエーションによるロバストな追跡をすでに実現している. さらに, 名前情報のないステレオストリームや音ストリームは, 2 本のストリームが交差したり, 近接したりするような場合には, 誤りが生じる可能性がある. このような場合, 顔ストリームや話者ストリームといった名前情報をもったストリームとアソシエーションを行うことによって, 誤り訂正の実現が期待できる.

図 6.16 において, t_6 から始まる話者ストリームの後半部分は, t_8 から始まる音ストリームとアソシエーションすることが妥当であり, この部分に関しては, アソシエーションの構築もしくはストリームの生成に失敗しているといえる. このような場合に対処す

るには, 一度アソシエーションモジュール内で, 高位の処理であるアソシエーションストリーム生成部から低位の処理であるストリーム生成部へフィードバックを行い, ストリームの再構築を行うような機構が必要であろう.

6.7 まとめ

本章では, 視覚情報, 聴覚情報, モータ制御を統合し, 複数の顔を実環境で実時間で追跡できるシステムを説明した. 実時間で実環境で動作させるため, 個々の処理は, 速度を優先した実装を行っている. この速度優先の実装と実環境であるが故の難しさのため, 各処理はある程度, 処理精度の不足が生じてしまう. 本章では, この問題を様々な情報を統合することによって, 解決できることを, 実際にロボットへの実装を通じて示した.

例えば, 音源定位モジュールでは, 各周波数ごとの音源定位を正確に行うのではなく, 音の倍音構造, IPD, IID といった複数の聴覚情報を利用して, 精度の不足を補っている. また, 正確な顔識別や定位を行う代わりに, 聴覚情報, モータ情報などマルチモーダルな情報をイベントの流れ (ストリーム) に基づき, 統合することにより, 精度不足を補っている. 実環境の実時間アプリケーションにおける情報統合の有効性は, 様々な状況に対応できるロバストな処理が実現できた事によって, 証明できた. 特に, 音環境理解 (CASA) に基づいた聴覚情報処理は, ロボットに対してそれ自体有効であるばかりか, 視覚情報の視野の不足を補うという意味でも有効であることを示す事ができた.

情報統合については, 2 章で述べたようにノンパラメトリックカルマンフィルタを用いたような情報理論的に最適な統合であることが保障される情報統合の枠組みも検討されており, 様々な適用を通じてその有効性が報告されている. 本章で示した情報統合は, 最適な情報統合であることは保障せず, アソシエーションされたストリームの中から最も精度の高いストリーム情報を選択するといるシンボリックな手法であるため, 統合による定位精度は, 高々, 最も高精度なモダリティの精度である. しかし, 実験結果から明らかなように, このような統合手法でも実環境で十分有効である. さらに, 様々なモダリティの階層的な統合, 文脈を考慮したトップダウンな統合制御とセンサベースのボトムアップな統合制御を同時に行うような統合, 様々な処理で混入するノイズの影響にロバストでスケーラビリティのある統合の実装が容易であるという利点がある. これは, 実際に, 音, 顔情報以外にステレオ視, 話者同定情報といった抽象度の異なるモダリティを追加し, 容易にスケールアップが実現できることにより示されている.

実装したシステムの定量的な評価については, 今後の課題である. 実時間で動作するシステムの場合, 刻々と周囲の状況が変化するため, 実験を完全に再現することが難しい. 例えば, 静的なシステムでは, 一度, マイクロホンやカメラで録音・録画してしまえば, 何度でもオフラインで同じデータを用いた実験を行うことができるが, *SIG* では, マイクロホ

ンやカメラが動作しながら録音・録画を行うため, 全く同じ条件で, 行動制御の方法を変えるような実験を正確に行うことは難しい. 従って, システムの定量的な評価については, 評価の手法と共に検討する必要があり, 大きな今後の課題である.

将来, ロボットに人間との知的なソーシャルインタラクションを求めるためには, 触覚情報など, さらに多くのセンサ情報を統合し, 知覚のロバスト性を高めることが求められよう. また, 未知環境や動的に変化する環境へ対応するための学習の枠組みや, より複雑な状況に対応できる注意制御ポリシーも必要であろう. これらも今後の検討課題である.