第7章

アクティブ方向通過型フィルタによる音源分離

本章では, 聴覚中心窩を積極的に利用できるよう, 方向通過型フィルタを拡張したアクティブ方向通過型フィルタ (*Active Direction-Pass Filter*, ADPF) を提案する. アクティブ方向通過型フィルタは, 聴覚中心窩に基づきフィルタの通過帯域をアクティブに制御し, 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. また, 実際にアクティブ方向通過型フィルタを利用した音源分離システムを, 2 本のマイクを搭載したロボットに実装し, 評価を行う.

7.1 聴覚中心窩

2 本のマイクロホンを利用したロボットでは, 本研究で, 聴覚中心窩と呼んでいる現象が存在する. アクティブ方法通過型フィルタは, 聴覚中心窩に基づいて構成されるため, まず, 聴覚中心窩について説明する.

7.1.1 聴覚中心窩とは

霊長類の視覚は, 中心窩と呼ばれる解像度が高い部分が中心部に存在し, 周辺部では解像度が低くなる代わりに, 広範囲な視野を得ている. このような構造を用いれば, 対象物を中心窩で捕らえることにより, 高解像度の情報を取得することができる. つまり, 広い視野と高い解像度を併せ持ち, かつ脳の情報処理量を劇的に削減できる効率的な構造を有している. ロボットでも, 同様の構造により計算量を削減できることから, 中心窩を利用した視覚処理はアクティブビジョン (*Active Vision*) [5] の典型的な例として, しばしば利用されている [68, 107].

人間の聴覚においても, 水平方向の音源定位の精度は正面方向で最も高く, 周辺部に行くに従い低くなることは, 古くから知られている [20]. 耳に2つのマイクを備えたロボットによる音源定位でも, 人間と同様の傾向が見られる. 図 7.1 は, 実時間人物追跡システム [79] における 3 つの定位モジュール**音源定位**, **顔定位**, **ステレオ物体定位**による定位結果の平均値, 図 7.2 は, **音源定位**による定位結果の分布を表している.

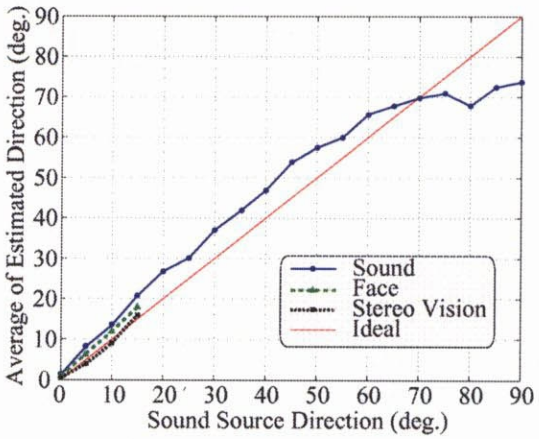


図 7.1: 「顔認識・定位」, 「ステレオビジョン」, 「音源定位」における音源方向に対する定位精度

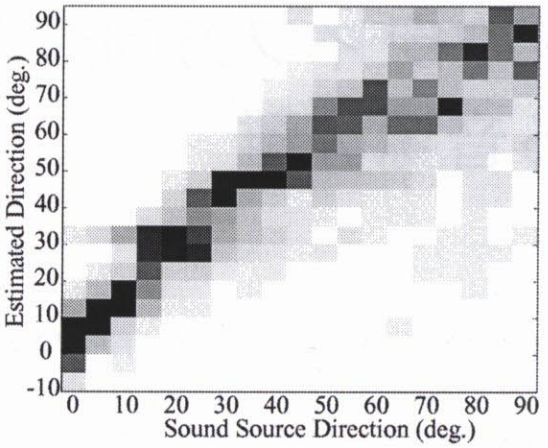


図 7.2: 音源方向に対する音源定位結果の分布

図 7.1 から, **音源定位**による定位誤差は, 正面方向から 20° 付近まで増加した後, 20° から 70° 付近までは 6° 程度で一定だが, それ以降は大きく悪化し, 90° では, 15° 以上になる. また, 図 7.2 から, 正面方向のばらつきは少なく, 正面から離れるにつれ, ばらつきが目立ち, 分散が大きくなる. このように定位結果の平均, 分散は, ともに正面方向で音源定位の精度が高くなることを示しているので, 本現象をロボットにおける聴覚中心窩と呼ぶ.

なお, 神経行動学 (neuroethology) では, ドップラー効果によるエコー音の周波数変化を抽出するため, キクガシラコウモリの蝸牛殻で特定の周波数に対する感度が高くなっている部分を聴覚中心窩と呼んでいる [110]. 選択的注意という広義の意味では, 両者は似ているが, 本稿では, ロボット頭部の正面方向で感度が高いという意味で聴覚中心窩という言葉を使用する.

図 7.1 では, ステレオビジョンによる定位誤差は 1°, 顔定位による誤差は 2° 程度と, 聴覚処理よりも正確であることがわかる. これは, 音源方向が正面に近く, 視覚情報が利用できる場合では, 高精度の視覚情報によって, 聴覚の精度不足を補うことが可能であることを示している.

これらから, 音源定位では, 視覚の中心窩と同様に, 聴覚中心窩を利用して音源に正対するようなアクティブな動作を行えば, システムの精度の向上が期待できる. さらに正面方

向で視覚情報が利用できれば, 視聴覚統合により分離のロバスト性を向上できると考えられる.

7.1.2 聴覚中心窩の方向通過型フィルタへの適用

方向情報を利用した音源の分離抽出を考えた場合, 正面方向の音源であれば, 正確な音源方向を利用することができるが, 音源方向が正面から離れるにつれ, 方向情報に精度を期待できなくなる.

方向通過型フィルタは, スペクトルの各サブバンドで, IPD と IID に対する仮説推論を行うことによって, 特定方向の音を抽出するものである [96]. そこで, 方向通過型フィルタを拡張し, 正面方向の音源については, 通過させる方向の範囲 (帯域) を狭く, 正面から離れた音源では, 通過させる帯域を広く取るように通過帯域を制御することにより, 音源の分離抽出精度の向上が期待できる. アクティブ方向通過型フィルタでは, このような通過帯域制御を行って, 正面方向では S/N 比の高い音響信号を抽出し, 正面方向から離れた音源に対しては帯域を広く取り, 背景雑音の混入により S/N 比は多少落ちるものの, 必要な情報をできるだけ抑制せずに, 特定の音源を強調することを目的とする. 正面方向から離れた音源を精度よく抽出する必要がある場合は, 聴覚中心窩を利用できるように, 音源方向を向くような制御を行う.

7.2 アクティブ方向通過型フィルタの詳細

図 7.3 にアクティブ方向通過型フィルタを用いた音源分離システムの構成図を示す. 図中の網掛け部分がアクティブ方向通過型フィルタの構成に対応する. アクティブ方向通過型フィルタへの入力は 4 つあり, 入力のスペクトル, 入力スペクトルから計算される IPD と 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となっている. 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, 聴覚中心窩に基づく通過帯域制御とロボットの伝達関数を利用した仮説生成により, 方向通過型フィルタでは難しかった高速な実環境での音源抽出を可能にしている. ここで, ロボットの伝達関数は, 部屋の伝達関数, ロボット頭部による音の歪みなどを考慮して, 特定方向の IPD および IID を推定するための関数である. 以下では, アクティブ方向通過型フィルタのアルゴリズムの詳細について説明する.

7.2.1 アクティブ方向通過型フィルタのアルゴリズム

アクティブ方向通過型フィルタのアルゴリズムは以下の 6 ステップで構成される.

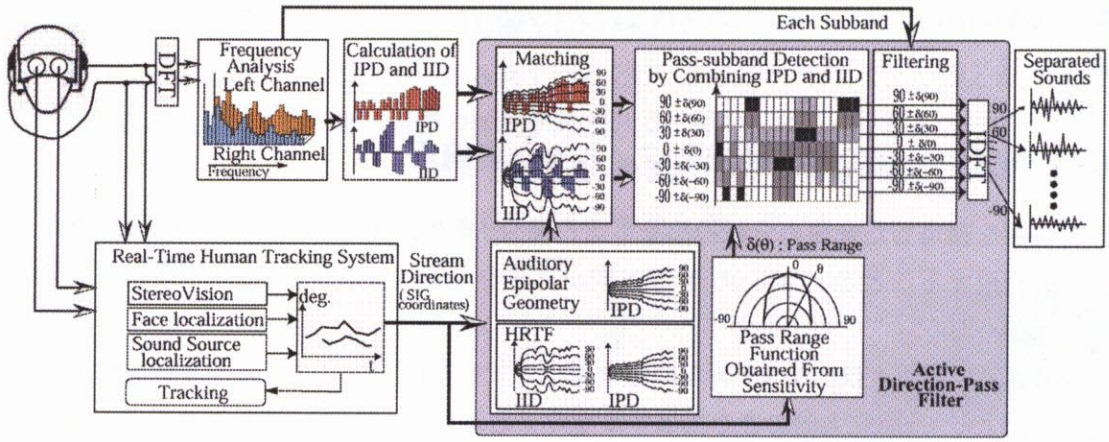


図 7.3: アクティブ方向通過型フィルタによる音源分離システム

1. 入力音のスペクトルから, 各サブバンドの IPD $\Delta\varphi'$ と IID $\Delta\rho'$ を計算する. ここで, Sp_l, Sp_r は, それぞれある時刻に左右のマイク入力信号から得られたスペクトルである.

$$\Delta\varphi' = \arctan\left(\frac{\Im[Sp_l]}{\Re[Sp_l]}\right) - \arctan\left(\frac{\Im[Sp_r]}{\Re[Sp_r]}\right) \tag{7.1}$$

$$\Delta\rho' = 20\log_{10}\left(\frac{|Sp_l|}{|Sp_r|}\right) \tag{7.2}$$

2. 抽出すべき音源の方向を θ_s とする. θ_s は 6 章で述べた実時間人物追跡システムから, ロボット座標系での水平角として得られる.
3. 通過帯域関数に従って, θ_s に対応するアクティブ方向通過型フィルタの通過帯域 $\delta(\theta_s)$ が選択される. 通過帯域関数は, 聴覚中心窩に基づき, ロボットの正面方向で最小となり, 周辺部で大きな値をとる関数である. 詳細は 7.2.2 で述べる. 選択された通過帯域 $\delta(\theta_s)$ を用いて, $\theta_l = \theta_s - \delta(\theta_s)$, $\theta_h = \theta_s + \delta(\theta_s)$ と定義すると, θ_l から θ_h の範囲にある音響信号を抽出するのがアクティブ方向通過型フィルタの基本的な動作である.
4. θ_l と θ_h に対する IPD, IID を推定する. これらの推定には, ロボットの伝達関数を利用する. ロボットの伝達関数には, 無響室で水平方向について 5° 刻みで, インパルス応答を計測することによって得られる計測伝達関数を用いる. また, 計測を行わなくてもかまわないように, 計算的にロボットの伝達関数を求めるよう拡張した方法についても 7.2.3 節で述べる.
5. 音源方向 θ に対して, ロボットの伝達関数を利用して, 入力スペクトルから以下の条件式 **A** を満たすサブバンドを選択する.

A: $f < f_{th}$: $\Delta\varphi_H(\theta_l) \leq \Delta\varphi' \leq \Delta\varphi_H(\theta_h)$,
$f \geq f_{th}$: $\Delta\rho_H(\theta_l) \leq \Delta\rho' \leq \Delta\rho_H(\theta_h)$

$\Delta\varphi_H(\theta)$, $\Delta\rho_H(\theta)$ は, それぞれロボットの測定伝達関数から推定される IPD, IID である. f_{th} は, フィルタリングの判断基準に IPD と IID のどちらを用いるかを決める閾値である. 一般に, 低周波数域では IPD, 高周波数域では IID が大きく影響し, この閾値はマイク間距離に依存する. 我々のロボットでは, 理論的にも, 実験的にも f_{th} として 1500 Hz が妥当であることが報告されている [80].

6. 選択されたサブバンドから, 音響信号を再合成し, 該当範囲にある音響信号を抽出する.

実際には, 音源方向 θ_s は時間 t の関数であるため, 特定音源を抽出し続ける際には, 時間方向の連続性を考慮する必要がある. 本稿では, 6 章に述べる実時間人物追跡システムから音源方向を得ることでこれを解決している. 実時間人物追跡システムでは, すべての情報をストリームという時間的な流れを考慮した表現を用いて表しているので, 同時に複数の音源が存在したり, 音源や自分自身が移動する場合でも, 一つのストリームに注目することによって, 特定音源からの方向情報を連続的に得ることができる. また, ストリームは視聴覚情報を統合するためにも使用しており, これにより, 7.1.1 節で述べた視覚情報による, 音源定位精度向上を実現している.

7.2.2 通過帯域制御

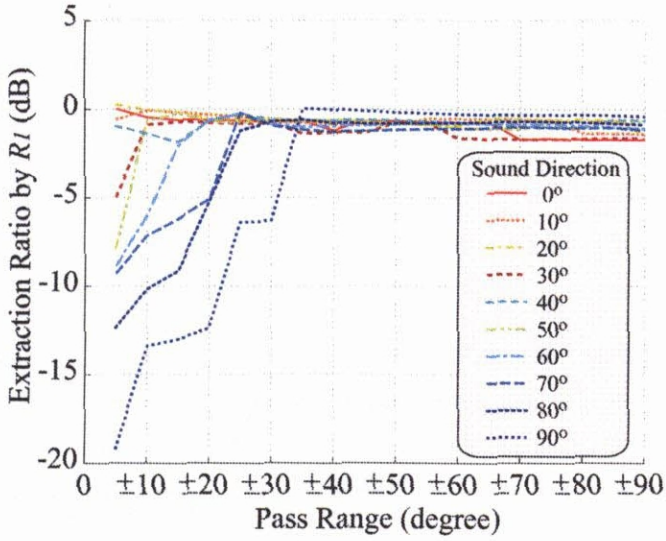


図 7.4: 通過帯域に対する単一音源の抽出率

アクティブ方向通過型フィルタにおいて通過帯域制御に使用する通過帯域関数を求めるために, 音源数 1 の場合に音源方向や通過帯域を様々に変化させて, 抽出精度の違いを調べた.

音源には, スピーカから出力される音声信号を用いた. スピーカとロボットの距離 は 1m とし, スピーカの水平方向を, ロボットの正面から, 0°～90° まで 10° おきに変化させた. また, 音源を抽出する際には, スピーカ方向は既知であるものとし, 方向通過フィルタの通過帯域を ±5°～±90° まで ±5° 単位で変化させて音源を抽出し, S/N 比による比較を行った. S/N 比の算出には, 7.3 節の式 (7.3) で定義する R_1 を用いた.

図 7.4 に結果を示す. 実験では, 背景雑音は無視できる程度に小さかったため, 音源数が 1 の場合は, R_1 が $0\,dB$ となった時に, 元波形が完全に抽出できたと解釈する. 音源方向が 0°～30° と正面方向に近い場合には, 通過帯域が ±10° 程度で元波形を抽出できているが, 音源方向が正面から離れるに従い, 元の波形に含まれるパワーを抽出するために, 広い通過帯域を必要とし, 音源方向が 90° の場合には, 最低でも ±35° 程度の通過帯域が必要であることがわかる.

音源数が 1 の場合には, 通過帯域が広ければ広いほど, S/N 比の高い信号を抽出することができるが, 実環境では, 背景雑音を含め, 複数の音源を考慮する必要があるため, なるべく通過帯域を狭くとることが望ましい. そこで, 図 7.4 から, ほぼ元波形が抽出でき, かつ極力狭い通過帯域を音源方向ごとに抽出し, 図 7.5 のように通過帯域関数を導出した. 通過帯域は正面方向では狭く, 周辺部では広くなっていることがわかる. これは, 音源定位と同様に, 音源分離でも聴覚中心窩を利用することが可能であることを示している. 実際の利用では, 他の音源の音を極力抽出したくない場合, 単なる音響信号の強調として利用したい場合など, 状況に応じたチューニングが必要な場合もあると考えられるが, 以後の実験では, 図 7.5 に示された通過帯域関数を利用するものとする.

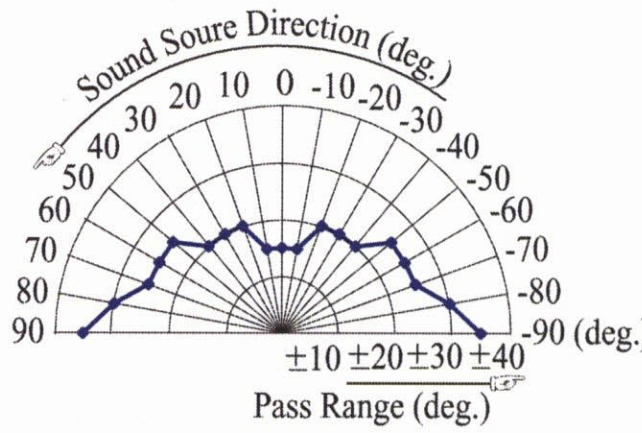


図 7.5: 通過帯域関数

7.2.3 ロボットの伝達関数の拡張

式 (5.16) を利用すれば, 音源方向 θ から IPD を推定することができる. 計算的に IID を推定することは難しいため, 完全に 7.2.1 節の条件 **A** の代替となるような計算的な推定法を得ることはできないが, 式 (5.16) から推定された IPD を $\Delta\varphi_E(\theta)$ とすると, 下記の条件式を利用することも可能である.

B: $f < f_{th}$: $\Delta\varphi_E(\theta_l) \leq \Delta\varphi' \leq \Delta\varphi_E(\theta_h)$,
$f \geq f_{th}$: $\Delta\rho_H(\theta_l) \leq \Delta\rho' \leq \Delta\rho_H(\theta_h)$

C: $f < f_{th}$: $\Delta\varphi_E(\theta_l) \leq \Delta\varphi' \leq \Delta\varphi_E(\theta_h)$

条件 **B** は IPD には聴覚用エピポーラ幾何, IID には測定ロボット伝達関数を使用する場合, 条件 **C** は聴覚用エピポーラ幾何のみを使う場合に相当する. これらの条件を利用した場合については, 評価実験で比較検討する.

7.3 アクティブ方向通過型フィルタの評価

アクティブ方向通過型フィルタの効果を調べるため, 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 1 m 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, 実験 1 の音響信号には, 純音を用い, 実験 2, 3 の音響信号には, "音声認識システム"[125] に付属する毎日新聞記事の読上げデータを使用した.

評価には, 以下の 3 つの評価指標 $R_1 \sim R_3$ を用いた. また, 試聴による評価も行った.

1. 周波数領域における入力信号と分離信号の S/N 比の差 (dB)

$$R_1 = 10\log_{10}\left(\frac{\sum_{j=1}^{n}\sum_{i=1}^{m}\left(|sp(i,j)| - \beta|sp_o(i,j)|\right)^2}{\sum_{j=1}^{n}\sum_{i=1}^{m}\left(|sp(i,j)| - \beta|sp_s(i,j)|\right)^2}\right) \tag{7.3}$$

$sp(i,j)$, $sp_o(i,j)$, $sp_s(i,j)$ はそれぞれ, スピーカから出力される原信号, ロボットのマイクで収音された観測信号, 分離信号のスペクトルを示す. また, m と n はスペクトルのサブバンド数と時間方向の平滑化のサンプル数, β は原信号と観測信号の減衰比を示す.

2. 信号損失比 (dB)

$$R_2 = 10\log_{10}\left(\frac{\sum_{n\in S}(s(n) - \beta s_o(n))^2}{\sum_{n\in S}(s(n) - \beta s_s(n))^2}\right) \tag{7.4}$$

$s(n)$, $s_o(n)$, $s_s(n)$ は, それぞれ, スピーカから出力される原信号, ロボットのマイクで収音された観測信号, 分離信号を示す. S は, 信号のうち, 分離すべき信号が含まれている部分を示し, $s(i) - \beta s_o(i) \geq 0$ を満たす i の集合として定義される.

3. ノイズ抑制比 (dB)

$$R_3 = 10\log_{10}\left(\frac{\sum_{n\in N}(s(n) - \beta s_o(n))^2}{\sum_{n\in N}(s(n) - \beta s_s(n))^2}\right) \tag{7.5}$$

$s(n)$, $s_o(n)$, $s_s(n)$ は上述の通りである. N は, $s(i) - \beta s_o(i) < 0$ を満たす i の集合であり, 信号中の分離すべき信号が含まれていないノイズ部分を表す.

実験 7-1: アクティブ方向通過型フィルタの基本性能を調べるため, 1 音源に対する音源抽出を行った. 音源には, 周波数帯域ごとのパフォーマンスを調べるため, 100, 200, 500, 1000, 2000 Hz の 5 つの純音を使用した. スピーカ 1 台を用いて, 5 つの純音を 0° から 90° まで, 10° おきに動かし, 抽出率を調べた. 抽出率には, R_1 を用いている. 分離の条件は 7.2.3 節で述べた条件 **C** を用いた.

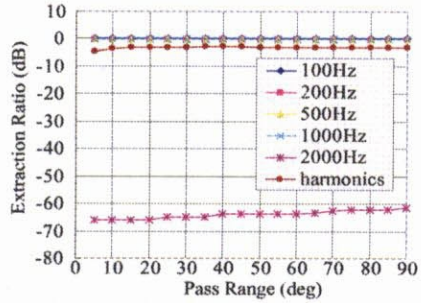
実験 7-2: 2 話者が同時に発話する場合に, それぞれの音源の分離抽出を行った. 音源には 2 台のスピーカを使用し, 一方のスピーカを 0° の方向に, もう一方のスピーカを 30°, 60°, 90° と 3 段階に変化させ, 同時に等音量で異なる音声を出力した. それぞれに対して, R_1, R_2, R_3 を計測した. また, 比較のため分離には 7.2.1 節で述べた条件 **A** の他に 7.2.3 節で述べた条件 **B**,**C** の 3 種類を用いた.

実験 7-3: 3 話者が同時に発話する場合の音源分離抽出を行った. 3 台のスピーカを用い, スピーカの 1 台を 0° に固定し, 残る 2 台のスピーカの位置を ±30°, ±60°, ±90° と変化させた. それぞれの場合について R_3 を計測した. 出力音声やフィルタリングの条件は実験 1 と同様である.

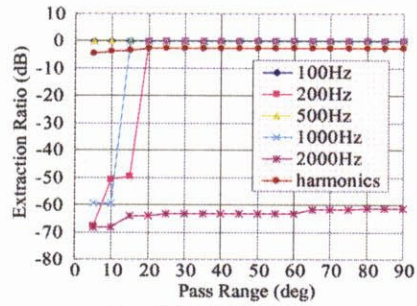
実験 7-1 の結果を図 7.6 に示す. 実験 7-2 の結果を表 7.1 に, 2 話者同時発話の分離の一例を図 7.7 に示す. また, 実験 7-3 の結果を表 7.2 に示す.

図 7.6 から, 周波数帯域によって抽出精度が異なることがわかる. 特に周波数が高くなるに従って, 抽出精度が落ちることがわかる. これは, 高周波数域では位相差情報に曖昧性が生じるためであることが考察される. また, どの周波数帯域でも, 音源方向が正面から離れるにつれ, より広範囲の通過帯域が必要であることがわかる. 抽出率が 0 dB, つまり, 完全に音源が抽出できたとみなせる通過帯域は, 図 7.5 で定義した通過帯域関数とよく対応しており, 通過帯域関数の妥当性を示すことができた.

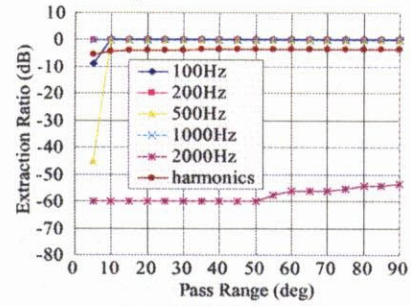
図 7.7a) は, スピーカから流したオリジナルの波形である. 正面のスピーカからは「ある日の二人の食事を紹介しよう」, 30 度のスピーカからは「打線は一番から四番まで左打者」という音声を流した. *SIG* のマイクロホンで収音した波形は, 図 7.7b) に示されてい



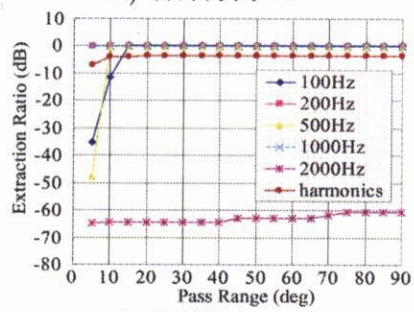
a) 音源方向 0°



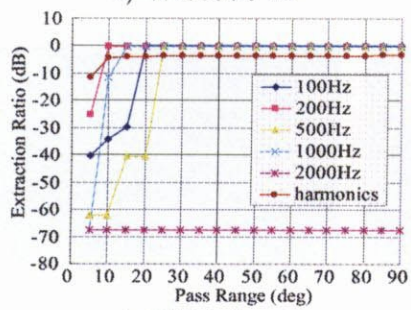
b) 音源方向 10°



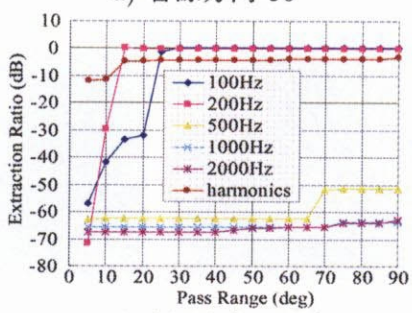
c) 音源方向 20°



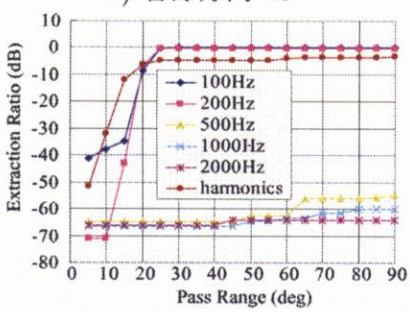
d) 音源方向 30°



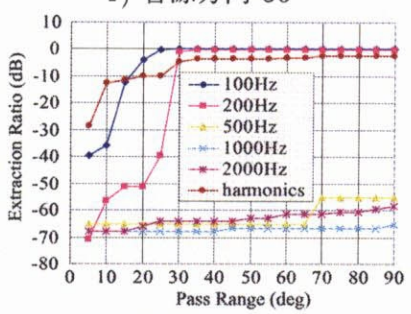
e) 音源方向 40°



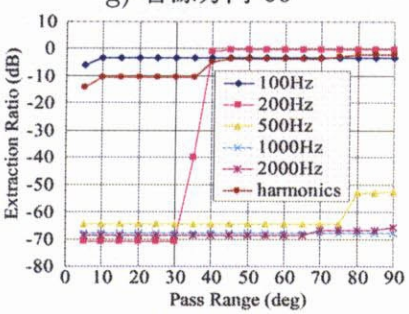
f) 音源方向 50°



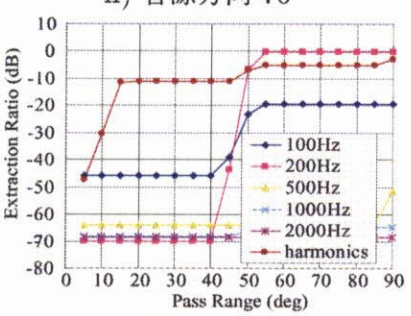
g) 音源方向 60°



h) 音源方向 70°



i) 音源方向 80°



j) 音源方向 90°

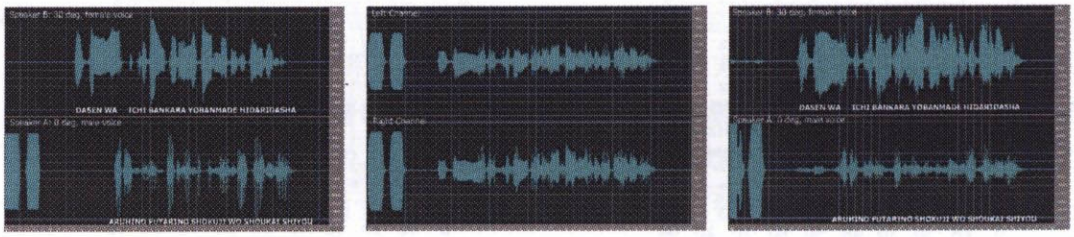
図 7.6: 音源方向, 周波数に対する音源抽出率

表 7.1: R_1, R_2, R_3 による 2 話者同時発話分離の評価

	$R_1(dB)$				$R_2(dB)$				$R_3(dB)$			
	0°	30°	60°	90°	0°	30°	60°	90°	0°	30°	60°	90°
A	2.2	1.4	1.6	0.8	-2.1	-3.4	-3.8	-7.3	9.1	4.6	3.4	-2.8
B	2.2	1.1	2.1	0.6	-2.5	-4.0	-3.3	-7.7	10.3	6.8	2.6	-3.5
C	2.0	1.3	2.2	0.5	-2.8	-3.1	-3.3	-7.7	10.4	4.7	2.6	-3.5

表 7.2: R_3 による 3 話者同時発話分離の評価

Interval	30°			60°			90°		
Direction	−30°	0°	30°	−60°	0°	60°	−90°	0°	−90°
A	4.8	9.1	4.7	3.7	9.2	3.8	-1.7	9.1	-1.3
B	5.7	7.4	5.9	3.5	9.6	3.1	-2.0	9.8	-2.0
C	4.9	8.1	5.1	3.2	9.6	3.1	-1.9	10.5	-1.7



a) 元波形　　b) 収音波形　　c) 分離波形

図 7.7: 同時 2 話者発話の分離例

る．この波形から，分離フィルタで分離したものが，図 7.7c) である．音は歪んでいるが，それぞれの音声が理解できる程度の分離が達成できている．

表 7.1 における S/N 比の向上 (R_1) を見ると，音源方向が 90° の時は，ほとんど効果が見られないが，その他の方向では，1〜2 dB 程度の向上が見られ，特に正面方向の音源に対する抽出精度が高く 2.2dB に達している．信号損失比 (R_2) とノイズ抑制比 (R_3) という観点から見ても，やはり，音源方向が 90° の時は，R_2 が 7 dB 以上も悪化しており，R_3 も負の値であることから，音源の抽出がうまく行われていないことを示している．これは，外装の影響でマイクが *SIG* 前方に指向性を有しているため，収音時にすでに信号が劣化してしまっていることが一因と考えられる．しかし，その他の方向では，R_2 は 2〜4 dB であり，R_3 は 3〜10 dB と概ね良好な数字を示している．特に，R_3 は正面方向ほど高くなっており，音源方向を向くという動作を行うことによって，抽出精度が劇的に上がることを示

している．表 7.2 においても，ほぼ同様のことが言える．特に，正面方向の音源の抽出では，他の 2 つの音源の方向にかかわらず，10 dB 程度のノイズ抑制効果があることがわかる．しかし，スピーカ間の角度が 30° の場合には，正面方向のノイズ抑制効果が若干落ちていることから，30° 以内に複数の話者がいる場合の分離は難しいと考えられる．

フィルタリングの条件式による差異については，両実験とも条件に関係なく，同様の傾向が見られる．IPD のみを用いた条件 **C** でも IPD と IID の両方を用いた場合とあまり差が見られないのは，IID を用いる 1500 Hz 以上の周波数域の音声パワーが小さいためと考えられる．従って，S/N 比という観点からは，IPD だけを利用する聴覚エピポーラ幾何のような完全に計算的な手法でも実環境での音源分離に有効であるといえる．しかし，音響学の識者 2 名による試聴では，条件 **C** の明瞭度が最も悪く，明瞭度という観点では，1500 Hz 以上の周波数域も抽出精度に影響している．また，試聴では，条件 **B** が最も明瞭度が高く，ベストケースでは 14ch の線形，もしくは 16ch 円形マイクロホンアレイにも匹敵するとの評価を得た．明瞭度でも条件 **A** と条件 **B** の差は僅少で，IPD に関しては，計算的手法である聴覚エピポーラ幾何でも測定関数に匹敵する性能を得られることがわかる．

計算時間的には，使用している Pentium III 1 GHz のマシン上で，実時間で，バッファオバーフローが発生することなく動作が可能である．

以上より，アクティブな動作や通過帯域制御により，聴覚中心窩をうまく利用すれば，音源分離を向上できることが明らかである．また，現時点では，ロボットの伝達関数は予め測定する必要があるものの，IPD に関しては，計算的な手法による推定も可能であることがわかった．なお，IID に関する計算的な推定法は今後の課題である．

7.4 まとめ

本章では，ロボットに搭載できる音源分離システムを実現するために，アクティブ方向通過型フィルタを提案・評価し，その有効性を示した．

マイクロホン数以上の音源を実環境で分離できる能力を備えたロボットは，これまで報告されていないが，アクティブ方向通過型フィルタは，2 本のマイクロホンを備えたロボットを対象に，少なくとも 3 音源以上の実時間分離が可能であることを示した．これは，ICA，ビームフォーミングといった一般的な音源分離手法は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を大きく緩和したといえる．

2 本のマイクロホンでマイクロホン数以上の高速な音源分離を可能にした鍵は，聴覚中心窩に基づき，音源方向に応じたアクティブな通過帯域制御と音源方向を向くというアクティブな動作である．これは，従来の音源分離手法が，マイクロホンと音源の相対関係が変化する場合をあまり考慮していないことに対し，積極的にマイクロホンのパラメータを変化させるという全く逆の発想により音源分離向上を実現したといえる．つまり，ロボット

では, アクティブオーディションが知覚向上に対して本質的であることを示している. 音源方向を向くという動作は, ロボット聴覚の向上だけではなく, 人間とのフレンドリなインタラクションを実現したり, テレイグジスタンスによる会議では, 相手の注意を向けさせたりという意味でも重要である.

アクティブ方向通過型フィルタでは, 正面方向では, 9 dB 程度のノイズ除去率を示しているが, 横方向では 3 dB 以下まで悪化する. これは, 正面方向では, 音源間角度に対する分離分解能が高いためである. 実際に通過帯域関数が示すように, 正面方向でも, 音源間角度が 30° 以内では分離が難しく, 横方向では, 70° 程度の分解能となる. このことから, 原理上, 最大 3 から 5 音源程度の分離が可能である. また, 音源定位と組み合わせて利用した場合, 音源定位の限界である 3 から 4 音源が, 実質上の最大分離音源数となる. 現時点では, 強度差情報には, HRTF を利用しているため, 予め部屋の音響環境がわかっていることが望ましい. より反響の強い部屋での動作, 方位角方向に対する分離, 前後問題の解決は今後の課題である.

次章では, アクティブ方向通過型フィルタが, 本研究で求められる「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」を満たしているのかを検証するために, その応用として分離音を認識する音声認識システムを説明する.

第 8 章

複数の音響モデルを利用した音声認識

本章では, 提案したロボット聴覚システムの応用として, アクティブ方向通過型フィルタで分離した分離音声の音声認識を説明する. この場合, 音源方向, 話者ごとに音響モデルを作成し, 複数の音声認識を用いた方法が有効であることを示し, これを利用した音声認識例を紹介する.

8.1 分離音の音声認識

一般に, ノイズに対しロバストな音声認識には missing data や missing fearute など一部区間が信頼できないことを前提にした音声認識手法が有効とされている [18, 105]. これらの方法は, ノイズが少ない場合は有効であるが, S/N 比が 0 dB 程度まで小さくなると認識率を向上させることは難しい. これは, 同時発話など, 複数の音声が混在している場合にも同様のことが言える. このような場合には, 音声認識のフロントエンドとして音源分離が必要である.

ところが, 7 章で紹介した通過帯域をアクティブに制御するアクティブ方向通過型フィルタでは, その性質上, 音源方向ごと, 話者ごとに分離音の音響的な性質が大きく変化する. そのため, 汎用の音声認識の音響モデルを作成することは難しい. そこで, 音源方向, 話者ごとに, 音響モデルを作成し, 複数の音声認識を用いた方法を提案する.

8.2 使用した音声認識システム

音声認識システムは, 一般に, 音声認識エンジン, 音響モデル, 言語モデル, 辞書類からなっている.

このうち, 音声認識エンジンのアルゴリズムは, ほぼ確立されつつあり, フリーで入手可能なソフトウェアも複数存在する. このような音声認識システムのうち, 京都大学で開発された "Julian"[125] は, 日本語による音声認識が可能で, ツール類や音響モデル, 言語モデルも提供されている数少ないシステムである.

Julian で使用する音響モデルは, 標準的な Hidden Markov Model Toolkit (HTK) のフォーマットとコンパチブルであるため, HTK を用いた音響モデル作成が可能である. HTK では, 音響モデルの作成方法が, ドキュメント化されており第3者による音響モデル作成が比較的容易である.

言語モデルについては, 自由に文法を記述できるようになっており, ディクテーションから孤立単語認識まで簡単に行うことができるように設計されている.

本研究では, 音源分離のアプリケーションとして利用するため, 音声認識エンジンを改良するアプローチは避け, Julian を利用して, 音響モデル, 言語モデル, 辞書類をチューニングすることによるアプローチを試みた.

8.3 言語モデル・音響モデル・辞書のチューニング

まず, 単語辞書については, 色, 数字, 食べ物といったすべて名詞で 150 単語を使用している.

言語モデルについては, 話者が単語を喋ることを想定し, 孤立単語認識を行っている.

音響モデルは, 話者, 音源方向ごとに作成した. 使用した音声は, 男性 2 名 (A 氏, C 氏), 女性 1 名 (B 氏) の計 3 名のデータである.

3 名のデータを, クリーンな環境でハンドマイクで録音した後, ロボットのマイクロホンを用い, 以下に示す条件で, 収録した.

4 章に示した $3m \times 3m$ の部屋で, SIG から $1m$ の距離にスピーカを設置した. スピーカは, 3 台使用し, SIG から見て, $0, \pm 60°$ の位置に設置した.

この状態で, 各スピーカから同時に音声を出力し, これを録音した. 録音時の各スピーカからの出力音声は, A, B, C の何れかの話者の音声, もしくは「音声なし」のうちどれか 1 つであり, 同じ話者の音声を同時には出力しないという条件で, すべての組み合わせを行った. 「音声なし」を含めているのは, 1 つのスピーカしか使用しない場合から, 3 つを同時に使用する場合まで, すべての組み合わせを行うためである.

次に, 各音声の方向を既知として, 録音データからアクティブ方向通過型フィルタによる音声抽出を行った.

抽出した音声を話者, 発話方向ごとに整理し, 音響モデルのトレーニングセットとした. また, 各トレーニングセットには, 情報落ちがないクリーン環境で録音したデータも合わせて加えた.

各トレーニングセットごとに, Hidden Markov Model Toolkit (HTK) を用いてトライフォンによる音響モデルを作成した. この結果, 3 話者, 3 方向の組合せ, つまり 9 種類の音響モデルを作成した.

8.4 複数の音響モデルを利用した音声認識

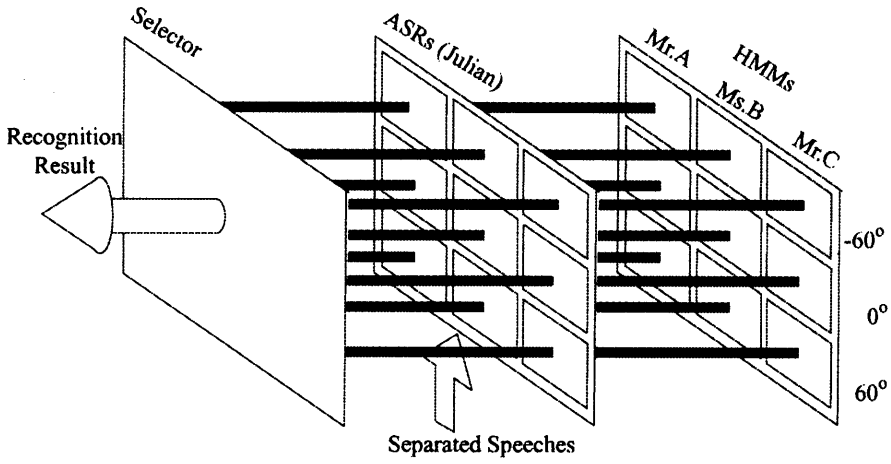


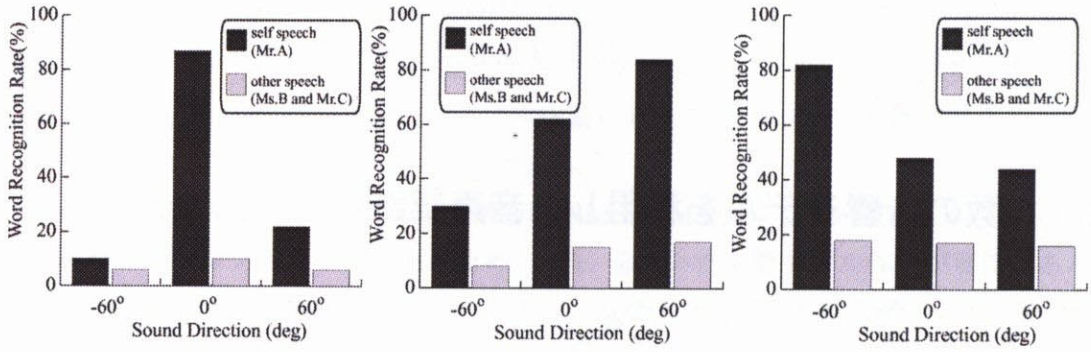
図 8.1: 方向・話者依存の音響モデルを利用した音声認識

音声認識では, 分離音声に対して, 音響モデルごとに, 並列に 9 つの音声認識プロセスが実行される. 図 8.1 はその様子を示している. 従って, 一つの分離音声に対して, 9 つの音声認識結果が得られる. そこで, セレクタはすべての音声認識結果を統合し, 最も信頼性が高いと判断される結果を出力する.

セレクタの精度が音声認識結果に大きな影響をもたらすため, 以下のようにして, 統合のアルゴリズムを定義した. まず, 特定話者の音響モデルに対する単語認識率を調べた. 図 8.2 に示した結果から, 話者よりも方向の違いによる認識率の低下が少ないことがわかる. また, 話者も方向もあっている場合は 80% 以上の認識率であることがわかる. この結果を踏まえ, 音声認識の際には, 音源方向は既知であることを利用し, セレクタには, 式 (8.1) に示すコスト関数を統合のために使用している.

$$V(p_e)=\left(\sum_d r(p_e,d)\cdot v(p_e,d)+\sum_p r(p,d_e)\cdot v(p,d_e) - r(p_e,d_e)\right)\cdot P_v(p_e). \qquad (8.1)$$

$$v(p,d)=\begin{cases}1 & if \quad Res(p,d)=Res(p_e,d_e),\\ 0 & if \quad Res(p,d)\neq Res(p_e,d_e).\end{cases}$$



a) A 氏, 0° の音響モデル　b) A 氏, 60° の音響モデル　c) A 氏, -60° の音響モデル

図 8.2: A 氏の音響モデルを利用した場合の単語認識率

ここで $r(p,d)$, $Res(p,d)$ は, 話者 p, 方向 d の音響モデルを使用した場合の 単語認識率と入力音声に対する認識結果を示している. また, d_e は実時間人物追跡システムから得られた音源方向であり, p_e は, 評価対象の人物である. $P_v(p_e)$ は顔認識モジュールで生成される確率であり, 顔認識ができない場合は, 常に 1.0 となる. 最終的に, セレクタは最も大きな $V(p_e)$ を持つ人物 p_e と認識結果 $Res(p_e,d_e)$ を出力する.

$V(p_e)$ の最大値が 1.0 以下もしくは, 2 番目に大きい値と近い場合は, SIG は認識が失敗もしくは, 一つの候補に絞りきれなかったと判断して, 音源方向を向き, 該当の人物に再度, 質問を行う. このように, 複数の音響モデルを利用して, 分離音と話者の認識を行う. また, 顔認識が利用可能であれば, 人物名がわかるためロバスト性を向上できる.

8.5 音声認識の評価

同時 3 話者発話のシナリオを通じてロボット聴覚システムを評価した. シナリオの内容を以下に示す.

1. ロボットから 1 m の距離に 60 度間隔 (SIG から見て, 0°, ±60°) で 3 名の人間が並んでいる (図 8.3 参照).
2. SIG は,3 名に質問をする.
3. 各話者は 3 人同時に質問に対する回答を行う.
4. SIG は 3 話者の混合音声の定位・分離・認識を行う.
5. 最終的に,SIG は各話者に向きながら, 向いた方向の人が誰で何を言ったかを答えていく.
6. 音声認識に失敗したと判断した場合は, 該当話者の方向を向いた時に再び尋ねなおす.

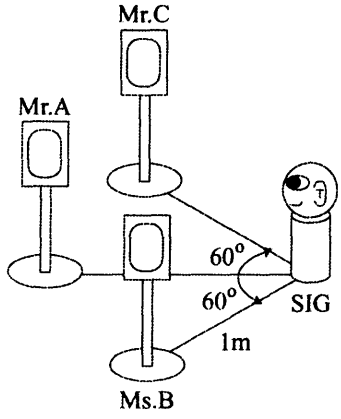


図 8.3: 同時発話時の音声認識実験

なお, 実験では, 実際の人間の代わりにスピーカとその前面に貼られた写真を用いている. スピーカは音響モデル作成時に使用したスピーカと同じものである. 各スピーカから流れる音声は, そのスピーカに貼られた写真と同じ人物のものである. 以下にこのシナリオの結果を 4 例示す.

I. *SIG* が一度ですべての音声を認識できる場合

1. *SIG* が好きな数字に関する質問をする (図 8.4a)).
2. 各スピーカから 1 から 10 までの互いに異なる数字が同時に流れる. ただし, 数字の組合せはトレーニングセットに含まれる組合せと同じものである (図 8.4b)).
3. *SIG* は各音声を実時間人物追跡システムを利用して各音源を定位する. 定位情報を利用して ADPF によりその方向の音声を抽出する. 各分離音に対し 9 つの音声認識プロセスが同時に実行される.
4. 正面の音声の認識を行う. まず, 正面が Nakadai 氏であることを仮定し, 評価を行う (図 8.4c)). 次に, 正面が Arai 氏であることを仮定し, 評価を行う (図 8.4d)). 最後に, 正面が Kyoda 氏であることを仮定し, 評価を行う (図 8.4e)).
5. 結果を統合した結果, 最も適合のよい話者名 (Nakadai), 認識結果 (1) が定まる (図 8.4f)). その結果を正面の話者に告げる (図 8.4g))
6. 同様の処理を他の方向の話者について行う. 各話者に向きながら, 求めた話者名, 認識結果を答える (図 8.4h)-i)).

II. *SIG* が一度ですべての音声を認識できない場合

1. 初期状態は I と同じ. I と同様に, *SIG* が好きな数字を尋ね (図 8.5a)), 各スピーカから同時に回答音声が流れる (図 8.5b)).

a) SIG asks a question.

b) Everyone replies simultaneously.

c) SIG checks if the front person is Nakadai-san.

d) SIG checks if the front person is Arai-san.

e) SIG checks if the front person is Kyoda-san.

f) SIG understood that he is Nakadai-san.

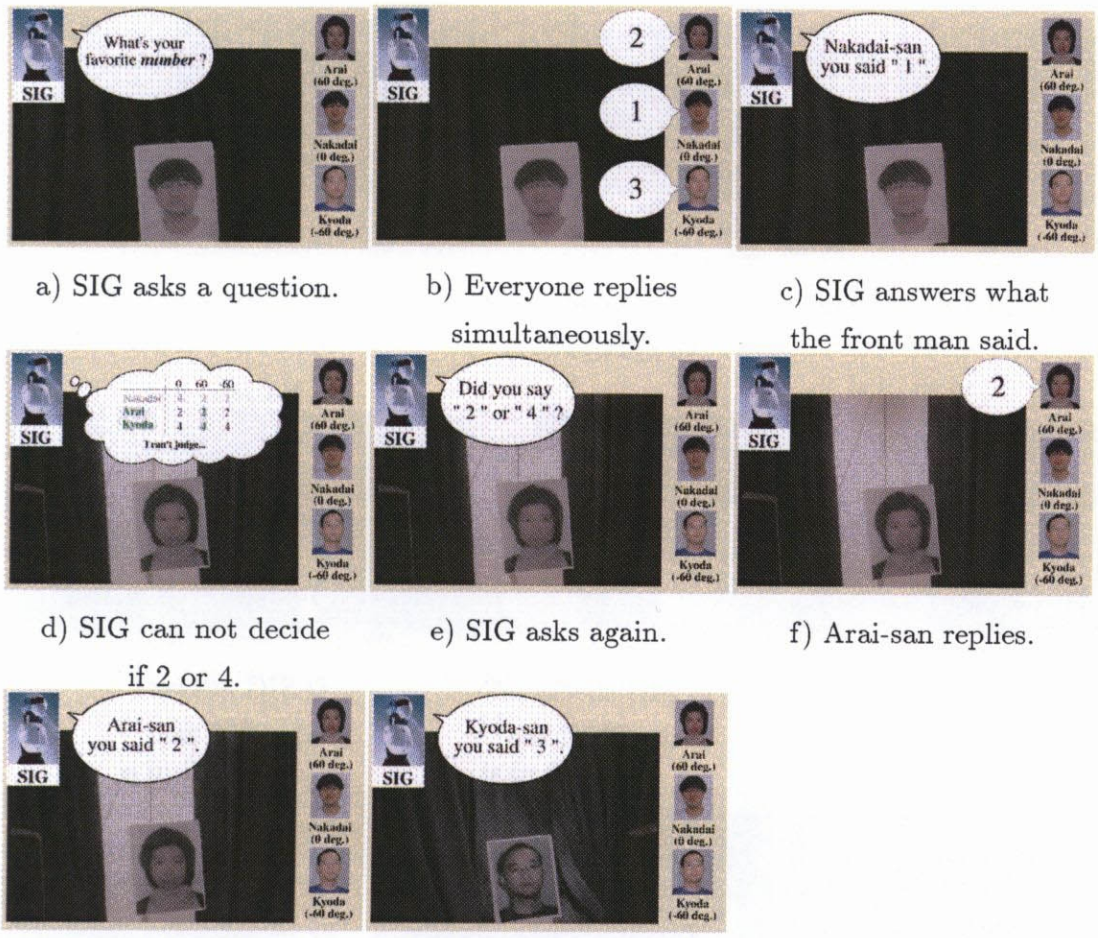
g) SIG answers what he said.

h) SIG answers what the 60° woman said.

i) SIG answers what the -60° man said.

図 8.4: 3 話者同時発話のスナップショット I

2. *SIG* は **I** と同様に, この混合音を分離, 認識する. 正面の音声については, 問題なく認識が可能である (図 8.5c)).
3. しかし, 60° の位置のスピーカから流れた音声が "2" か "4" か区別がつかない (図 8.5d)).
4. そこで, *SIG* は 60° のスピーカの方に体をむけ, 「2 ですか? 4 ですか?」と尋ね直す (図 8.5e)).
5. スピーカから正しい答え "2" が流れる (図 8.5f)). この場合, スピーカは正面方向にあるため, 分離, 認識が容易である (図 8.5g)).
6. **I** と同様, 他のスピーカについても話者名, 認識結果を答える (図 8.5h)).



a) SIG asks a question. b) Everyone replies simultaneously. c) SIG answers what the front man said.

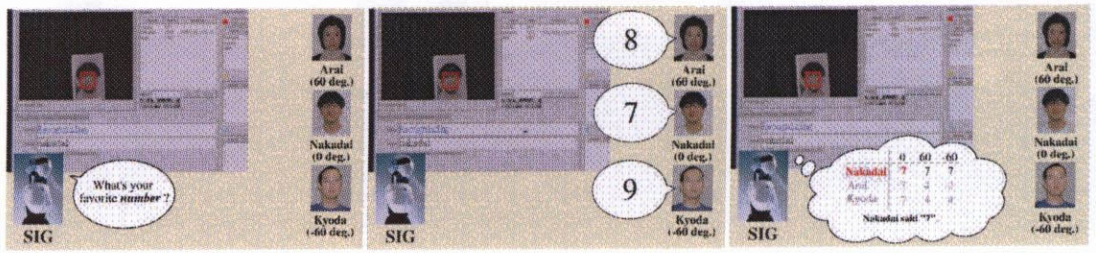
d) SIG can not decide if 2 or 4. e) SIG asks again. f) Arai-san replies.

g) SIG understood what she said. h) SIG answers what the -60° man said.

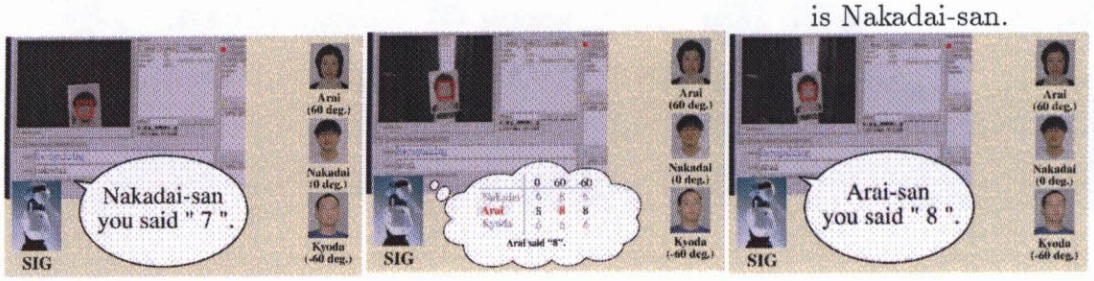
図 8.5: 3 話者同時発話のスナップショット II

III. 顔認識と音声認識を統合した場合

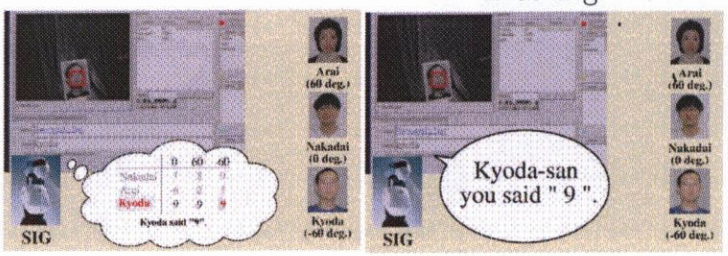
1. 初期状態は **I** と同じ. **I** と同様に, *SIG* が好きな数字を尋ね (図 8.6a)), 各スピーカから同時に回答音声 (7,8,9) が流れる (図 8.6b)).
2. *SIG* は **I** と同様に, この混合音を分離, 認識する. 正面の人物は Nakadai 氏である確率が高いという情報から, その確からしさを統合時に考慮する (図 8.5c)). これにより, より正確な認識が可能である (図 8.5d)).
3. 向きを変えると, 60 度の人物は Arai 氏であることがわかるので, 正面の人物の場合と同様に, その確からしさを考慮する (図 8.5e)). この場合は, 顔認識の情報が利用できないと 6 か 8 か区別がつかない状況であるが, 顔認識の情報により, 曖昧性が解消されている (図 8.5f)).



a) SIG asks a question. b) Everyone replies simultaneously. c) SIG uses info that the front person is Nakadai-san.



d) SIG answers what he said. e) SIG checks info that Arai-san is 60 deg. f) SIG answers what she said.



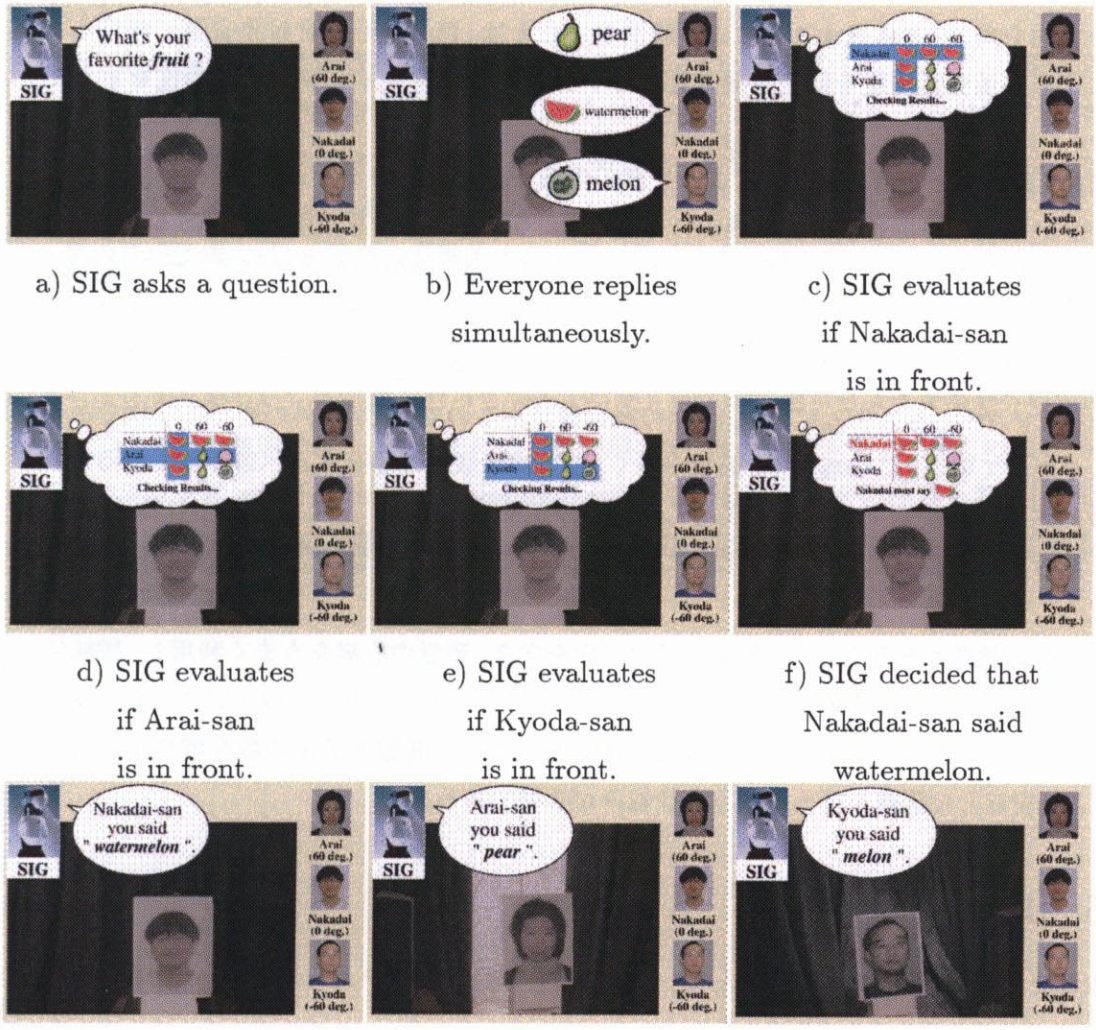
g) SIG uses info that Kyoda-san is -60deg. h) SIG answers what he said.

図 8.6: 3 話者同時発話のスナップショット III

4. 他の話者についても同様の処理を行い正しい認識が行える (図 8.5g),h)).

IV. 数字以外 (フルーツ名) の場合

1. 初期状態は **I** と同じ. 今度は, *SIG* が好きなフルーツを尋ね (図 8.7a)), 各スピーカから同時に回答音声 (スイカ, なし, メロン) が流れる (図 8.7b)).
2. *SIG* は **I** と同様に, この混合音を分離し, 正面の人物が Nakadai 氏であるという前提の下で, 認識結果を評価する (図 8.5c)). 同様に, Arai 氏, Kyoda 氏についても結果の評価を行う (図 8.5d),e)).



a) SIG asks a question.

b) Everyone replies simultaneously.

c) SIG evaluates if Nakadai-san is in front.

d) SIG evaluates if Arai-san is in front.

e) SIG evaluates if Kyoda-san is in front.

f) SIG decided that Nakadai-san said watermelon.

g) SIG answers what he said.

h) SIG answers what she said.

i) SIG answers what he said.

図 8.7: 3 話者同時発話のスナップショット IV

3. 評価の結果, 正面の音声は, Nakadai 氏がスイカといっている可能性が高いと判断し (図 8.5f)), 正面の人に回答する (図 8.5g)).
4. 他の方向の音声についても同様の処理を行う (図 8.5h),i)).

結果 **I** では, 3 話者の認識がすべて成功しており, 同時発話の場合でも, ロボット自身のマイクを使った音声の定位・分離・認識を行うロボット聴覚システムの有効性を示すことができた.

結果 **II** では, 側面方向の音声の認識に曖昧性が生じている. これは, 側面方向の音声の分離精度が悪いという聴覚中心窩の影響を示すものである. また, この曖昧性が対象音

源の方向を向き, 訊き返すことにより解消することは, 聴覚中心窩を対象音源に向けることにより音源分離精度を向上させるというアクティブオーディションの有効性を示している.

結果 III では, 事前に顔認識によって顔の名前がわかっているときには, 誰が話しているかという情報が確信度つきで得られるため, より精度の高い認識が可能である. 特定の環境での利用を前提とした場合などでほぼ 100%近い顔認識処理の精度が得られる場合は, 顔認識情報を信頼できる情報とみなすことができるので, 音声認識で使用する音響モデルの数を削減することができる. この場合, より高速で正確な認識が可能であろう.

結果 IV では, システムには, 150 語の語彙数があるので, 数字以外の認識も可能であることを示している. この場合, 音節数が多く, 認識率は数字の場合と比べ若干低くなる.

最終的にどの場合でも認識が成功している事例を示したが, 図 8.2 に見られるように, 各分離音声の認識率は高々 80 %程度である. 従って, しばしば音声認識に失敗する場合が生じる.

顔認識の情報を利用したり, 対象音源の方向を向き, 聴覚中心窩をうまく利用し, 曖昧性を解消するように訊き返すようなアクティブオーディションを用いて解決できるものもあるが, すべてのケースにおいて当てはまるわけではない. 音響モデルの数を増やすなどして, より安定した認識を行えるような仕組みが必要であろう.

8.6 まとめ

本章では, 音源分離の応用例として, 分離音の音声認識システムを構築し, その有効性を示した.

混合音から一つの音声を完全に分離抽出することは難しく, 分離音声には, なんらかの歪みが生じてしまう. 従来の音声認識研究では, ノイズの混ざった音声から音声を分離して, 認識率を高めるという手法は見られたが, このような分離音声の認識は難しく, 複数の音声が同時に存在するという状況で, その両方を認識するという研究はほとんど見受けられなかった. また, 実際にロボットに搭載して実環境で分離音認識を行うという試みもほとんど報告されていない.

本章では, これを話者・方向毎にトレーニングした音響モデルを用い, これらの結果を統合することによって解決の糸口を探った. 実際に, いくつかの例を示し, その有効性を明らかするとともに, 前処理として使用した音源分離が「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」を持っていることを示している. また, 音声認識においても, これまでと同様, 音源方向を向くというアクティブな動作を行ったり, 顔認識と組み合わせたりすることによって認識率が向上することを示した. これは, 音声認識におけるアクティブオーディションの有効性, 視聴覚統合の有効性を示

している.

音源方向を向くという動作は, ロボット聴覚の向上という意味だけではなく, 人間とのフレンドリーなインタラクションを実現したり, テレイグジスタンスによる会議では, 相手の注意を向けさせたりという意味でも重要であろう. より自然なインタラクションという面については, 次章で述べる.

本章の音声認識では, 9 種類の音響モデルを用いた手法を紹介した. しかし, 話者数や音源方向を増やした場合の検証は行っておらず, このようなパラメータのスケーラビリティの検証は今後の課題である. また, 今回は, 話者をスピーカで代用して, 方向および, 発話のタイミングを固定にして実験を行っている. パラメータが変動するような場合への対応についても今後の課題である. より制約の少ない環境で, 音声認識率を上げるためには, 音声認識エンジンに missing data や missing feature など分離データの性質を考慮した改良 [18, 105] が有効であると考えられる. このような対応についても, 今後行う予定である.