実環境下における
ロバスト音声インタフェースの研究

システム情報工学研究科
筑波大学

2006年7月

山本潔

目 次

図 目 次

表 目 次

第1章　序論

1.1　研究の背景

近年，家庭などで使われる家電製品が急速に発展している．特に，テレビやビデオは地上波デジタル放送やインターネット接続に対応するなど，多機能化が進んでいる．一方，パーソナルコンピュータもテレビ放送の視聴や録画ができるなど，家電との融合が進んでいる．これらの機器を操作するインタフェースとしては，現在の所，キーボードやマウス，リモコンが主流である．しかし，これらのインタフェースは，利用時に手がふさがってしまう，特にリモコンはボタンの数が多く操作が煩雑であるといった問題がある．このため，より人間にとって違和感のないインタフェースを用いる事が必要となる．この様なインタフェースとして音声を用いたインタフェースが考えられる．このように音声を用いて機器に対して操作や指示を行うためのインタフェースをこれ以降音声インタフェースと呼ぶ事とする．

人間が音声インタフェースを介して機器を操作する場合，口元にマイクロホンを設置し，音声を高いSN比（例えば60 dB以上）で収音できる理想的な場合は少なく，

- 機器と人間の間にある程度の距離がある．
- 室内のような，反射や残響が存在する環境下で使用する．
- 話者の音声以外にも何らかの雑音が存在する．

などの理由により，収音時のSN比が低下し，音声の認識が困難な場合が多い．図

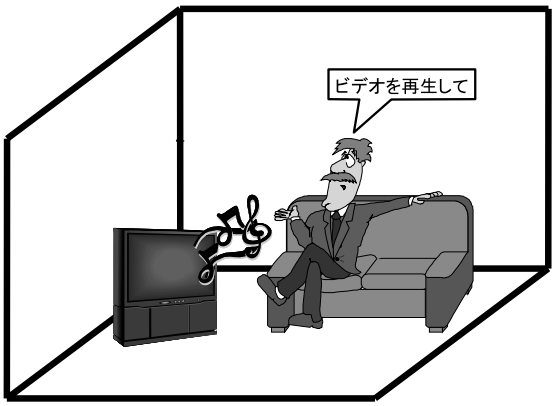

図 1.1: 実環境下における音声インタフェースの利用例

1.1 は本論文が想定している実環境下における音声インタフェースの利用例である．この図では，テレビが発音している状況下で人間が音声を用いてビデオの再生を指示している．本論文では，この図で示すように，室内などの反射や残響が存在する環境において，雑音源が発音し続ける中，話者が機器に対して発話するという状況を想定している．このような状況下では，任意のタイミングで発せられる話者の発話を検出する必要がある．また，話者の音声の他に雑音が存在する場合は話者の音声のみを分離・抽出する必要がある．

以上で述べたように，実環境下で音声を用いるためには，音声認識の技術のみならず，発話検出，雑音抑圧といった様々な技術との融合が不可欠である．次節では，音声インタフェースで用いられる従来の要素技術を概観し，その問題について述べる．

1.2　音声インタフェースのための要素技術

本節では，音声インタフェースを実現する上で必要となる要素技術について，それらの技術を音声インタフェースに用いる際に問題となる点について述べる．

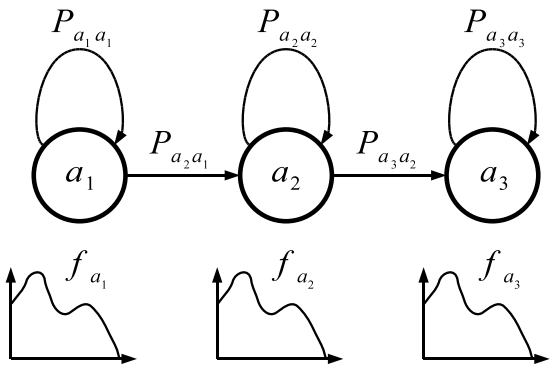

図 1.2: HMM の例

1.2.1 音声認識

音声インタフェースでは，話者の発話した内容を認識するために，音声認識の技術が必要となる．一般に人間は，まず最初に発話する内容（文章）を考え，それに従って実際の音声を発話する．音声認識の処理は，これに対し，まず最初に話者が発した音響信号から適当な特徴量 y を算出する．この，入力音声の特徴量に対して，認識結果 $\hat{w}$ の事後確率 $P(\hat{w}|y)$ を最大にするような $\hat{w}$ を求める．式で表せば以下の通りである．

$$P(\hat{w}|y) = \max_{\hat{w}} \frac{P(y|\hat{w})P(\hat{w})}{P(y)} \tag{1.1}$$

ここで，$P(y)$ は $\hat{w}$ とは無関係であるため，上述の最大化問題においては，$P(y|\hat{w})$ と $P(\hat{w})$ が分かればよい．$P(y|\hat{w})$ を算出するモデルを音響モデル，$P(\hat{w})$ を算出するモデルを言語モデルと呼ぶ．言語モデルは認識結果の文や単語がどの程度確からしいかを表現するものであり，音響モデルは認識結果と音響信号の対応を表現したものである．

式（1.1）を求める際に必要となる音響モデルについては，現在，隠れマルコフモデル（HMM; Hidden Markov Model）による手法が広く用いられている．HMM は音素や単語などの，モデルを表現する単位に対して複数の状態を割り当て，各

モデルを状態間の遷移確率及び各状態におけるシンボルの出現確率を用いて表現したものである．図 1.2 は HMM の一例を示した図である．図において，a_1，a_2，a_3 が状態，$P_{a_j a_i}$ が状態 a_i から a_j へ遷移する確率，f_{a_i} は状態 a_i におけるシンボルの出現確率を確率分布で表したものである．このように，HMM では状態間の遷移とシンボルの出現を確率値で表現する事により，発話の時間的な揺らぎ，及びスペクトル上での揺らぎを含む形でモデルを表現する事ができる．

現在，音声認識においては HMM を用いて音響モデルを表現する手法が確立されている．この手法は基本的に接話マイクロホンを使用して高い SN 比で観測された音声から作られた音響モデルを用いている．しかし，周囲に雑音源が存在し，認識対象となる目的音声が高い SN 比で観測できない場合は，音響モデルと入力信号の間にミスマッチが生じ，認識性能が低下する．この問題に対処するために，音響モデルを雑音が重畳した音声を用いて構築するというアプローチが考えられる．このアプローチでは，事前にどのような雑音が重畳するか既知である必要がある．しかし，一般にそのような情報を事前に知る事は困難である．また，この他にも，音響モデルに対してオンラインでモデル適応の処理を行う技術が存在する．しかしこの方法も，特に SN 比が極端に悪い場合は，音声の特徴量が雑音のそれに埋もれてしまい，音声認識が困難となる．

1.2.2　音源分離

先にも述べた通り，音声認識において用いられる音響モデルは通常，接話マイクロホンを用いて目的音声のみが収録されたデータで学習する事が多い．このため，入力信号に雑音が重畳すると，クリーンな音声を用いて学習したモデルとミスマッチが生じ，音声認識率が低下する．従って，音声認識器に入力する前に，音声と雑音を分離する技術が必要となる．

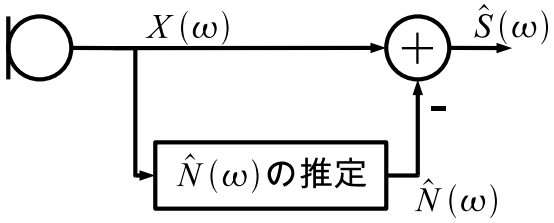

図 1.3: シングルチャネル信号処理のブロック図

シングルチャネル信号処理

図 1.3 はシングルチャネル信号処理の処理の流れを示したものである．シングルチャネル信号処理では単一のマイクロホンを用いて信号を観測し，雑音除去を行う，例えばスペクトルサブトラクション [1] においては，以下の式で示す方法で音源分離を実現している．

$$\hat{S}(\omega) = X(\omega) - \hat{N}(\omega) \tag{1.2}$$

ここで，$X(\omega)$ はマイクロホン入力，$\hat{N}(\omega)$ は雑音の推定値，$\hat{S}(\omega)$ は分離結果のスペクトルをそれぞれ表す．シングルチャネル信号処理においては $\hat{N}(\omega)$ を正確に推定する事が重要となる．例えば，[1] では，入力信号の先頭部分を用いて雑音の推定を行っている．これは，入力信号の先頭部分は雑音のみが重畳しており，この雑音が目的音部分にも連続して重畳している，という仮定に基づいている．

しかしながら，雑音が非定常である場合は，適用が困難である．また，SN 比が低く目的音声のスペクトルが雑音のそれに埋もれてしまうような場合は効果が期待できない．

マルチチャネル信号処理

図 1.4 はマルチチャネル信号処理のブロック図を示したものである．マルチチャネル信号処理では複数のマイクロホンを用いて，音源からの音がマイクロホンに到達する際のマイクロホン間の時間差を用いて音源分離を実現している．図 1.4 に

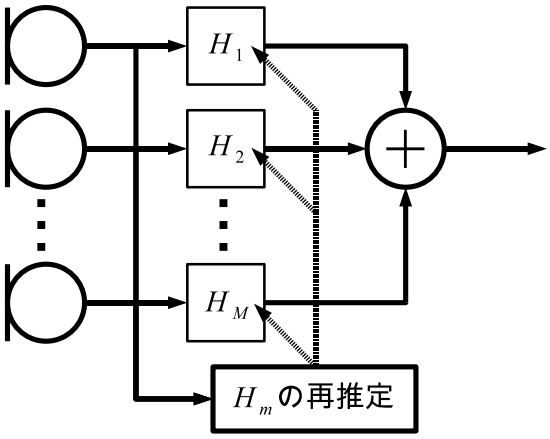

図 1.4: マルチチャネル信号処理のブロック図

おいて $H_1, \ldots, H_M$ は各マイクロホンに対するフィルタ係数を表している．また，M はマイクロホンの数である．マルチチャネル信号処理ではこの H_m を制御する事により分離対象となる音源に向けて指向性を形成したり，雑音源の方向に死角を形成して音源分離を実現している．

しかし，$H_1, \ldots, H_M$ が固定であると，目的音源や雑音源の位置が移動するなど環境が変化した場合に，その変化に追従する事ができない．この問題に対応するため，環境の変化を検出し，それに対応して H_m を更新する事が必要となる．

例えば，ブラインド信号分離（BSS; Blind Source Separation）[2] においては，図 1.4 の出力が独立となるよう H_m が制御される．BSS では，分離対象となる音源同士が独立である事のみを用いているため，音源の種類や位置，マイクロホンアレイの形状に関して事前に知識を必要としない．また，最近では分離対象となる音源（例えば，目的音源と雑音源）以外に，背景雑音が存在する環境下で BSS を実現する手法 [3] も提案されている．しかし，BSS では一般に学習が収束するまでに数秒程度の時間が必要とされ，音声インタフェースのように短いコマンドを発話する場合などは，学習に必要な十分な長さの入力信号が得られない場合もある．また，音声インタフェースなどの応用では，マイクロホン配置などの情報は既知であるため，これらの事前知識を活用する手法の法が合理的である．

また，適応ビームフォーマ（ABF; Adaptive BeamFormer）[4] では，事前情報を用いて音源分離を行う．例えば最小分散（MV; Minimun Variance）法では，以下の式で分離フィルタが得られる．

$$\mathbf{w}_{MV}(\omega) = \frac{\mathbf{R}^{-1}(\omega)\hat{\mathbf{a}}(\theta,\omega)}{\hat{\mathbf{a}}^{H}(\theta,\omega)\mathbf{R}^{-1}(\omega)\hat{\mathbf{a}}(\theta,\omega)} \tag{1.3}$$

ここで，$\mathbf{w}_{MV}(\omega) = \begin{bmatrix} H_1(\omega) \\ \vdots \\ H_M(\omega) \end{bmatrix}$ である．また，$\mathbf{R}(\omega)$ はマイクロホン入力の空間相関行列であり，入力信号から求められる．一方，$\hat{\mathbf{a}}(\theta,\omega)$ は分離対象となる音源から各マイクロホンへの伝達関数を要素とする位置ベクトルであり，事前情報として与える必要がある．このようにMV 法はBSS に比べ，分離対象となる音源の位置に関する情報を必要とする．このため，この情報を別途推定する必要があるが，分離性能はBSS より高い事が期待される．

また，MV 法よりもさらに事前情報を必要とする手法として最尤推定（ML; Maximun Likelihood）法がある．ML 法では，以下の式で分離フィルタが得られる．

$$\mathbf{w}_{ML}(\omega) = \frac{\mathbf{K}^{-1}(\omega)\hat{\mathbf{a}}(\theta,\omega)}{\hat{\mathbf{a}}^{H}(\theta,\omega)\mathbf{K}^{-1}(\omega)\hat{\mathbf{a}}(\theta,\omega)} \tag{1.4}$$

ここで，$\mathbf{K}(\omega)$ は雑音のみの空間相関行列であり，雑音源の位置が定常であれば，目的音源の休止区間から推定される．このようにML 法はMV 法で必要となる分離対象の音源位置に加え，分離対象の音源の休止区間の情報も必要とする．このため，実際にML 法を用いる際には事前に推定すべき情報が多く煩雑である．しかし仮にそれらの情報がうまく推定できれば，MV 法よりもさらに高い分離性能が期待される．

1.2.3 発話区間・目的音源検出

そもそも音声認識器は，入力された信号に対して認識処理を行い何らかの結果を出力する．しかし，観測信号のうち実際に話者が発話しているのは音声インタ

フェースへの入力の一部である．観測信号のうち話者が発話していない区間の信号を音声認識器への入力とすると，意図しない認識結果が出力され，認識誤りとなる．また，先に述べた音源分離の技術において発話区間情報は，分離を行う際の事前情報となる．さらに，空間に複数の音源が存在する場合は，どの音源が音源分離や音声認識の対象となる目的音源（話者）であるかを同定する事で，同様に事前情報として用いる事ができる．

接話マイクロホンなどを用いてある程度高い SN 比で目的音声を観測できる場合には，入力信号のエネルギーをもとに発話区間を検出 [5, 6] する事が可能である．この手法は，入力信号のエネルギーに対して閾値を設け，閾値を越えた区間を発話区間として検出する手法である．しかし，発話の始端と終端は子音であるなどのためパワーが小さく，正確に検出する事は困難である．この問題を解決するために，閾値によって検出された発話区間の前後数フレームも発話区間に含める事が考えられる．また，その他にも，発話の始端と終端に存在する事が考えられる摩擦音を零交差回数に対する閾値で検出（例えば，[7]）する方法がある．しかし，これらの方法は全て入力信号のエネルギーを基準としているため，雑音のエネルギーが目的音声のエネルギーと同程度である場合は，雑音の区間を発話区間と誤って検出してしまう．そこで，音声と雑音の音響的な違いに着目した VAD の手法が提案されている．

例えば [8] では音声信号中に含まれるホルマントを検出し，ホルマントが存在する区間を発話区間とする手法である．また [9] では，あらかじめ雑音区間においてケプストラムやスペクトルの平均値を求めておき，入力信号のケプストラムやスペクトルとのユークリッド距離を求めている．距離が近ければ入力信号は雑音の特性により近いと判断し，距離が遠ければ雑音とは異なる，すなわち目的音声と判断し，その区間を発話区間として検出する手法である．さらに，[10] では入力信号の尖度（kurtosis）を求め，その形状によって信号の種類を識別している．入力信号の尖度は，信号の種類（例えば人間の声と物音）によってその形状が異なる事

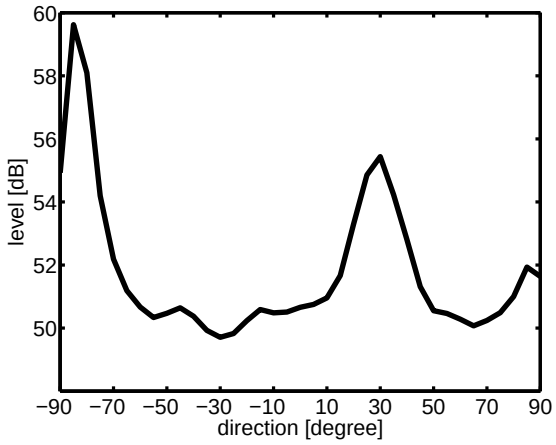

図 1.5: 空間スペクトルの例

が知られている [2]．そこで，[10] ではこの形状の違いを用いて信号を識別し，音声と識別された区間を発話区間として検出している．

これらの手法は，音声と雑音の音響的な特徴を用いて発話区間を検出しているため，SN 比が比較的悪い条件下でも目的音声の区間のみを発話区間として検出する事が可能である．しかし，例えばテレビから音声が流れるように，雑音が音声であった場合は目的音声だけでなく，雑音の区間も発話区間として検出してしまうという問題がある．

一方，音源定位の技術 [11] を用いる事により，環境に存在する音源がいつ，どの方向で発音したのかを知る事ができる．例えば，カーナビゲーションのように，目的音源の位置が事前にある程度限定されている場合は，ある角度範囲で発音した音源を目的音源として検出する事が可能である [12]．

図 1.5 は音源定位の一手法である MUSIC 法 [11]（サブスペース法）により得られる空間スペクトルの例である．図において，空間スペクトルのピークが存在する角度が音源の推定位置になる．この例では，−80 度と 30 度の方向に音源が存在すると推定される．

このように，サブスペース法を用いる事で，音源の位置を推定する事が可能で

ある．例えば，目的音源と雑音源が同時に発音している場合，音源の位置を 2 つ推定する事は可能である．しかし，目的音源の存在範囲が限定できないような場合は，推定された 2 つの音源位置のうち，どちらが目的音源で，どちらが雑音であるかという事は分からない．そのため，この手法を用いて目的音源から発音しているかを判別する事は困難である．

このように，雑音環境下で音声インタフェースを用いる上で必要となる音声認識，音源分離，発話区間検出の技術は，それぞれの手法を単独で用いる事が困難である場合が少なくない．上述のように，音声認識技術は基本的に高い SN 比で話者の発話を収録できる事を前提としており，本論文で想定するような雑音が存在する環境下において認識性能が低下してしまう．また，音源分離において，特に高い性能が期待される ML 法では，分離対象となる音源の位置や発話区間に関する情報を必要とする．発話区間検出においては，雑音源に音声が含まれる場合，雑音を発話区間として誤検出してしまう可能性がある．目的音源の検出においても，複数の音源が存在する場合，推定された音源位置のどれが目的音源に対応するかを特定するのが困難である．しかしながら，ここで述べた要素技術及び他の要素技術とを融合する事により，それぞれの要素技術の問題を克服できる可能性がある．このような研究アプローチは情報統合と呼ばれ，近年盛んに研究がなされている．

1.3　音声インタフェースの応用

続いて，音声インタフェースの応用例を概観する．まず，自動車の車内における例である．自動車内には空調やカーステレオやカーナビゲーションといった機器が存在する．これらの機器の操作を行う場合，手を機器のスイッチへ動かし，視線も機器の方に向ける必要がある．しかし，走行中に運転手がこれを行う事は，交通安全の観点から好ましくない．このような場合に音声インタフェースを用いる事が可能であれば，運転手は視線などを動かす事なく音声による指示のみで機器を

操作する事ができる．このため，車内における音声インタフェースの研究が行われている．例えば [13] では，車内にマイクロホンアレイを設置し，ハンズフリー音声認識を実現するための音声インタフェースを提案している．この際，カーステレオから発音している環境下で，話者の音源位置を推定するために，音源定位の技術に加え調波構造を用いて音源位置を推定する手法を提案している．この手法を用いる事により，カーステレオと区別して話者の音源位置を推定する事が可能である．また，[14] では音声インタフェースを実現する上で重要となる音源分離の技術について，車内で収録したデータを用いて検討を行っている．

また，家庭内環境で人間に対して様々なサービスを提供する事を目的としたロボットの開発が近年盛んに行われている．ホンダの ASIMO [15]，産業技術総合研究所の HRP-2 [16]，NEC の PaPeRo [17] などはその一例である．これらのロボットは人間に対して何らかのサービスを提供する事を目的としている．このため，人間とロボットの間でコミュニケーションをとる必要がある．この際，人間にとって自然な形態である，音声を用いたコミュニケーションが極めて重要である．例えば [18] では，ユーザからの観光案内の問い合わせに対して身振りを交えて対話するロボットの構築を行っている．また，[19] では複数の話者との対話において，会話の状態を把握し適切なタイミングで会話に参加するロボットの構築を行っている．このように，人間といかに対話を行うかという事は，スムーズなコミュニケーションには重要な要素である．また，家庭やオフィス環境では，雑音源が存在する事も多く，音源分離などの技術をロボットに実装した研究 [20, 21, 22] も行われている．ロボットとの対話の研究である [19] では，人間の音声の入力に接話マイクロホンを用いている．このようなロボットとの対話において，雑音源が存在する環境下でも人間同士の会話と同等な方法で音声入力が行えれば，人間にとって違和感のないより自然な対話になる事が期待される．このように，人間にとって違和感のない自然な対話をロボットとの間で実現するためには，このような音声インタフェース技術の向上が切望されている．

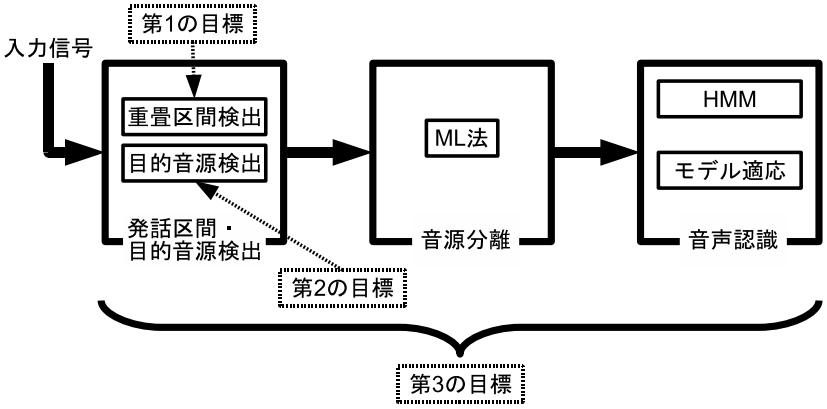

図 1.6: 音声インタフェースで用いる要素技術

また，これらのロボットは人間の要求に応じて室内を移動したり，場合によっては戸外に出る事も考えられる．この場合，ロボットに実装された機能をロボットの外部にあるハードウェアで処理すると，データの送受信の遅延が問題になる可能性がある．また，ロボットが任意の地点で外部のハードウェアと通信できるとも限らない．そのため，各種機能を実装する際は，それらを実現するハードウェアがロボット内部に実装されている事が望ましい．しかし，ロボットの内部に搭載できるハードウェアは物理的なスペースや消費電力の点で著しい制約を受ける．そのため，[18, 19] ではロボットの外部に PC を用意し，音声認識や画像処理を行っている．また，[22] ではロボットの内部に実装するために，処理量の少ない雑音抑圧の手法を提案している．このように，音声インタフェースでは限られた資源・制約の中でこれを実現する実装技術に関する研究も重要である．

1.4　研究の目的

1.2 節で述べた通り，本論文が想定する環境下で音声インタフェースを実現する場合，様々な問題が存在する．発話区間検出，音源位置推定，音源分離，音声認識など音声インターフェースを実現する上で有効な要素技術は従来から研究されている．しかし，環境に存在する雑音が非音声であるなどの制約が存在する，使用

可能な環境が著しく限定される，学習に時間がかかるなど，実環境でのインターフェースとしては効果が不十分であるものが少なくない．しかしながら，これらの要素技術をお互いに組み合わせたり，新たな要素技術と融合する事により，実環境でも実用に耐えうるインターフェースの構築が可能であると考えられる．前節でも述べたように，特に異なるモダリティ間の情報や要素技術を融合する研究は，情報統合（Information Fusion）と呼ばれ，新たな研究領域として注目されている．本論文では，この情報統合の考え方に基づいた実環境でロバストに動作する新たな音声インターフェースの構築を最終的な目的とする．以下に，この最終目的を達成するための具体的な目標を述べる．

図 1.6 は本論文で用いる音声インタフェースの要素技術とその関係を示した図である．本論文では，1.2 で述べたように，雑音の分離性能が高く目的となる音声の歪みも少ない，マイクロホンアレイを用いたマルチチャネル信号処理を音声認識の前処理として採用する事を考える．特に ML 法は，環境における事前情報を与える事により高い性能が期待される．ML 法を用いるためには，目的音源および雑音源の空間的な性質が事前情報として必要である．これらを推定するためには，目的音源に雑音源が重畳しているかどうかの情報が重要となる．目的音源と雑音源が重畳しているかどうかは，環境に存在する有効音源数を推定する事により判定する事ができる．しかしながら，後述するように，従来の手法は主にレーダやソナーなどの分野で開発され，本論文が想定するような部屋の反射や残響が存在する環境下では高い性能が期待できない．そこで，雑音と音声の重畳区間を検出するあらたな要素技術の開発を本論文の第 1 の目標とする．

また，ML 法を用いた雑音抑制法では環境に存在する音源のうち，どの音源が目的音源かを同定する事が必要不可欠である．従来の研究では，目的音源の位置に制約を設けるなどの方法をとってきた．しかし，特にロボットのように動的な環境で動作するアプリケーションのための音声インターフェースにおいては，このような制約は音声インタフェースとしての利便性を損なう．そこで，本論文では，

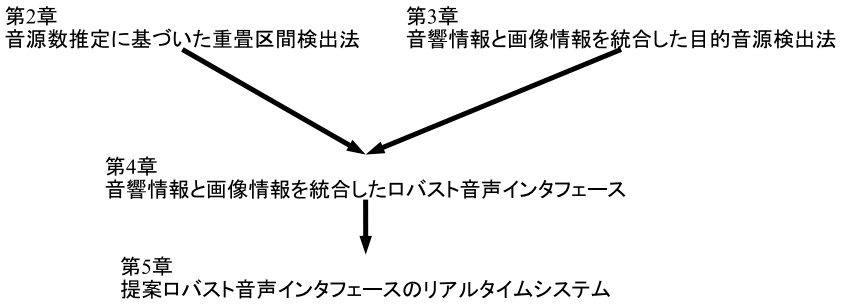

図 1.7: 章間の関係

マイクロホンアレイから得られる音響情報と，カメラから得られる画像情報とを統合する事により，音源の位置に制約を設けずに目的音源を同定する事が可能な新たな手法を開発する事を第 2 の目標とする．

第 3 の目標は，上述の 2 つの目標を達成する事で開発される新たな要素技術と従来の音源定位，音源分離（ML 法），音声認識における音響モデルの適応技術などをさらに統合し，本論文の最終目的である音声インターフェースを構築する事である．ここでは，要素技術の統合により相乗効果が得られるようなシステムを構築する事が目標である．また，特に音声インターフェースの研究では，実時間で動作するシステムを構築し，実環境で実証実験を行う事が極めて重要である．第 3 の目標では，本論文で想定しているロボットのプラットフォームに提案する音声インターフェースを実装し，実環境においてその性能の検証を行う．

1.5　本論文の構成

図 1.7 は本論文における章間の関係を示したものである．本論文ではまず，第 2 章において, 第 1 の目標を達成するため，音源数推定に基づいた重畳区間の検出手法を提案する．この手法は，複数の音源が同時に発音している場合に，これを重畳区間として検出する手法である．本章において行う重畳区間の検出は，話者の発話区間を検出する上での重要な情報となる．

次に，第 3 章では，第 2 の目標を達成するため，音響情報と画像情報を用いた目的音源の同定法を提案する．先に述べた通り，目的音源の音源位置を推定する際に複数の音源位置が得られた場合，そのどれが目的音源の音源の位置であるかを判別する事は困難である．この手法では，音源位置の推定結果，カメラ画像から得られる人物位置の推定結果を統合する事により，話者の位置を検出する．これにより，音響的な情報のみでは難しい，SN 比が低くテレビなどの人の声を含む音源が雑音として存在する場合でも，話者の声を雑音源とは区別して検出する事が可能となる．

第 4 章では，第 3 の目標を達成するために，第 2 章で提案した重畳区間の検出法及び第 3 章で提案した目的音源同定法を他の要素技術と統合し，実環境下でロバストに動作する音声インタフェースを提案する．提案するインタフェースでは，第 2 章と第 3 章で提案した手法を導入する事により，目的音源と雑音源がそれぞれどの音源位置で，いつ発音しているかを検出・同定する事が可能となる．このように，目的音源と雑音源に関する情報を推定する事により，音源分離においてこれらの情報を用いる事ができるようになるため，高い分離性能を達成する事が期待される．インターフェースの性能は、オフラインの実験により、詳細に評価する．

第 5 章では，第 3 の目標及び本論文の最終目的の達成度を評価するため，第 4 章で提案した音声インタフェースをリアルタイムシステムとして実装し実環境での評価を行う．本論文では，音声インタフェースの具体的な応用例として，提案する音声インターフェースをヒューマノイドロボットへ実装する．人間へ様々なサービスを提供するロボットでは，雑音源の存在する人間の生活環境下で音声を用いて人間とコミュニケーションをとる事が必要不可欠である．本論文では，提案する音声インターフェースを実装したロボットをさらに LAN よりテレビなどの情報家電機器などに接続した環境を構築し，実勢の生活環境に近い環境で提案法の評価を行う．

第 6 章は本論文の結論である．

第2章　音源数推定に基づいた重畳区間検出法

2.1　はじめに

第 1 章で述べた通り，本論文が想定する環境下で音声インタフェースを実現するためには，話者の音声を雑音から分離する必要がある．その際，入力信号中のどの区間で話者が発話しているかを特定する事が重要である．例えば，雑音源が常に発音し続ける環境下で話者が断続的に発話する場合は重畳区間の情報を以下のように用いる事ができる．

- 重畳区間の時は，雑音源と話者が同時に発音しており，発話区間となる．
- 非重畳区間の時は，雑音源のみから発音しており，非発話区間となる．

また，雑音源も断続的になる場合は，

1. 話者のみが発話している．
2. 雑音源のみが発音している．
3. 雑音源が発音し続ける中で話者が発話している．

の 3 つのケースが考えられる．この 3 つのケースを重畳区間情報のみで識別する事は出来ないが，1，2 と 3 を区別するための情報としては有用である．

その他にも例えば，数人が参加する小規模な会議において，音声認識技術を用いて議事録や会議の概要を作成する場合にも重畳区間情報は重要な情報となる．この場合，複数の話者がそれぞれ断続的に発話するため，重畳区間であっても非重

畳区間であっても話者が発話している区間である可能性がある．しかし，重畳区間は複数の話者が同時に発話している区間のため，そのままでは音声認識に失敗してしまう．そのため，音源分離技術を用いて重畳区間における発話をそれぞれの話者に分離する必要があるが，重畳区間情報を用いる事で音源分離を行う区間を特定する事ができる．

後述するように，マイクロホンアレイから求めた空間相関行列の固有値が音源数や音源間のパワー差の情報を反映している事が知られている．空間相関行列を固有値分解すると，マイクロホンの数と同じだけの固有値が得られる．このうち，その場で実際に音を出している音源の数と同じ数の固有値が，他の固有値よりもある程度大きい．この性質を用いる事で音源の数を推定し，重畳区間を検出する事が考えられる．背景雑音が白色である場合は，AIC や MDL などの情報量規範を固有値に適用して音源数を推定する手法が提案されている [23]．しかし，本論文で想定する部屋の反射や残響が存在するような実環境下では，これらの手法をそのまま用いる事はできない．背景雑音を白色化する手法も提案されているが，この手法を適用するためには背景雑音が単独で観測できる必要がある．しかし，部屋の反射や残響は音源の信号により生じ，常に同時に存在するものであるため，これらを単独で観測する事は不可能である．また，この他にも背景雑音のレベルを推定し，これを越える固有値の数を数える事で音源数を推定する方法も考えられる．しかしながら，この方法も音源のパワーに依存する背景雑音のレベルを推定する必要があり，実際には困難である．

本章ではこの問題に対処するために Support Vector Machines（SVM）を用いた重畳区間検出法を提案する．空間相関行列を固有値分解する事で得られる固有値分布は音源数に関する情報を反映しているため，パターン識別の手法（ここでは，固有値のパターンを識別）を用いて，音源数を推定する．SVM は 2 つのクラスにクラス分けされた学習データを用いて識別関数を設計する手法として知られている．この識別関数を用いる事で属するクラスが未知であるデータを識別する事が

可能である．そこで，提案手法では，あらかじめ音源数が既知である固有値分布を学習データとして用い，

- 音源数が 1（非重畳区間）
- 音源数が 2 以上（重畳区間）

の 2 つのクラスに分類し，SVM を用いて識別関数を設計する．この識別関数を用いる事で音源数が未知である固有値分布に対して音源数を推定でき，重畳区間であるかどうかが判別可能である．この手法では，学習データに実環境下で得られた固有値分布を用いる．この固有値分布には部屋の反射や残響といった情報も含まれているため，得られる識別関数はこのような環境に対して頑健である事が期待される．

SVM の手法は学習データを用意する段階で固有値分布を 2 つのクラスに分類するため，音源間のパワー差のバリエーションが多い場合には，重畳区間の検出に失敗する可能性がある．また，重畳区間の基準を学習データ生成時に決定する必要があるため，この基準を変更するためには再度学習データを用意し，識別関数を設計し直す必要がある．そこで，本章ではこの問題に対処するためにさらに，Support Vector Regression（SVR）を用いた重畳区間検出法も提案する．

SVR はスカラー値が付与されたベクトルデータである学習データに対して，このデータを回帰する回帰曲線を求める手法として知られている．一方，固有値分布は音源数のみならず，音源間のパワー差の情報も反映している．そこで，SVR による手法では，まず，何らかの方法で音源間のパワー差が既知である固有値分布を用意する．次に，この固有値分布に対して SVR を用いて回帰曲線を求める．このようにして，得られた回帰曲線を用いる事で音源間のパワー差が未知である固有値分布に対して，パワー差を推定する事が可能である．さらに，推定されたパワー差に基づいて重畳区間であるかを判別する事ができる．この SVR による手法は SVM による手法と同様の理由で，実環境に対して頑健である事が期待される．

また，回帰曲線でパワー差を推定後に重畳区間であるかを判別するため，重畳区間の基準が変更されても直ちに回帰曲線を求め直す必要がない．

以下の節では，まずはじめに本論文で重畳区間をどのように定義するかを述べる．次に，重畳区間を検出する上で必要となる固有値分布について述べ，固有値分布を用いた手法における問題点について説明する．さらに，SVM を用いた手法及び SVR を用いた手法について説明を行い，最後に提案手法の評価実験について述べる．

2.2　音源数と空間相関行列の固有値分布との関係

本節では，本章で重要となる音源数と空間相関行列の固有値分布の関係に関する基本的な理論を述べる．この理論は，もともとレーダーやソナーなどの分野において狭帯域信号に対して確立したものであり，この理論を本論文で扱う音声などの広帯域信号に適用するため，入力信号をフーリエ変換し，周波数領域で考えるものとする．各周波数には，上述の理論がそのまま適用できるため，まず，周波数ごとに独立に音源数の推定を行い，この情報をもとに，2.4.5 において，最終的な広帯域信号に対する音源数推定及び重畳区間推定を行う．

本論文では，音声インタフェースの入力としてマイクロホンアレイを用いる事を想定している．そのため，以下ではマイクロホンアレイの入力を利用できるという前提で説明を行う．今，マイクロホンアレイ入力の短区間フーリエ変換を $\mathbf{x}(\omega, T) = [x_1(\omega, T) \cdots x_M(\omega, T)]^T$ と書く事とする．ここで，ω は周波数，T はフレーム番号，M はマイクロホンの数，$\cdot^T$ は転置である．このとき，$\mathbf{x}(\omega, T)$ は

$$\mathbf{x}(\omega, T) = \mathbf{A}(\omega)\mathbf{s}(\omega, T) + \mathbf{n}(\omega, T) \tag{2.1}$$

と書ける．$\mathbf{s}(\omega, T) = [s_1(\omega, T) \cdots s_N(\omega, T)]^T$ と $\mathbf{n}(\omega, T) = [n_1(\omega, T) \cdots n_M(\omega, T)]^T$ はそれぞれ音源と背景雑音を短区間フーリエ変換したものであり，N は音源の数である．音源のパワーは背景雑音のパワーより大きいものとする．また，$\mathbf{A}(\omega)$ は

音源からマイクロホンまでの伝達関数の行列であり，要素 (m, n) が n 番目の音源から m 番目のマイクロホンへの伝達関数となる．

ここで，空間相関行列 $\mathbf{R}(\omega)$ を以下のように定義する．

$$\mathbf{R}(\omega) = E[\mathbf{x}(\omega, T)\mathbf{x}^H(\omega, T)] \tag{2.2}$$

$E[\cdot]$ は期待値を取る操作であり，$\cdot^H$ は複素共役転置である．ここで，音源と背景雑音の間に相関がなければ,

$$\mathbf{R}(\omega) = \mathbf{A}(\omega)\mathbf{P}(\omega)\mathbf{A}^H(\omega) + \mathbf{K}(\omega) \tag{2.3}$$

と記述できる．$\mathbf{P}(\omega) = E[\mathbf{s}(\omega, T)\mathbf{s}^H(\omega, T)]$，$\mathbf{K}(\omega) = E[\mathbf{n}(\omega, T)\mathbf{n}^H(\omega, T)]$ である．さらに，背景雑音間の相関が無相関であれば,

$$\mathbf{R}(\omega) = \mathbf{A}(\omega)\mathbf{P}(\omega)\mathbf{A}^H(\omega) + \sigma\mathbf{I} \tag{2.4}$$

となる [24]．ここで，$\mathbf{I}$ は単位行列，σ は背景雑音の分散（パワー）である．

式（2.4）が成り立つとき，$\mathbf{R}(\omega)$ を固有値展開して得られる固有値は

$$\lambda_1(\omega), \ldots, \lambda_M(\omega) = \overbrace{\gamma_1 + \sigma, \cdots, \gamma_N + \sigma}^{N \text{ 個}}, \overbrace{\sigma, \cdots, \sigma}^{(M-N) \text{ 個}} \tag{2.5}$$

となる．ただし，固有値は降順でソートされているものとする．ここで，式（2.5）により，次の性質が成り立つ事が分かる [24, 25].

1. N 個の音源 $\mathbf{s}(\omega, T)$ のエネルギーは N 個の主な固有値に集中する．
2. 雑音成分 $\mathbf{n}(\omega, T)$ のエネルギーはすべての固有値に分散する．

図 2.1 は上述の仮定

- 音源と背景雑音が無相関である．
- 背景雑音が空間的に白色である．

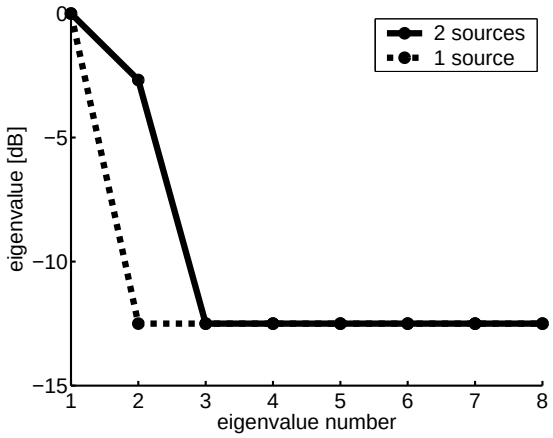

図 2.1: 空間的に白色の場合の固有値分布

- 音源のパワーが背景雑音のパワーより有意に大きい.

が成立する場合の固有値分布を模式的に示したものである（音源数は 2）．図を見ると，式（2.5）で示されている通り，音源数に対応した数の固有値がほかの固有値に比べ相対的に大きな値となっている．また，それ以外の固有値は全て等しい値である．このように，上述の仮定が成り立つ理想的な場合は固有値分布の形状から音源数を容易に推定する事が可能である．

一方，図 2.2 は実環境で得られた固有値分布である．この図を見ると，理想的な場合に得られる固有値分布である図 2.1 とは大きく異なる事が分かる．実環境では部屋の反射や残響が存在しており，これらは先の式（2.1）における背景雑音 $\mathbf{n}(\omega, T)$ に相当する．しかし，一般に反射や残響は音源に対して無相関ではない．また，背景雑音も各マイクロホンで独立ではなく，ある程度相関があり，空間的に白色ではない．このため，式（2.5）に至る過程で用いた前提が完全には成り立たない．しかし，2 音源と 1 音源の固有値分布を比べてみると，2 音源の場合は 2 個目の固有値が大きな値を取るなど，ある程度図 2.1 における固有値分布の性質が見られる部分もある．

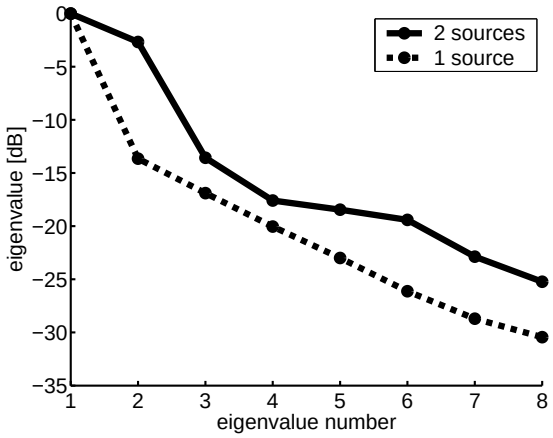

図 2.2: 空間的に白色ではない場合の固有値分布

2.3 固有値分布に基づいた従来の音源数推定法とその問題点

2.2 節で述べた固有値分布と音源数の関係を用いた音源数推定法としては AIC（Akaike Information Criterion）や MDL（Minimum Description Length）を用いる手法が提案されている [23]．AIC や MDL は実際に観測されたデータに最もよく合うモデルを選択するための基準である．そこで，マイクロホンで観測された固有値分布と音源数に対応するモデルを比較し，最もよく合うモデルの音源数を推定音源数とする事を考える．

今，マイクロホン入力 $\mathbf{x}(t)$ を短区間フーリエ変換し，フーリエ係数 $\mathbf{x}(\omega, T_1), \ldots, \mathbf{x}(\omega, T_L)$ が得られたとする．ここで，ω は周波数，T_i はフレーム番号である．さらに，空間相関行列

$$\mathbf{R}(\omega) = \frac{1}{L} \sum_{i=1}^{L} \mathbf{x}(\omega, T_i) \mathbf{x}^H(\omega, T_i) \tag{2.6}$$

を求め，これを固有値分解し固有値 $\lambda_1(\omega), \ldots, \lambda_M(\omega)$ を求める（M はマイクロホ

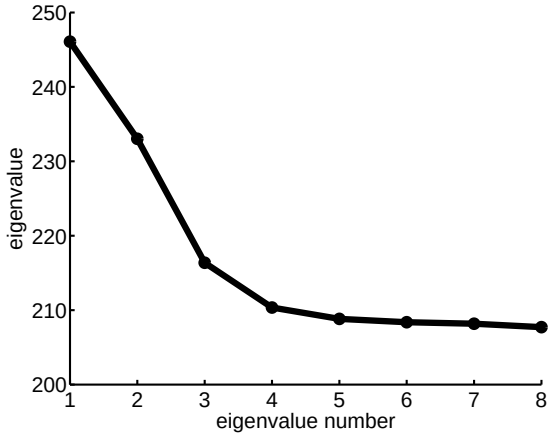

図 2.3: 背景雑音が白色の場合

ンの数）．このとき，AIC による音源数推定は次式で表される [23].

$$AIC(N_s) = -2\sum_{f=f_l}^{f_h} \ln \left[\frac{\left(\prod_{i=N_s+1}^{M} \lambda_i(\omega_f)\right)^{\frac{1}{M-N_s}}}{\frac{1}{M-N_s}\sum_{i=N_s+1}^{M} \lambda_i(\omega_f)} \right]^{(M-N_s)L} + 2(f_h - f_l)[N_s(2M - N_s)] \tag{2.7}$$

また，MDL を用いる場合は，

$$MDL(N_s) = -\sum_{f=f_l}^{f_h} \ln \left[\frac{\left(\prod_{i=N_s+1}^{M} \lambda_i(\omega_f)\right)^{\frac{1}{M-N_s}}}{\frac{1}{M-N_s}\sum_{i=N_s+1}^{M} \lambda_i(\omega_f)} \right]^{(M-N_s)L} + \frac{1}{2}(f_h - f_l)N_s(2M - N_s)\ln L \tag{2.8}$$

f_l と f_h は周波数帯域番号であり，N_s はモデルの音源数である．最終的な推定音源数はそれぞれ，$\mathrm{argmin}_{N_s} AIC(N_s)$，$\mathrm{argmin}_{N_s} MDL(N_s)$ で得られる．

この方法は，閾値などを明示的に与える必要がない．しかし，背景雑音 $\mathbf{n}(\omega, T)$ の各成分が無相関であるという事を仮定している．一般に実環境では，背景雑音として部屋の反射や残響といったものが考えられる．これは，通常，無相関ではない．そのため，この手法は実環境下で十分な性能が期待できない．

図 2.3 と図 2.4 は，それぞれ背景雑音が白色である場合と，白色でない場合の固有値分布の例である．どちらも，音源数が 2 の固有値分布である．この分布に対

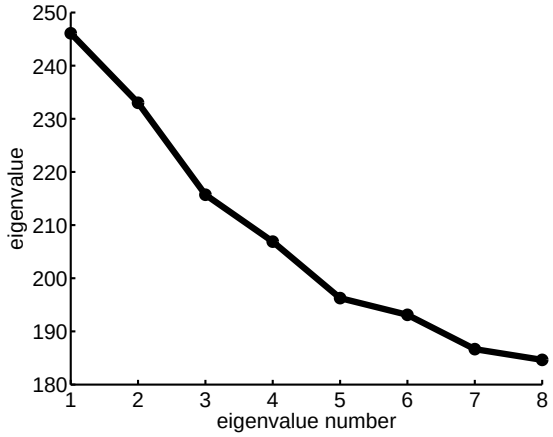

図 2.4: 背景雑音が白色ではない場合

して AIC で音源数を推定した結果を図 2.5 と図 2.6 に示す．図 2.5 では，推定音源数（$AIC(N_s)$ が最小になる N_s）が 2 と正しく推定できている．一方，背景雑音が白色でない場合（図 2.6）は推定音源数が 4 と，正しく推定できていない．

仮に背景雑音 $\mathbf{n}(\omega, T)$ が既知であれば，背景雑音の空間相関行列 $\mathbf{K}(\omega) = E[\mathbf{n}(\omega, T)\mathbf{n}^H(\omega, T)]$ を用いて，次式のように一般化固有値展開を行う事で，背景雑音を白色化できる [24].

$$\mathbf{R}(\omega) = \mathbf{K}(\omega)\mathbf{E}(\omega)\mathbf{\Lambda}(\omega)\mathbf{E}^{-1}(\omega) \tag{2.9}$$

図 2.7 は背景雑音が白色でない場合の固有値分布の例であり，これを式（2.9）に従って白色化した固有値分布が図 2.8 である．このように，背景雑音が既知であれば白色化を行う事は可能である．しかし，実環境では $\mathbf{n}(\omega, T)$ として部屋の反射や残響が想定される．これらは音源の直接音に付随して発生するものであり，これだけを単独で $\mathbf{n}(\omega, T)$ として観測する事は不可能である．

また，これ以外にも，固有値に対して閾値を設定し，閾値を越えた固有値の数を音源数とする方法も考えられる．図 2.9 はこの方法による音源数推定の例である．図において，破線が固有値に対して設定した閾値である．この例の場合，閾値を越えている（閾値より大きい）固有値は 2 個であるので，推定音源数は 2 となる．

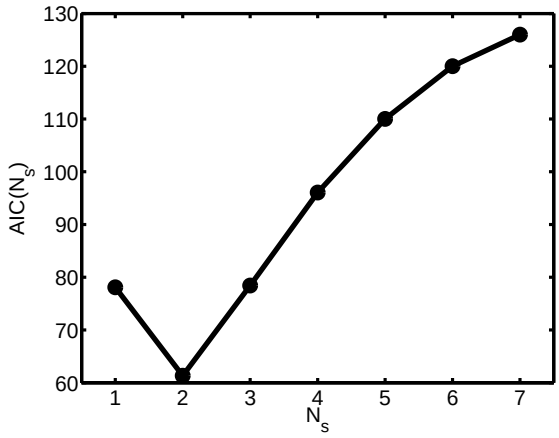

図 2.5: 図 2.3 に対する推定結果

図 2.10 はある帯域における 2 つの音源のパワー差の例である．また，図 2.11，図 2.12，図 2.13 はそれぞれ図 2.10 の 3 秒，4 秒，6 秒の時の固有値分布である．このように，実環境下では，音源のパワー差が時間とともに変化する．また，それにともない，固有値分布のパターンも様々に変化する．そのため，これらの固有値分布に対して単一閾値を設定する事は困難である．

2.4 Support Vector Machines を用いた重畳区間検出法

2.4.1 基本的な考え方

前節までで，実環境下での固有値分布は理想的な場合とは異なり大きく変動するため，従来の手法では音源数がうまく推定できない事を述べた．本節ではこの問題に対処するために，SVM を用いた重畳区間検出法を提案する．SVM は一般に，入力データを 2 つのクラスにクラスタリングする手法として知られている．ここでは，固有値分布を音源数に対する特徴量と考え，SVM への入力に固有値分布を用いる．SVM は入力された固有値分布を 2 つのクラスに分類する．この 2 つのクラスが音源数 1 と音源数 2 に対応していれば，分類されたクラスをその固有値

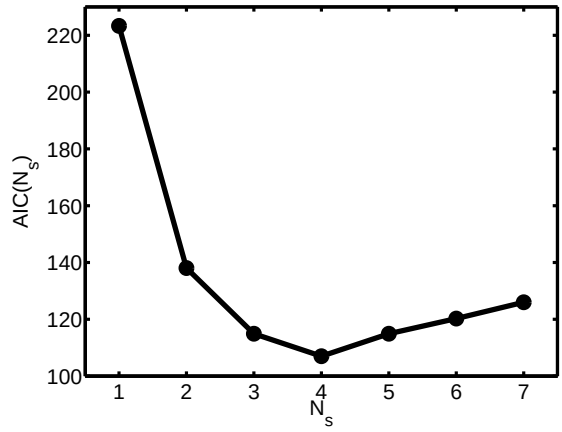

図 2.6: 図 2.4 に対する推定結果

分布の音源数とする事が可能である．重畳区間検出では，この推定音源数に基づいて，音源数が 2 と推定された場合を重畳区間と考える．

図 2.14 は SVM による重畳区間検出のイメージ図である．図において，左側にある 4 つの固有値分布は音源数が未知の固有値分布である．この固有値分布を図の中央の識別関数に入力する事で，2 つのクラスに分類する事ができる．分類後の属するクラスによってそれぞれの固有値分布を持つ区間がが重畳区間であるかどうかを推定できる．

2.4.2 Support Vector Machines について

本節で提案する重畳区間検出法は，SVM を用いて固有値分布をクラスタリングする事で実現される．ここでは，提案手法で必要となる SVM について詳述する．

SVM は，2 つのクラスのどちらに属するかラベル付けされた学習データを用いて，学習データをクラスタリングする識別関数を設計する手法として知られている．一度識別関数を設計すれば，この関数を用いて，属するクラスが未知のデータを学習データと同様にクラスタリングする事ができる．

図 2.15 は SVM による識別関数の設計の概念図である．この図では各学習デー

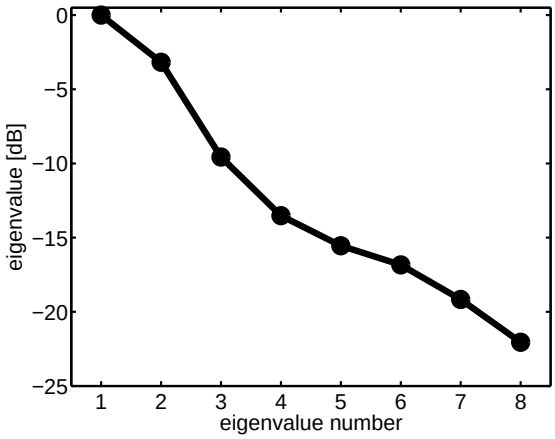

図 2.7: 背景雑音が白色でない場合の固有値分布

タは 2 次元のベクトルとなっている．クラスは 2 種類であり，それぞれのクラスに属するデータは + と・でプロットされている．SVM はこのように学習データがあたえられると，図の中央に optimal hyperplane と示されている識別超平面（もしくは識別超曲面）を与える．この超平面（曲面）を識別の基準とする事で，属するクラスが未知であるデータが与えられても，クラスタリングが可能となる．

今，2 つのクラス $+1$ と -1 にクラス分けされた学習データ $\mathbf{x}_i$ $(i = 1, \ldots, L)$ があるとする．L は学習データの数である．図 2.15 はこの学習データをプロットしたものである．クラス $+1$ のデータは + で，-1 のデータは $\cdot$ でプロットされている．また，データは線形分離可能であるとする．

線形分離可能であるので，適当な係数 $\mathbf{w}$ とバイアス b をとる事で，

$$\begin{cases} \mathbf{w}^T\mathbf{x}_i + b \geq 0 \text{ for } d_i = +1 \\ \mathbf{w}^T\mathbf{x}_i + b < 0 \text{ for } d_i = -1 \end{cases} \tag{2.10}$$

とする事が可能である．ただし，d_i は $\mathbf{x}_i$ が属するクラスで，ここでは $d_i = +1$ か $d_i = -1$ である．さらに，式（2.10）を $\min_i\{d_i(\mathbf{w}^T\mathbf{x}_i + b)\}$ で割る事で

$$\begin{cases} \mathbf{w}^T\mathbf{x}_i + b \geq +1 \text{ for } d_i = +1 \\ \mathbf{w}^T\mathbf{x}_i + b < -1 \text{ for } d_i = -1 \end{cases} \tag{2.11}$$

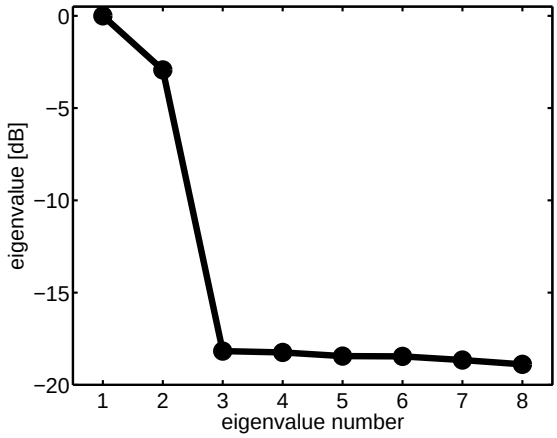

図 2.8: 図 2.7 を白色化した固有値分布

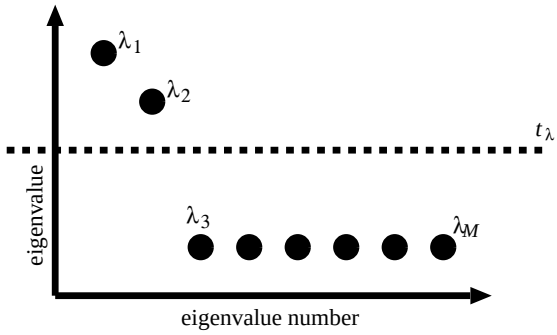

図 2.9: 閾値を用いる音源数推定法の例

とする事ができる．式（2.11）を 1 つの式にまとめれば,

$$d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \tag{2.12}$$

となる．ここで，特に $\mathbf{w}^T\mathbf{x}_i + b = +1$ や $\mathbf{w}^T\mathbf{x}_i + b = -1$ となる $\mathbf{x}_i$ をサポートベクターと呼ぶ．図 2.15 では，矢印で support vectors と指された点がそれである．

線形分離可能なデータであれば，式（2.11）を満たす $\mathbf{w}$ と b は無数に存在する．そこで，サポートベクターから分離平面への距離（図 2.15 では r になる），すなわちもう一方のクラスまでのマージンを最大化するという規範の下で，分離平面を求める事を考える．式（2.11）が成り立つとき，r の値は $r = \dfrac{1}{||\mathbf{w}||}$ となるので,

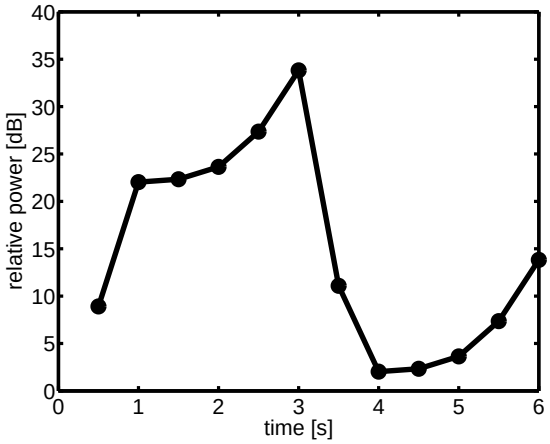

図 2.10: ある帯域における音源のパワー差の変化例

これを最大化すればよい事になる．すなわち，以下の二次計画問題を解く事で，最適な分離平面が得られる．

$$\text{最小化 } \frac{1}{2}||\mathbf{w}||^2 \tag{2.13}$$

$$\text{条件 } d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1,\ i = 1, \ldots, L \tag{2.14}$$

この問題の最適解 $\mathbf{w}^*$，b^* を用いる事で，学習データが超平面からマージン $r \geq 1$ の領域に入るよう，超平面 $\mathbf{w}^{*T}\mathbf{x} + b^* = 0$ が決定される．これを式で表せば，$d_i(\mathbf{w}^{*T}\mathbf{x} + b^*) \geq 1$ となる．

こうして得られた超平面を識別の基準とする事で，属するクラスが未知のデータを識別する事ができる．具体的には，属するクラスが未知のデータ $\mathbf{x}'$ に対して以下の基準で識別可能である．

- $\mathbf{w}^{*T}\mathbf{x}' + b^* \geq 0$ であれば属するクラス $+1$
- $\mathbf{w}^{*T}\mathbf{x}' + b^* < 0$ であれば属するはクラス -1

図 2.16 は，線形分離可能なデータの例である．図で，丸印で示されたデータがクラス -1，星印で示されたデータがクラス $+1$ のデータになる．これに対して，式

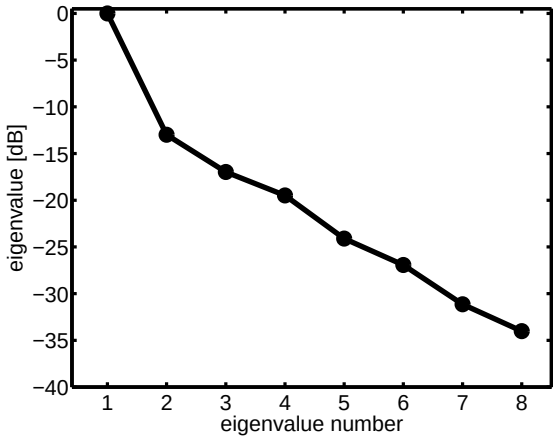

図 2.11: 図 2.10 の 3 秒の時の固有値分布

(2.13)，式（2.14）を用いて識別平面を求めた結果が図 2.17 である．図 2.17 において，実線が識別平面，破線が support vector を通過する線である．つまり，実線と破線で挟まれた領域が，SVM の学習によって最大化されたマージンになる．なお，この例で support vector は，破線上に乗っているデータで，クラス -1（丸印）では 1 つ，クラス $+2$（星印）では 2 つ存在する．

2.4.3 識別の難しいデータを含む場合の学習

2.4.2 節では，学習データが完全に 2 つのクラスに分離可能な場合を述べた．しかし，学習データが常に線形分離可能であるとは限らない．そこで，式（2.14）を $d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$ と拡張する（$\xi_i \geq 0$ である）．これは，図 2.18 のように，マージン内のデータを識別エラーとして許容する事に相当する（図 2.18 では星型で示しているデータが識別エラーのデータ）．

ただし，識別エラーやエラーの度合い（ξ_i の大きさ）はなるべく抑えたいので，目的関数（式（2.13））にペナルティパラメータ C を導入する．

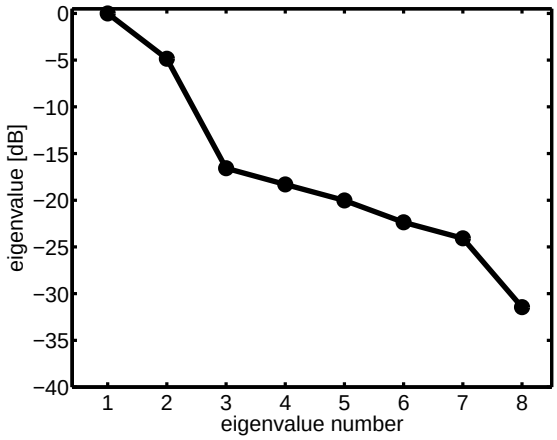

図 2.12: 図 2.10 の 4 秒の時の固有値分布

以上をまとめると，先の二次計画問題は以下のように変形される．

$$\text{最小化 } \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{L}\xi_i \tag{2.15}$$

$$\text{条件 } d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i = 1, \ldots, L \tag{2.16}$$

ここで，C は適当な正の値である．C の値が大きくなると，ξ_i が大きくなる事に対するペナルティーが重くなり，学習データに対してより識別誤りを起こしにくい分離平面が得られる．

図 2.19 は線形分離不可能なデータの例である．具体的には，図 2.16 に対してクラス $+1$（星形）のデータが $(1, 0.5)$ と $(0.5, 1.3)$ に追加されたものである．これに対して，式（2.15），式（2.16）を用いて識別平面を求めた結果が，図 2.19 から図 2.22 である．この 3 つの図はそれぞれ C の値が異なる．図 2.19 が $C = 0.1$，図 2.21 が $C = 1$，図 2.22 が $C = 1000$ で識別平面を求めた結果である．

$C = 0.1$ の場合は，識別マージン内にデータが入る事に対するコストが低いため，マージン最大化の影響が強く出ている．$C = 1$ になると，識別マージン内にデータが入るコストが高くなるため，先の例よりもマージン内に入るデータの数が減っている．$C = 1000$ では，このコストが極めて高いため，マージン最大化よりも，マージンの外にデータが出るように識別平面が設計される．この結果，学

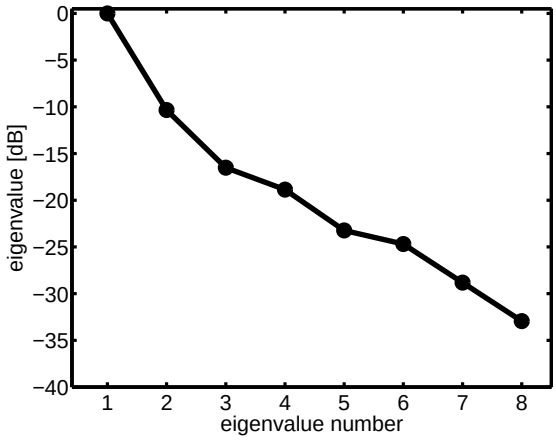

図 2.13: 図 2.10 の 6 秒の時の固有値分布

習データを極めて高い精度で分離する識別平面が得られる．$C = 1000$ の場合，学習データに対して識別誤りを起こしているのは，クラス -1（丸印）のデータのうちの 1 つだけである．

実際の最適化の計算では，式（2.15），式（2.16）を直接用いるのではなく，次式で表される，式（2.15），式（2.16）の双対問題を考える．

$$\text{最大化} \ -\frac{1}{2}\sum_{i,j=1}^{L}\alpha_i\alpha_j d_i d_j(\mathbf{x}_i\cdot\mathbf{x}_j)+\sum_{i=1}^{L}\alpha_i \tag{2.17}$$

$$\text{条件} \ \sum_{i=1}^{L}\alpha_i d_i = 0 \tag{2.18}$$

$$0 \le \alpha_i \le C,\ i = 1,\ldots,L \tag{2.19}$$

図 2.22 の例でも分かるように，線形分離が不可能な学習データに対しては，C をどんなに大きくしても完璧に分離する事は不可能である．そこで，学習データを特徴空間に写像し，この空間で線形分離を行う事を考える．特徴空間に写像する事で，元の空間では線形分離が不可能な場合でも，高い精度でクラスタリングが可能な識別関数が得られる事が期待される．

今，適当な関数 $f_{feature}(\cdot)$ を用いて，学習データを $f_{feature}(\mathbf{x}_i)$ のように特徴空間

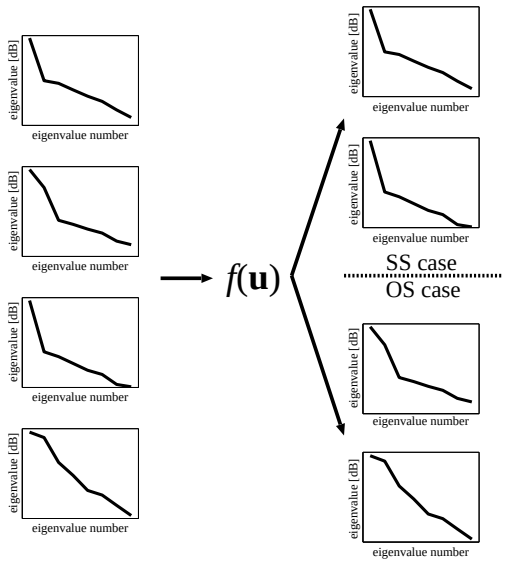

図 2.14: SVM による重畳区間検出法のイメージ

へ写像する．式（2.17），式（2.19）は,

$$\text{最大化}\ -\frac{1}{2}\sum_{i,j=1}^{L}\alpha_i\alpha_j d_i d_j (f_{feature}(\mathbf{x}_i)\cdot f_{feature}(\mathbf{x}_j)) + \sum_{i=1}^{L}\alpha_i \tag{2.20}$$

$$\text{条件}\ \sum_{i=1}^{L}\alpha_i d_i = 0 \tag{2.21}$$

$$0 \le \alpha_i \le C,\ i = 1,\ldots,L \tag{2.22}$$

となる．実際には，特徴空間での内積 $f_{feature}(\mathbf{x}_i)\cdot f_{feature}(\mathbf{x}_j)$ をカーネル関数 $K(\mathbf{x}_i,\mathbf{x}_j)$ で置き換えた式が使われる．カーネル関数を用いる事により，元の空間のデータから直接内積を計算する事が可能となる．

一般に，カーネル関数は Mercer の条件を満たしている必要がある．このようなカーネル関数の代表的なものとして以下のものがある [26].

$$d\text{ 次の多項式カーネル: } K(\mathbf{x},\mathbf{y}) = (1+\mathbf{x}\cdot\mathbf{y})^d \tag{2.23}$$

$$\text{RBF (Radial Basis Function) カーネル: } K(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2}\right) \tag{2.24}$$

$$\text{シグモイドカーネル: } K(\mathbf{x},\mathbf{y}) = \tanh(k\mathbf{x}\cdot\mathbf{y}-\theta) \tag{2.25}$$

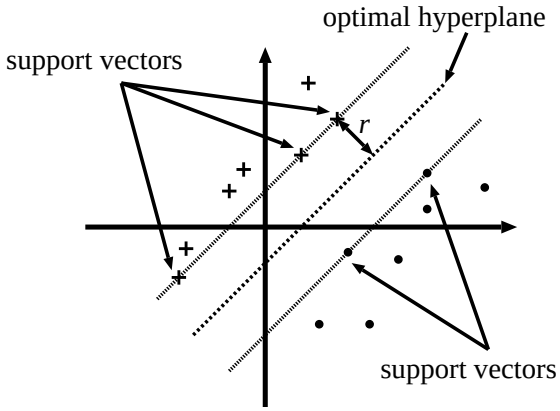

図 2.15: 線形分離可能な例

ここで，d，σ，k，θ はそれぞれ各カーネル関数のパラメータである．

以上の事より，最終的に次式のように定式化される [27, 28].

$$\text{最大化} \ -\frac{1}{2}\sum_{i,j=1}^{L}\alpha_i\alpha_j d_i d_j K(\mathbf{x}_i,\mathbf{x}_j)+\sum_{i=1}^{L}\alpha_i \tag{2.26}$$

$$\text{条件} \ \sum_{i=1}^{L}\alpha_i d_i = 0 \tag{2.27}$$

$$0 \le \alpha_i \le C,\ i = 1,\ldots,L \tag{2.28}$$

この二次計画問題の最適解 α_i^* を用いる事で識別関数 $f(\cdot)$ は

$$f(\mathbf{x}) = \sum_{i=1}^{L}\alpha_i^* d_i K(\mathbf{x},\mathbf{x}_i) + b^* \tag{2.29}$$

と表せる．ただし，b^* は $0 < \alpha_i^* < C$ なる適当な i を用いて

$$b^* = d_i - \sum_{j=1}^{L}\alpha_j^* d_j K(\mathbf{x}_i,\mathbf{x}_j) \tag{2.30}$$

となる．

図 2.23 から図 2.25 は図 2.19 に対して，カーネル関数を用いて識別曲面を求めた場合の結果である．黒い線が識別曲面である．また，赤い線と黒い線で挟まれた領域が識別マージンとなる．

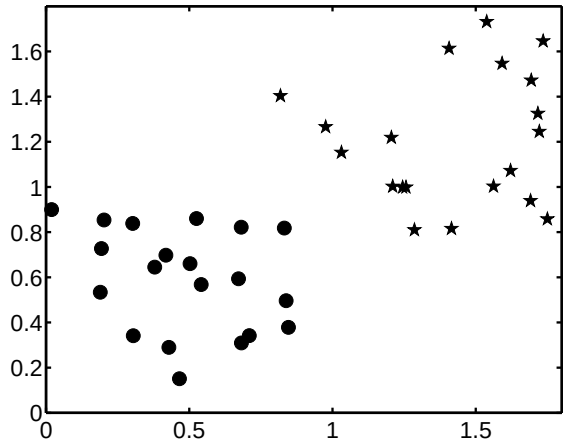

図 2.16: 線形分離可能な実例

カーネル関数には RBF（$\sigma = 2$）を用いており，図 2.23 が $C = 10$，図 2.24 が $C = 100$，図 2.25 が $C = 1000$ の場合の結果である．基本的には，図 2.19 から図 2.22 と同じ傾向が見られる．すなわち，C の値が小さいときは，より多くのデータが識別マージン内に入っているのに対して，C が大きくなるとこの数が減ってゆくという事である．特に今回の例は，カーネル関数を用いているため，識別の境界が曲面となっている．そして，$C = 1000$ の場合は学習データを 2 つのクラスに正しく識別できる識別曲面が得られている．これは，図 2.19 から図 2.22 の例では不可能である．これが，カーネル関数を用いるメリットである．

2.4.4 Support Vector Machines を用いた重畳区間検出法の提案

2.2 節において固有値分布が重畳区間に関する情報を反映しているという事を述べた．この性質より，各固有値分布が重畳区間，非重畳区間のいずれに属するかを決定できると考えられる．また，2.4.2 節と 2.4.3 節において，SVM を用いて学習データを基に識別関数を設計できる事を述べた．そこでここでは，空間相関行列の固有値分布を重畳区間に関する特徴量とみなして，SVM で識別する事を考える．固有値分布として実環境で得られたデータを用いる事で，十分な性能で重畳

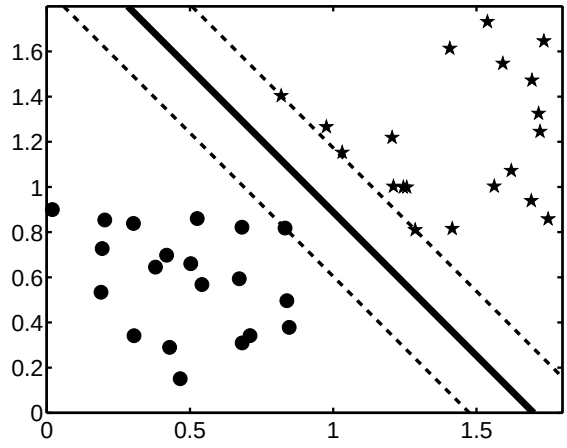

図 2.17: 図 2.16 に対して求めた識別平面

区間の推定が行える事が期待される．また，SVM を用いる事で，自動的に学習を行う事が可能となる．

今，重畳区間であるかが既知である固有値分布 $\boldsymbol{\lambda}_i$ $(i = 1, \ldots, L)$ が存在したとする．ここで，$\boldsymbol{\lambda}_i$ は M 次元のベクトルであり，式（2.26）における，$\mathbf{x}_i$（もしくは $\mathbf{x}_j$）に相当する．さらに，$\boldsymbol{\lambda}_i$ が非重畳区間の場合は属するクラスを -1，重畳区間の場合は属するクラスを $+1$ とする．このようにして，SVM において識別関数を設計するために必要となる学習データが用意できる．ここで，各固有値分布が重畳区間に属するかをどのように決定するかが問題となるが，この基準については 2.4.5 節において述べる．

この学習データと SVM を用いる事で，学習データをクラスタリング可能な識別関数 $f(\boldsymbol{\lambda})$ が得られる．この関数を用いれば，重畳区間であるかが未知である固有値分布 $\boldsymbol{\lambda}'$ に対しても $f(\boldsymbol{\lambda}')$ を計算する事で重畳区間の推定が可能となる．すなわち，

- $f(\boldsymbol{\lambda}') < 0$ であれば，$\boldsymbol{\lambda}'$ は非重畳区間の固有値分布．
- $f(\boldsymbol{\lambda}') > 0$ であれば，$\boldsymbol{\lambda}'$ は重畳区間の固有値分布．

である．

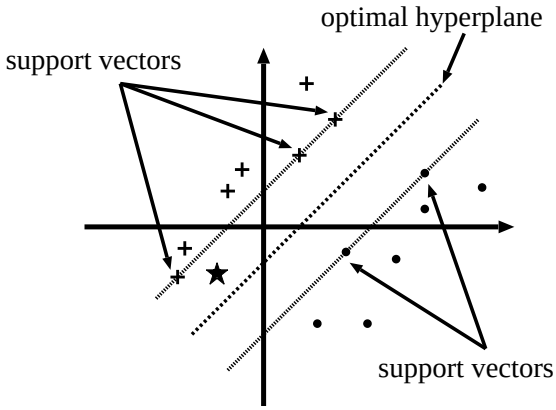

図 2.18: 線形分離不可能な例

続いて，SVM を用いて識別関数を設計する際に必要なパラメータについて検討する．2.4.2 節や 2.4.3 節において，C の値を変化させる事により，より正確に学習データを識別できる関数を設計できる事を述べた．図 2.26 は実環境下で得られた固有値分布を学習データとして設計した識別関数を用いた，固有値分布のクラスタリングの識別率を示したものである．ただし，学習と識別実験に用いた固有値分布は音源が異なるものである．図の横軸は学習時のパラメータである C で，この値を変化させた時の性能を示している．図において，C の値が増加すると性能が下がっている事が分かる．C の値が増加すると，学習データに対する識別性能は増加するが，それに対し汎化能力が下がるため，この影響が出ていると考えられる．また，SVM におけるパラメータとしては C 以外にもカーネル関数のパラメータ（RBF の場合は σ）がある．これらのパラメータは，学習データとは異なるテストデータに対する性能が最適になるように決定する事ができる．

2.4.5　広帯域信号への拡張

本節では，狭帯域において推定された重畳区間情報から，どのように広帯域信号における重畳区間情報を推定するかを述べる．まず，重畳区間をどのように定義するかについて述べる．重畳区間は基本的に 2 つ以上の音源が同時に音を出し

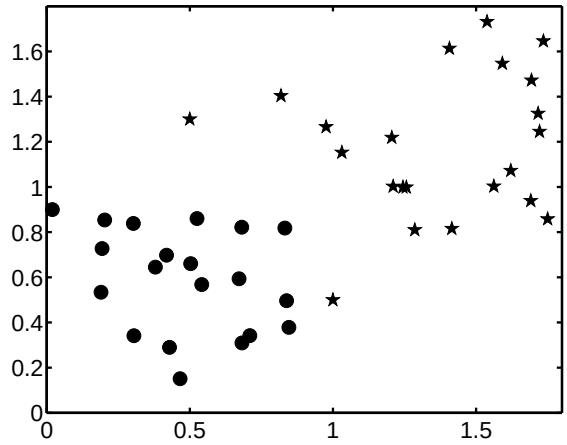

図 2.19: 線形分離不可能な実例

ている区間である．しかし，例えば 2 つの音源から同時に音を出している場合であっても，音源間のパワー差は常に変化している．図 2.27 は 2 つの音から音が発音している時のパワー差を示した図である．図の上段と中段はそれぞれ，2 つの音源の波形である．図の下段はこの 2 つの音源について，ある周波数でのパワー差を示したものである．音源 1 は全区間において同程度のパワーで発音しており，音源 2 は，後半の 3 秒間は音源 1 と同程度で発音しており，前半の 3 秒間はそれに比べ弱いパワーで発音している．ここで，2 つの音源から同時に発音している意味では図の全区間が重畳区間となるべきだが，図の 3 秒までの区間は音源のパワー差が極めて大きいため，事実上 1 つの音源が単独で発音している，非重畳区間と見なせる．

そこで，本論文では次のような基準で狭帯域における重畳区間を定義する．今，周波数領域における音源 1 と音源 2 のパワーをそれぞれ $s1_{power}(\omega, T)$，$s2_{power}(\omega, T)$ とする．T は短区間フーリエ変換におけるフレーム番号である．このとき，適当な閾値 $P_{threshold}$ を設定し，以下の基準で各周波数における重畳区間を定義する．

- $|s1_{power}(\omega, T) - s2_{power}(\omega, T)| > P_{threshold}$ ならば非重畳区間.
- $|s1_{power}(\omega, T) - s2_{power}(\omega, T)| \leq P_{threshold}$ ならば重畳区間.

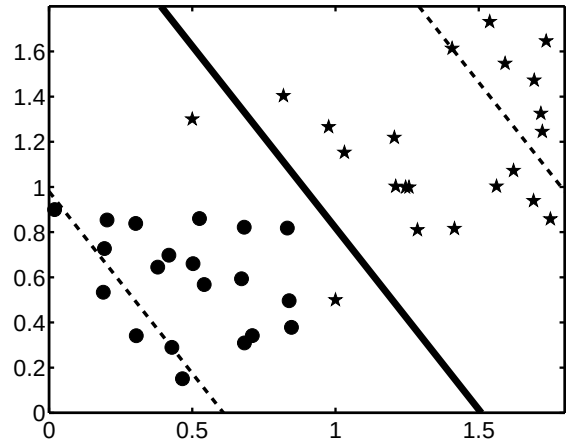

図 2.20: 図 2.19 に対して求めた識別平面（$C = 0.1$ の場合）

これは，音源のパワー差が $P_{threshold}$ を基準として，それより大きければ事実上 1 つの音源が発音している考え，非重畳区間と見なしている．

この方法により，周波数ごとに重畳区間であるかを決定できる．しかし，実際には時間領域において該当する区間が重畳区間かどうかを決定する必要がある．図 2.28 は，同程度のパワーで同時に発音している 2 つの音源の帯域別のパワーを示したものである．この図において，矢印で示した 2 つの帯域では音源のパワー差が大きく異なり，それぞれ重畳区間，非重畳区間に属する帯域である．このように，時間領域において重畳区間であるデータであっても，帯域ごとには重畳区間，非重畳区間が混在する．

そこで，本論文では周波数ごとに得られた重畳区間情報から図 2.29 のようなヒストグラムを生成する．そして，より多くの周波数が分類された方を時間領域における結果とする．例えば，この図では重畳区間に分類された帯域の数がそれ以外の帯域の数より多い事を示している．この時，時間領域における該当する区間は重畳区間という事になる．

ここで，この節で述べた周波数領域において付与した重畳区間情報から時間領域での重畳区間を定義する方法が妥当であるかが重要となる．図 2.30 は時間領域

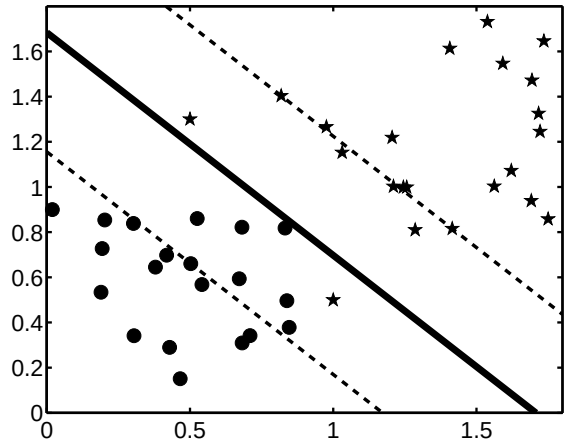

図 2.21: 図 2.19 に対して求めた識別平面（$C = 1$ の場合）

における重畳区間と非重畳区間において，それぞれの区間に含まれる帯域ごとの音源のパワー差（$|s1_{power}(\omega, T) - s2_{power}(\omega, T)|$）の分布を示した図である．図において重畳区間における音源のパワー差は非重畳区間における音源のパワー差よりも全体的に小さい．特にこの例では，$P_{threshold}$ を 18 dB 辺りに設定すれば，時間領域における重畳区間を正確に検出できる．このように，周波数領域において付与した重畳区間情報を用いて時間領域における重畳区間を検出する事が可能である．

狭帯域において推定された重畳区間情報から，広帯域の重畳区間を推定する際にも，上述の考え方を用いる．すなわち，周波数ごとに得られた固有値分布の重畳区間情報から図 2.29 のようなヒストグラムを生成し，より多くの周波数が分類された方を時間領域における結果とする．図 2.31 は図 2.27 の 2 つの音源が入力された時の広帯域の信号に対する重畳区間検出の例である．図の上段は重畳区間区間の正解を示しており，下段は狭帯域ごとに推定された重畳区間のヒストグラムを示している．図の下段において，最初の 3 秒間は始めのフレームを除いて，非重畳区間と推定された固有値分布の数が重畳区間と推定された固有値分布の数を上回っている．このため，提案手法は始めのフレームを除いて，正しく非重畳区

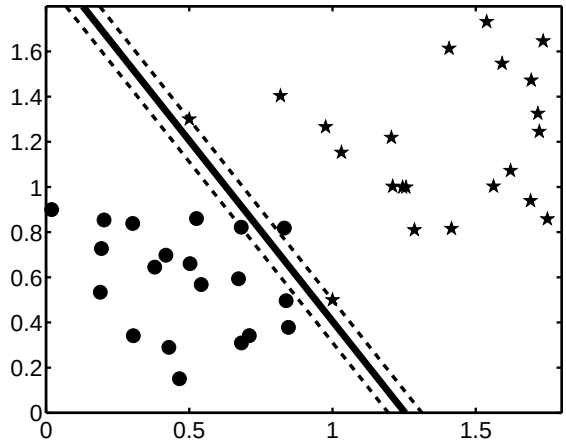

図 2.22: 図 2.19 に対して求めた識別平面（$C = 1000$ の場合）

間を検出できている．また，3 秒以降については，重畳区間と推定された固有値分布の方が多く，最終的に重畳区間を正しく検出できている．

2.5 Support Vector Regression を用いた重畳区間検出法

2.5.1 Support Vector Machines を用いた重畳区間検出法の問題点

前節では，SVM を用いた重畳区間検出法を提案した．しかしながら，この手法では重畳区間の検出時に問題が生じる可能性がある．ここでは，その問題について述べる．

図 2.32 は 2 つの音源において，音源間のパワー差が 0 dB，10 dB，20 dB と変化した時の固有値分布を示したものである．この図を見ると，明らかに重畳区間である 0 dB の固有値分布が他の 2 つとは異なり，図 2.2 と同等の分布を示している．また，他の 2 つの固有値分布についてもある程度の違いが見受けられる．このように，固有値分布は音源のパワー差の情報を反映している．SVM による手法ではこれらの分布をパワー差に対する閾値で 2 つのクラスに分類し，学習データ

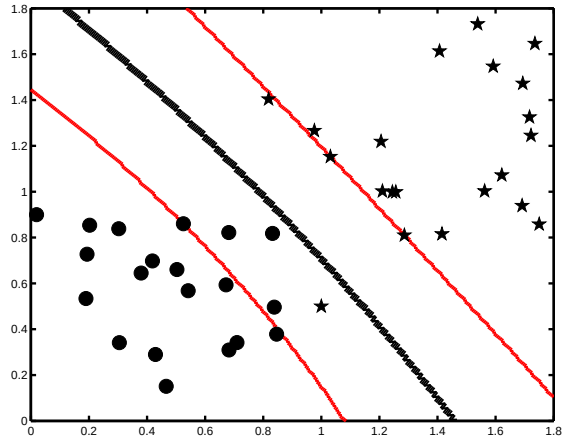

図 2.23: 図 2.19 に対して求めた識別曲面（RBF $\sigma = 2$，$C = 10$ の場合）

として用いていた．しかし，実際にはパワー差は連続値であり，これに対応して固有値分布にも様々な形状が存在する．そのため，これを 2 つのクラスで完全に表現する事は困難である．また，特に閾値付近での固有値分布に対し SVM がクラスタリングを誤る可能性がある．

この問題に対処するため，本節では，固有値分布を 2 つのクラスに分類するのではなく，固有値分布が生成された時の音源間のパワー差を連続値としてそのまま推定する事を考える．パワー差が推定できれば，重畳区間の定義に基づき，適当な閾値を用いて各固有値分布が重畳区間の分布であるかを決定できるため，より正確に重畳区間を検出できる事が期待される．また，SVM による手法ではパワー差の閾値を設定後に識別関数を設計する．このため，閾値を変更するためには改めて学習データを用意し SVM で学習を行う必要がある．しかし，パワー差を推定する事が可能となれば，パワー差の推定後に重畳区間であるかを決定する事ができ，閾値を変更してもパワー差の推定部分には影響が及ばない．そのため，より簡便に閾値を変更する事が可能である．本節では，音源間のパワー差を推定する手法として，SVR を用いる事を提案する．SVR は，学習データから回帰曲線を設計する手法であり，SVR を用いる事で固有値分布からパワー差を推定する事が可

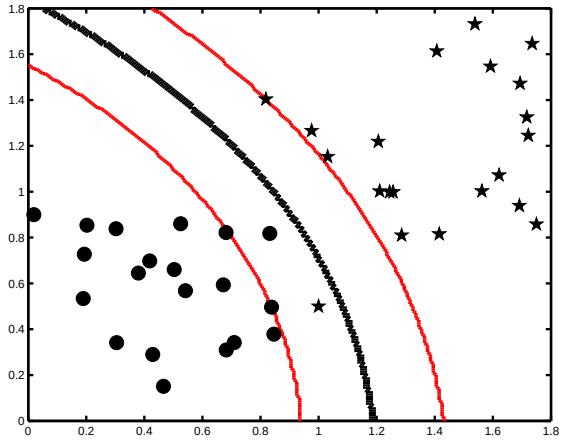

図 2.24: 図 2.19 に対して求めた識別曲面（RBF $\sigma = 2$，$C = 100$ の場合）

能になると考えられる．

2.5.2 Support Vector Regression について

ここでは，SVR を導入するための準備として，一般的な SVR について述べる．SVR は，真値としてスカラー値が与えられた学習データを用いて，学習データを回帰する直線（平面）を設計する手法として知られている．一度回帰関数を設計すれば，この関数を用いて学習データと同じ属性を持つ，真値が未知のデータを推定する事ができる．

今，学習データ $\{\mathbf{x}_i\}$ と $\{\mathbf{x}_i\}$ に対応するスカラー値 $\{d_i\}$ があるとする．ここで，SVR で求められる回帰直線 $f(\cdot)$ は以下の式で表される．

$$f(\mathbf{u}) = \sum_{i=1}^{L} (\alpha_i^* - \alpha_i) G(\mathbf{u}, \mathbf{x}_i) + b, \tag{2.31}$$

ここで，$G(\cdot, \cdot)$ はカーネル関数である．α_i^* と α_i は以下の数理計画問題の最適解で

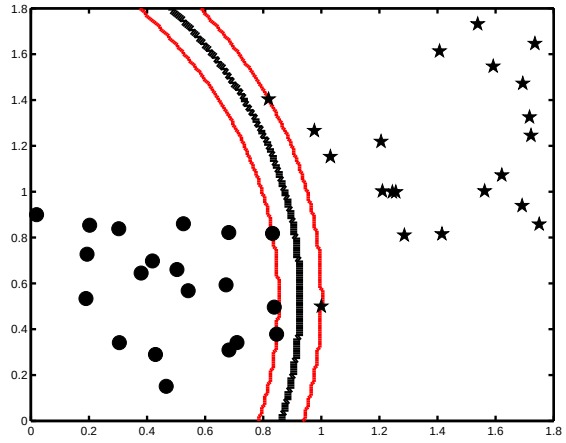

図 2.25: 図 2.19 に対して求めた識別曲面（RBF $\sigma = 2$，$C = 1000$ の場合）

ある．また，$C > 0$ と $\epsilon \geq 0$ は定数である．

$$\begin{aligned} \text{maximize} \quad & -\epsilon \sum_{i=1}^{L} (\alpha_i^* - \alpha) + \sum_{i=1}^{L} (\alpha_i^* - \alpha_i) d_i \\ & - \frac{1}{2} \sum_{i,j=1}^{L} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(\mathbf{x}_i, \mathbf{x}_j) \qquad (2.32) \\ \text{subject to} \quad & \sum_{i=1}^{L} (\alpha_i^* - \alpha_i) = 0, \qquad (2.33) \\ & 0 \leq \alpha_i^*, \alpha_i \leq C \ (i = 1, \ldots, L), \qquad (2.34) \end{aligned}$$

b は以下の式で求められる．

$$b = d_k - \sum_{j=1}^{L} (\alpha_j^* - \alpha_j) G(\mathbf{x}_k, \mathbf{x}_j), \qquad (2.35)$$

ここで，k は $0 < \alpha_k^*, \alpha_k < C$ を満たす任意の値である．

また，ϵ は回帰時に誤差として許容する範囲を指定するパラメータである．図 2.33 は黒丸で示された一次元の学習データに対する SVR の例である．図の横軸は学習データ（$\mathbf{x}_i$）の値，縦軸は学習データに与えられた真値（d_i）である．また，図の実線は SVR で求めた回帰直線，破線は実線からそれぞれ $\pm\epsilon$ 動かした線であ

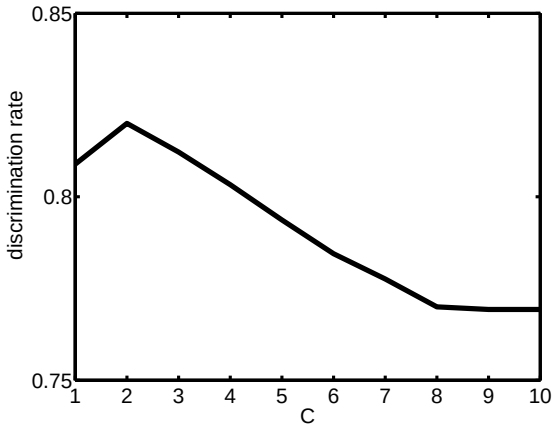

図 2.26: SVM における性能と C の関係

る．ここで，破線の範囲内にある学習データは誤差 0 で回帰されたとみなされている．このため，例えば，図の◯で示された点に関して，誤差は ξ のみとなる．

パラメータ C は，回帰直線との誤差に対するペナルティを与えるものである．すなわち，C の値が大きくなれば，なるべく誤差を小さくするように（多くの学習データが破線の範囲内に入るように）回帰直線が設計される．

図 2.34 は一次元の学習データの例である．図の横軸は学習データ（$\mathbf{x}_i$）の値，縦軸は学習データに与えられた真値（d_i）である．図 2.35 は図 2.34 に対して式（2.31）を用いて求めた回帰直線の例である．図において点で示されているのが図 2.34 の学習データであり，破線が SVR で求められた回帰直線である．図の縦軸は，学習データの真値，もしくは回帰直線で推定された学習データの値である．ここで，式（2.31）のカーネル関数には内積（$G(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$）を用いている．このように，SVR はカーネル関数に内積を用いる事で，学習データに対して回帰直線を求める事が可能である．また，ϵ は誤差として許容する範囲を与えるパラメータである．このため，この値を小さくする事により，学習データをより正確に回帰する直線を求める事ができる．図 2.36 は $\epsilon = 1$ とした場合の回帰直線である．先の $\epsilon = 4$ とした場合の図 2.35 に比べ，図 2.36 の結果は，より正確に回帰直線が求められている．

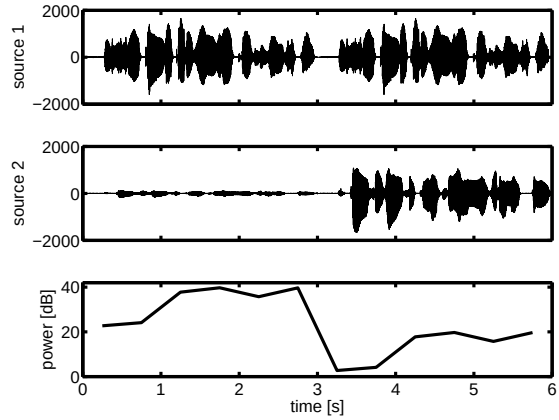

図 2.27: 2 つの音源のパワーの差

SVM の場合はカーネル関数に内積以外の関数を用いる事で，分離曲面を求める事が可能である．同様に，SVR の場合も，カーネル関数に用いる関数を変える事で回帰曲線を求める事ができる．図 2.37 はカーネル関数に RBF を用いて求めた回帰曲線である．また，ここで，パラメータ C は回帰曲線によって推定された値と学習データの値の誤差に対するペナルティである．この値を大きくする事でより学習データを正確に回帰する直線を求める事ができる．図 2.38 は $C = 10$ とした場合の回帰曲線である．先の図 2.37 に比べ，より正確に回帰曲線が求められている．図 2.39 は図 2.38 からさらに $\epsilon = 0.1$ に変更して得られた回帰曲線である．

このように，SVR においても，カーネル関数に内積以外の関数を用いる事で，より正確な回帰曲線を求める事ができる．

2.5.3 Support Vector Regression を用いた重畳区間検出法の提案

2.5.1 節においては，固有値分布が音源間のパワー差の情報を反映しているという事も述べた．仮に，音源間のパワー差が推定可能であれば，重畳区間の定義に従い，パワー差が小さい固有値分布は各音源が同程度の大きさで発音しているので重畳区間とみなし，そうでない場合は非重畳区間と判断する事ができる．また，

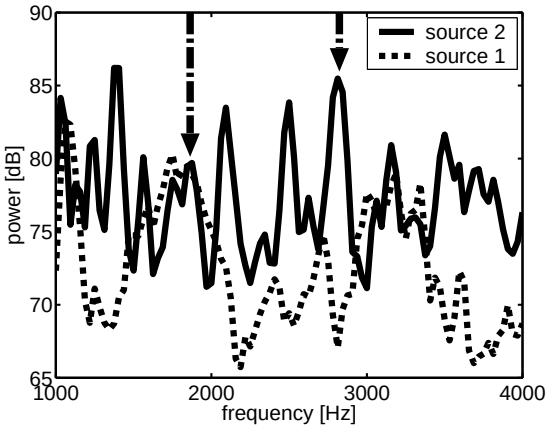

図 2.28: 各音源の帯域別のパワー

2.5.2 節では SVR を用いて，学習データを基に回帰曲線を求める事ができる事を述べた．そこで，ここでは，空間相関行列の固有値分布を音源間のパワー差の特徴量とみなして，SVR でパワー差を推定する事を考える．

今，音源間のパワー差が既知である固有値分布 $\boldsymbol{\lambda}_i$ $(i = 1, \ldots, L)$ が存在したとする．ここで，$\boldsymbol{\lambda}_i$ は M 次元のベクトルであり，式（2.26）における，$\mathbf{x}_i$（もしくは $\mathbf{x}_j$）に相当する．さらに，$\boldsymbol{\lambda}_i$ に対応する音源のパワー差を p_i $(i = 1, \ldots, L)$ とする．これは，式（2.32）における，d_i に相当する．このようにして，SVR において回帰曲線を求めるために必要となる学習データが用意できる．

この学習データと SVR を用いる事で学習データに対する回帰曲線 $f(\lambda)$ を求める事ができる．この回帰曲線 $f(\lambda)$ を用いる事で，パワー差が未知である固有値分布 λ' に対して $f(\lambda')$ を求める事でパワー差を推定できる．さらに，推定パワー差に対して閾値 P_{th} を設定する事で，

- $f(\lambda') \leq P_{th}$ の時，λ' は重畳区間の分布
- $f(\lambda') > P_{th}$ の時，λ' は非重畳区間の分布

と重畳区間を推定する事ができる．

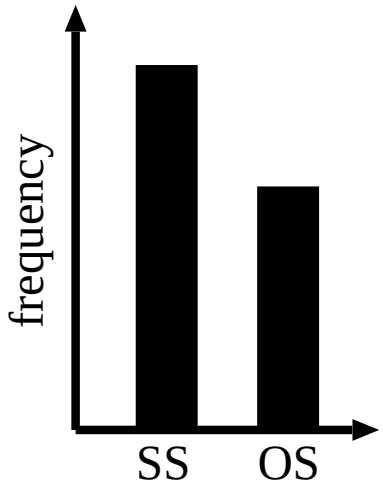

図 2.29: 帯域ごとに得られた重畳区間情報のヒストグラム

図 2.40 は SVR を用いたパワー差の推定例である．図は 100 個の学習データ（固有値分布）で学習した回帰曲線を用いて，学習に用いたのと同じデータに対してパワー差を推定した結果である．学習時のカーネル関数は RBF，$\sigma = 10$，$C = 10$ である．図において，横軸は学習データの通し番号，縦軸はパワー差を表している．図中の黒い点がパワー差の正解値であり，赤い破線が SVR による推定結果である．この図より，概ね 10 dB から 30 dB の範囲においてはパワー差が正しく推定できている事が分かる．また，図 2.41 は図 2.40 とは異なるデータに対してパワー差を推定した結果である．図において，横軸はパワー差の正解値，縦軸は推定値である．パワー差の推定が完全であれば，推定結果は赤い破線に一致する．この図より，パワー差はある程度推定されている事が分かる．

図 2.42 は実環境下で得られた固有値分布を学習データとして設計した回帰曲線を用いた，パワー差の推定性能を示したものである．ただし，学習と識別実験に用いた固有値分布は音源が異なる．図の横軸は学習時のパラメータである C で，この値を変化させた時の性能を二乗平均平方根誤差（RMSE; Root Mean Square

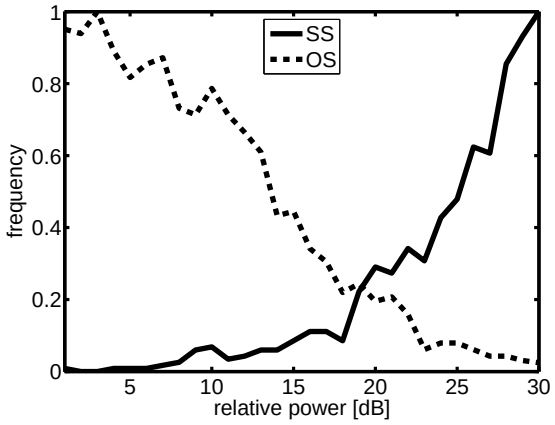

図 2.30: 重畳区間と非重畳区間におけるパワー差の分布

Error）で示している．RMSE は以下の式で定義される．

$$\sqrt{\frac{\sum_{i=1}^{N}(p_i - p_i')^2}{N}} \tag{2.36}$$

ここで，p_i はパワー差の正解値，p_i' はパワー差の推定値，N は推定に用いたデータの個数である．図において，C の値が増加すると RMSE の値が下がっている事が分かる．これは，推定精度が改善している事を示している．

2.5.4　広帯域信号への拡張

2.5.3 節において，固有値分布ごとに，その固有値分布のパワー差を推定し，重畳区間を検出する手法を提案した．しかし，これは SVM による手法と同様に，各固有値分布が重畳区間であるか否かを周波数ごとに推定している．そこで，これを時間領域の広帯域へ拡張する事を考える．これは，2.4.5 節と同様に，周波数ごとに得られた固有値分布の重畳区間情報から図 2.29 のようなヒストグラムを生成し，より多くの周波数が分類された方を時間領域における結果とする事で実現できる．

図 2.43 は図 2.27 の 2 つの音源に対して重畳区間を検出した例である．図 2.43 の上段はある帯域におけるパワー差の正解値，中段は上段と同じ帯域におけるパワー

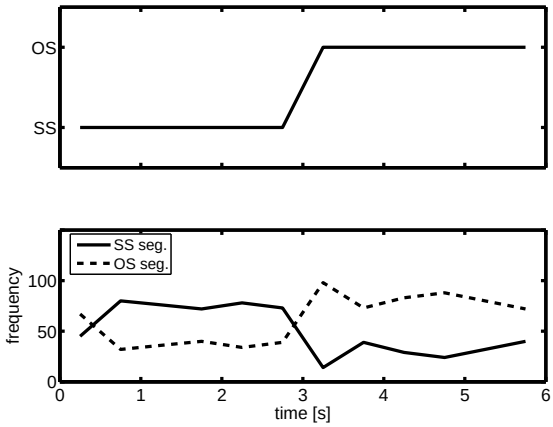

図 2.31: 広帯域信号におけ重畳区間検出の例

差の推定値，下段は全帯域での重畳区間と非重畳区間の頻度を示したものである．図より，パワー差の推定がある程度行えており，また，帯域ごとの結果を集計する事で，広帯域における重畳区間を正しく検出できている．図 2.44 は同一データに対して SVM と SVR で重畳区間を検出した例である．図の上段と中段はそれぞれ 2 つの音源の波形であり，下段は SVM 及び SVR を用いた重畳区間の検出結果である．下段の実線が SVM による結果であり，破線が SVR による結果である．SVM では 0 秒付近と 1.5 秒から 2 秒の間の非重畳区間を重畳区間と誤検出しているが，SVR ではこれを正しく検出している．また，前半の 3 秒間も重畳区間として検出したい場合，SVM では学習データからの再学習が必要になるが，SVR では，上述した P_{th} の値を変更するだけで可能である．

2.6 提案手法の評価

2.6.1 実験条件

本章では，複数人が出席した会議の様子をマイクロホンアレイで収録し，このデータを用いて実験を行った．実際に行われた会議は，市場調査の分野で用いられるグループインタビューというものである．グループインタビューは，調査対

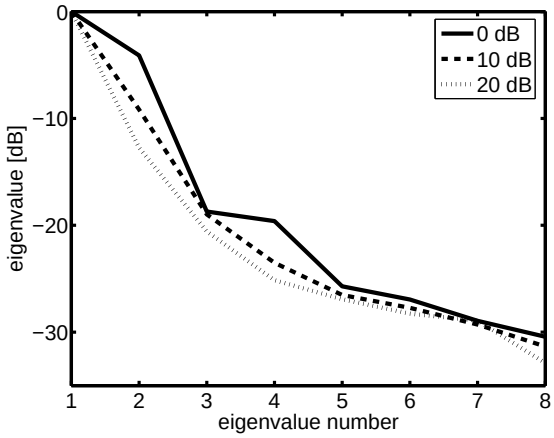

図 2.32: パワー差と固有値分布の関係

象となる参加者に対してプロの司会者が，調査項目に関して質問を行い，参加者がそれに対して答えるという形式で行われる．例えば，テーマが携帯電話であれば，司会者は「どんな種類の携帯電話を使っているのか」といった質問を行い，参加者がそれに対して答えていく．この収録における参加者は，司会者 1 名と 5 名の男性大学生である．この 6 名が会議室内にあるテーブルに着席し，会議の様子はそのテーブルの中央に置いたマイクロホンアレイで収録した．図 2.45 は収録に用いたマイクロホンアレイである．アレイは直径 0.2 m の円形であり，8 つのマイクロホンが取り付けられている．

固有値分布は帯域ごとに 1 つ得られる．今回用いる（主に想定している）音源は音声であり，音声のパワーが主に存在している帯域で，重畳区間の情報が固有値分布に反映していると考えられる．この事を踏まえ今回は 500 Hz から 4000 Hz に対応する固有値分布を用いている．この帯域は，ターゲットとしている音源のパワーがどの帯域に存在しているかを考慮して決定すればよい．また，今回は各手法の性能の上限値を調べるという観点から，学習時に用いるパラメータについては様々な値で実験を行い，その中で性能の高かったものを結果として採用している．その他の実験パラメータは表 2.1 に示している．

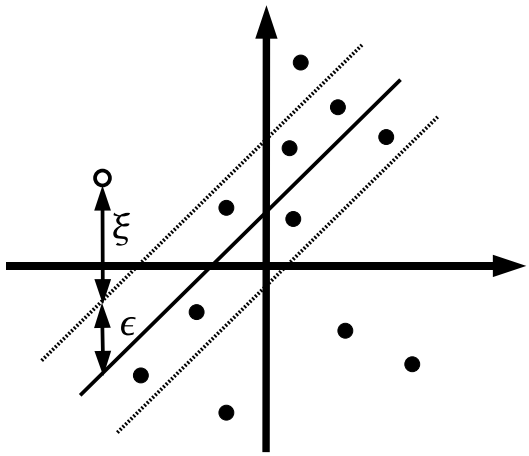

図 2.33: SVR の例

2.6.2 実験結果

表 2.2 は収録データにおける重畳区間検出実験の結果である．表において，R_r は再現率，R_p は適合率，R_f は誤検出率，F は F 値 [29] を表している．それぞれの値は以下の式で定義される．

$$R_r = \frac{\text{検出結果のうち正しく OS と検出できた数}}{\text{全ての OS の数}} \tag{2.37}$$

$$R_p = \frac{N_d \text{ のうち正しく OS が検出できた数}}{\text{OS と検出した数（}N_d\text{）}} \tag{2.38}$$

$$R_f = \frac{N_d \text{ のうち実際の正解が SS であった数}}{\text{全ての SS の数}} \tag{2.39}$$

$$F = \frac{2R_rR_p}{R_r + R_p} \tag{2.40}$$

また，表中の P_{th} は SVR を用いる手法において，推定パワー差に対して設定された閾値の値である．この値を小さくする事で，より多くの固有値分布が非重畳区間に分類され，逆に大きくする事で重畳区間と判別される固有値分布の数が多くなる．すなわち，P_{th} の値を変化させる事で重畳区間と非重畳区間の検出率をある程度制御する事ができる．

表において，SVR の手法による太字の結果と SVM による結果はそれぞれ F 値

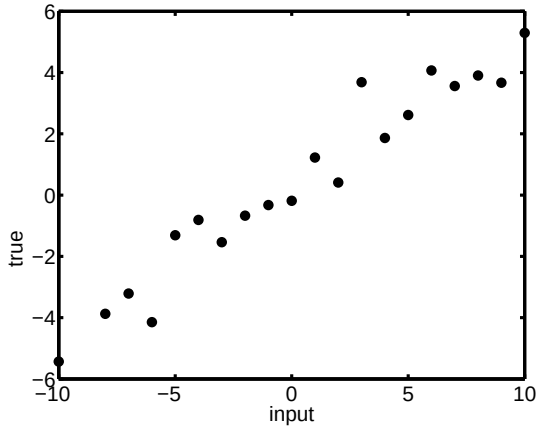

図 2.34: 回帰直線を求めるためのデータの例

が最大となったものである．この時，再現率が SVM において約 4 割，SVR において約 5 割となっている．誤検出率はほぼ同等である．また，SVR においては，P_{th} を変化させる事により，再現率と適合率の値を変える事ができる．例えば，再現率を SVM の結果と同じなるように P_{th} を変化させた場合，適合率は約 4 %上昇する．

さらに本実験では比較のために閾値法による重畳区間検出も行っている．閾値法における閾値は，F 値が最大になる値を用いており，本実験で採用された閾値は -7.0 である．閾値法の結果を見てみると，再現率が SVM や SVR による手法よりも高いがその一方で適合率は極めて低い．これは，より多くの区間を重畳区間と検出しているからである．

2.7　おわりに

本章ではまず，音源数や音源間のパワー差の情報を反映する空間相関行列の固有値分布の性質について述べた．また，固有値分布を用いた従来の手法について述べ，部屋の反射や残響が存在する実環境下では従来法が前提とする条件を満たす事ができず，音源数の推定に失敗する事を述べた．

続いて，SVM を用いた重畳区間検出法を提案した．SVM による手法では，SVM

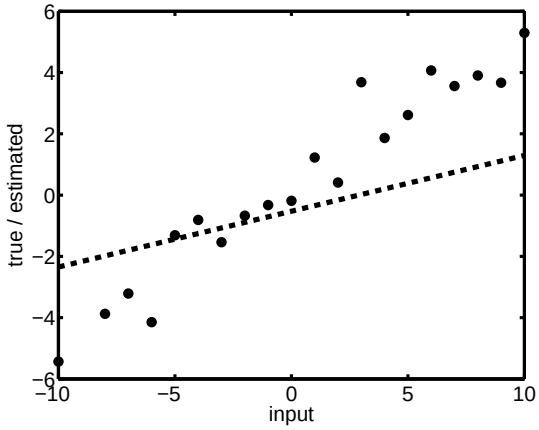

図 2.35: 図 2.34 に対して求めた回帰直線（$C = 1$，$\epsilon = 4$）

を用いて固有値分布を 2 つのクラスに分類し，分類されたクラスによって重畳区間であるかを判別する手法である．しかし，音源のパワー差は様々なケースが考えられ，固有値分布も多様な形状が存在する．SVM による手法はこれを 2 つのクラスで表現するため，クラスタリングを誤る可能性がある．

そこで，この問題を解決する手法としてさらに，SVR を用いた重畳区間検出法を提案した．SVR による手法では，固有値分布からパワー差を推定する手法である．最後に提案手法を実環境下で収録された会議音声に適用し，重畳区間の検出実験を行った．この結果，再現率において 4 割から 5 割程度の結果が得られた．また，SVR による手法では，重畳区間と非重畳区間の境界である閾値を再学習を行う事なく変更可能である．これにより，検出性能をある程度調整可能である事を示した．

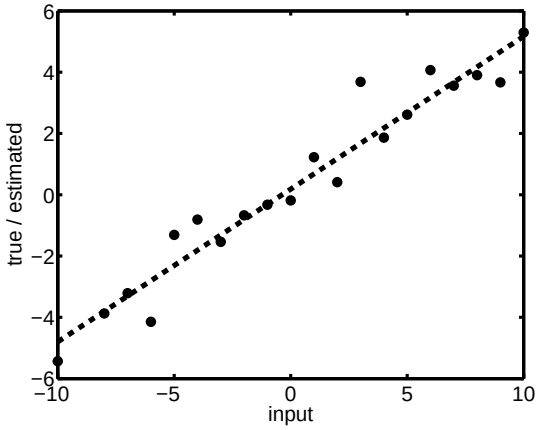

図 2.36: 図 2.34 に対して求めた回帰直線（$C = 1$，$\epsilon = 1$）

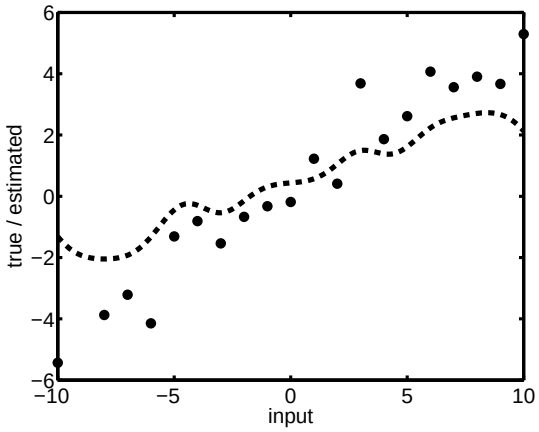

図 2.37: 図 2.34 に対して求めた回帰曲線（RBF $\sigma = 1$，$C = 1$，$\epsilon = 1$ の場合）

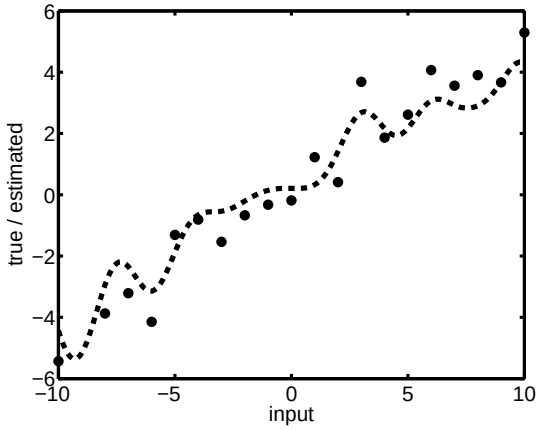

図 2.38: 図 2.34 に対して求めた回帰曲線（RBF $\sigma = 1$，$C = 10$，$\epsilon = 1$ の場合）

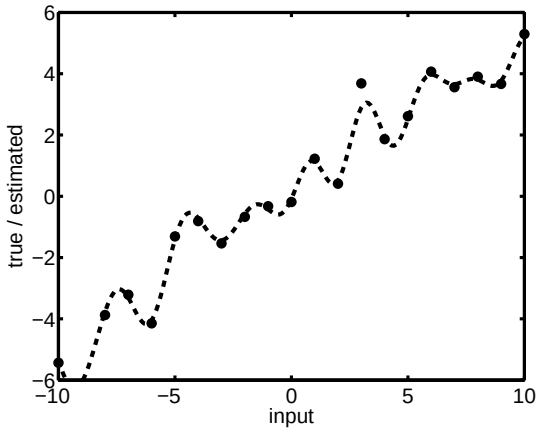

図 2.39: 図 2.34 に対して求めた回帰曲線（RBF $\sigma = 1$，$C = 10$，$\epsilon = 0.1$ の場合）

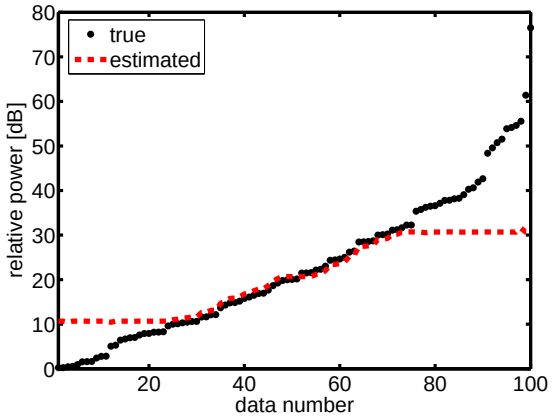

図 2.40: SVR を用いたパワー差の推定例

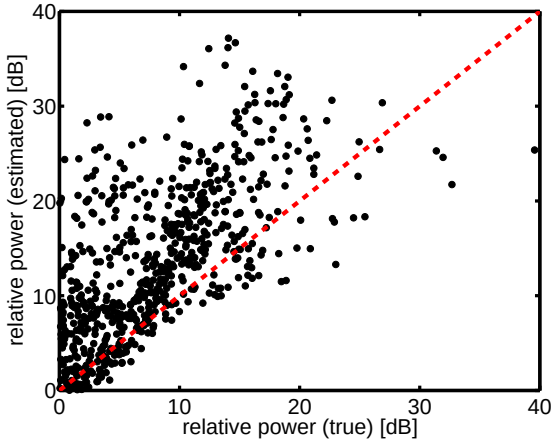

図 2.41: SVR を用いた学習データとは異なるデータに対するパワー差の推定例

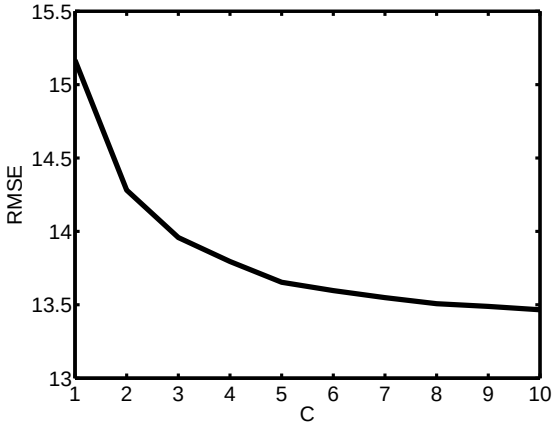

図 2.42: SVR における性能と C の関係

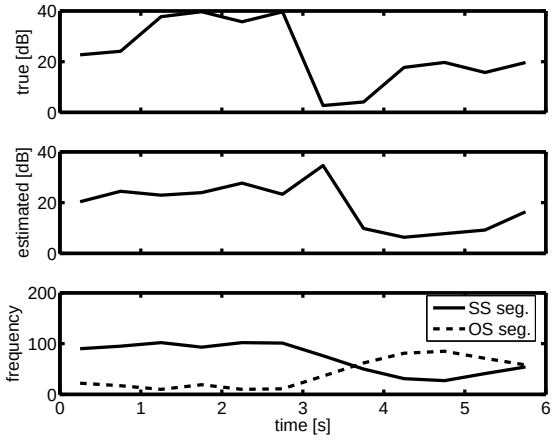

図 2.43: SVR による重畳区間検出例

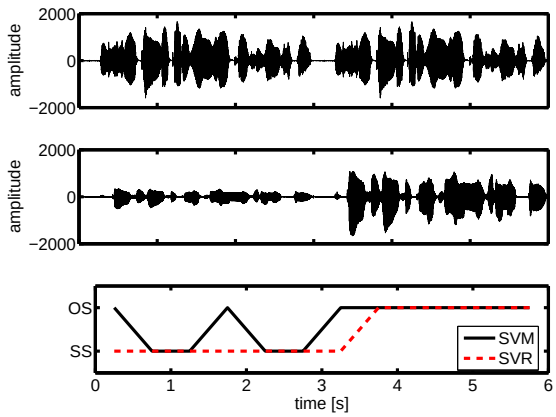

図 2.44: SVM と SVR による重畳区間検出例

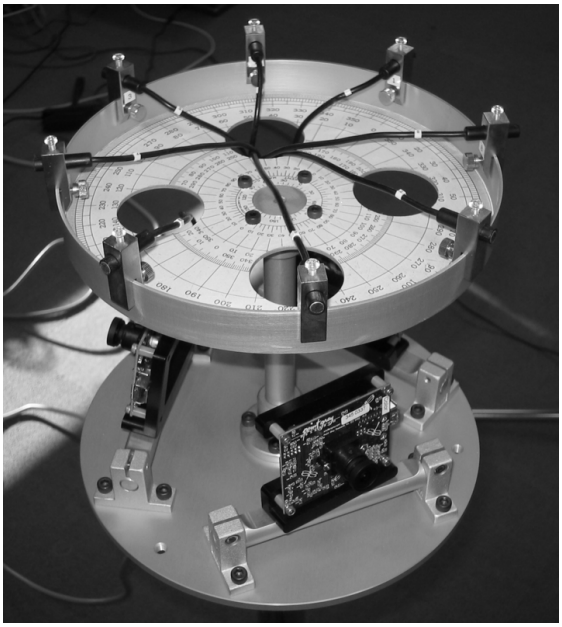

図 2.45: 収録に用いたマイクロホンアレイ

表 2.1: 実験時の各種パラメータ

サンプリング周波数	16000 Hz
FFT 点数	512
FFT シフト数	128
使用帯域	500 - 4000 Hz
フレーム長	0.5 秒
SVM と SVR のカーネル関数	RBF
SVM の学習データ数	300
SVM の σ	2
SVM の C	0.1
SVR の学習データ数	100
SVR の σ	10
SVR の C	8
SVR の ϵ	1

表 2.2: 実験結果

	R_r	R_p	R_f	F	閾値（SVR では P_{th}）[dB]
SVR	0.402	0.424	0.063	0.413	15.2
	0.423	0.423	0.066	0.423	15.3
	0.433	0.404	0.073	0.418	15.4
	0.454	0.400	0.078	0.425	15.5
	0.474	0.393	0.084	0.430	15.6
	0.495	**0.393**	**0.087**	**0.438**	**15.7**
	0.495	0.369	0.097	0.423	15.8
	0.526	0.347	0.113	0.418	15.9
SVM	0.402	0.361	0.081	0.381	-
閾値法	0.320	0.100	0.329	0.152	-5.0
	0.526	0.115	0.465	0.188	-6.0
	0.784	**0.134**	**0.579**	**0.229**	**-7.0**
	0.856	0.121	0.711	0.212	-8.0
	0.907	0.113	0.817	0.201	-9.0

第3章　音響情報と画像情報を統合した目的音源検出法

3.1　はじめに

本論文が想定する環境下で音声インタフェースを実現するためには，音源分離技術を用いて話者の声を雑音から分離する必要がある．第1章でも述べた通り，音源分離には様々な手法が存在する．例えば，BSS（Blind Source Separation）を用いれば，用いるマイクロホンアレイの形状や音源の位置に関する情報が未知であっても音源分離を行う事ができる．しかし，BSSでは学習に，ある程度の長さのデータを要する点や分離性能がABF（Adaptive BeamFormer）に比べ劣るという問題がある．一方，ABFでは高い分離性能が期待される反面，話者の音源位置などの情報を必要とする．この問題に対処するため，例えば[12]では話者が存在する位置をあらかじめ決めておき，その範囲内に存在する音源は話者とみなしてABFの分離フィルタを構成する事を行っている．この方法は，音響的な情報のみで話者の位置を特定できるが，決められた範囲内に話者が存在しなければならない．また，雑音源がこの範囲内に入れば話者として誤検出される．このように，この方法では話者および雑音源の位置が限定される．例えばこのような制約下で音声インタフェースをロボットに用いる場合，話者はロボットと一定の位置関係（例えばロボットの真正面）で発話する必要がある．しかし，実際にはロボットと人間は様々な位置関係になる可能性があり，発話に際してわざわざロボットの真正面に移動するのは煩わしい．また，音声の音響的な特徴を用いて話者の位置を特定する方法も考えられるが[13]，この方法では音声が発音している音源はすべて話者

として検出される．しかし，本論文で想定している環境下では，ニュース番組のようにテレビやその他のオーディオ機器からの音に人間の声が含まれる可能性がある．この場合には，テレビなどの雑音源を話者と誤って検出する可能性がある．このように，従来の手法を用いて本論文が想定する環境下で，目的音源である話者を検出する場合には，話者と音声インタフェース（マイクロホンアレイ）の間の位置関係や雑音源の種類に制限を加える必要がある．しかし，これはあまり現実的な制約ではない．

本章ではこれらの問題を解決するために，話者の位置や雑音源の種類に関する制約がより少ない目的音源検出法を提案する．提案手法ではマイクロホンアレイを用いた音響的な情報の他に，カメラを用いて画像情報も取得する．マイクロホンアレイからは音源の位置に関する情報を，カメラからは人物の位置を取得し，これらの情報から総合的に話者の位置や発話区間を検出する．以下の節ではまず提案法に必要な音源位置推定法及び人物位置推定法について述べ，続いて，これらを統合して，話者の位置と発話区間を検出する手法を提案する．最後に提案手法の評価実験を行う．

3.2　提案法の概要

図 3.1 に，提案法の概要を示す．上述のように，音響情報だけを用いて目的音源を検出する場合，従来の手法では，目的音源位置や信号の属性に制約を設ける必要があった．提案法では，図 3.1 に示すように，音響情報だけではなく，画像情報も利用する事により，この制約を緩和する．

図 3.1 (a) に示すような音響情報を用いる事により，音源の存在する方向を知る事が出来る．一方，図 3.1 (b) に示すような画像情報を用いる事により，人物の存在する方向を知る事ができる．提案法では，人物の存在する方向と，音源の存在する方向が一致する場合，これを目的音源として検出する．言い換えれば，ある場所に人物が存在する事を画像イベントの発生，ある場所で音源が発音する事を

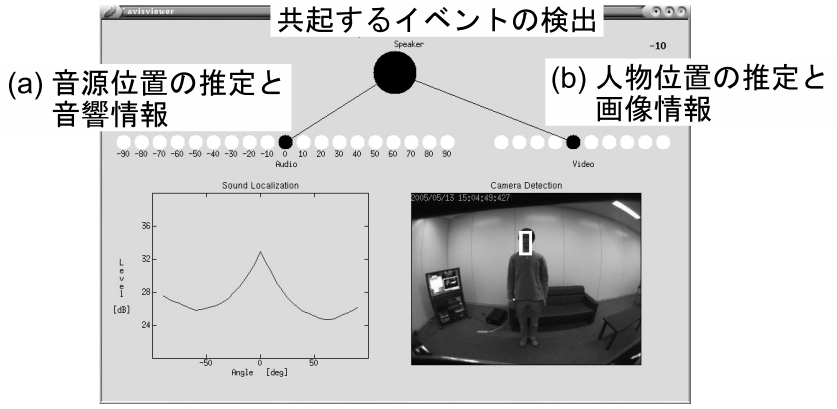

図 3.1: 提案法の概要

音響のイベントの発生と呼ぶ事にすると，両イベントが，時間的・空間的に共起する箇所を検出する事になる．このような規範を用いる事により，目的音源の位置や属性に関する制約を緩和する事が可能となる．

上述のような音響イベントと画像イベントの共起を検出するためには，音源位置の推定に用いる座標系と，人物位置の推定に用いる座標系とが一致していなくてはならない．一般に，両者を対応させるためには，キャリブレーションと呼ばれる作業が必要である．例えば，実際に観測される空間に，目印となる物体及び音源を置き，この目印の画像系での位置と，音源の音響系での位置を記録して，対応を取るなどの作業を行う．しかしながら，高精度のキャリブレーションを行ったとしても，例えば，同じ方向にありながら，人物からカメラまでの距離が異なる事により，画像系での位置が異なる事があり得る．また，音響による音源位置推定，及び画像による人物位置推定では，推定誤りが発生する可能性がある．このため，音源及び人物の位置が物理系で完全に同一でも，それぞれの観測座標系では，異なる事が起こり得る．このような曖昧性がある場合に，単純なキャリブレーションを用いて音響イベントと画像イベントの共起を検出するのは，困難である．そこで，提案法では，曖昧性を考慮した共起の検出法として，ベイジアンネットワークを用いて音響情報と画像情報を統合する手法を提案する．ベイジアンネッ

トワークは，確率的な推論を行う手法の一つであり，本手法で扱うような情報統合の問題にも，用いられている [30].

3.3　位置推定のための要素技術

本節では，3.4 節で目的音源同定法を提案するための準備として，本論文で用いる音響情報を得るための音源位置推定法と，画像情報を得るための人物位置推定法について述べる.

3.3.1　音響情報による音源位置推定

ここでは，マイクロホンアレイを用いて音源位置を推定する手法であるサブスペース法（MUSIC; Multiple Signal Classification）について述べる．今，マイクロホンアレイからの入力を短区間フーリエ変換したものを $\mathbf{x}(\omega, T)$（ω が周波数，T がフレーム番号）とすると，空間相関行列 $\mathbf{R}(\omega)$ が以下の式で求められる.

$$\mathbf{R}(\omega) = E[\mathbf{x}(\omega, T)\mathbf{x}^H(\omega, T)] \tag{3.1}$$

ここで，$E[\cdot]$ は期待値を求める操作で，$\cdot^H$ は複素共役転置である．さらに，空間相関行列の固有値分解を求める.

$$\mathbf{R}(\omega) = \mathbf{E}(\omega)\mathbf{\Lambda}(\omega)\mathbf{E}^{-1}(\omega) \tag{3.2}$$

固有値分解の結果を用いて，サブスペース法では以下の式で求められる空間スペクトルから音源の位置を推定する.

$$P_{MUSIC}(\theta, \omega) = \frac{\mathbf{a}^H(\theta, \omega)\mathbf{a}(\theta, \omega)}{|\mathbf{a}^H(\theta, \omega)\mathbf{E}_n(\omega)|^2} \tag{3.3}$$

ここで，$\mathbf{a}(\theta, \omega)$ はマイクロホンからの角度が θ の位置に対する位置ベクトルである．また，$\mathbf{E}_n(\omega)$ は固有ベクトル $\{\mathbf{e}_1(\omega) \ldots \mathbf{e}_M(\omega)\}$ のうち，小さい方から $M - N$

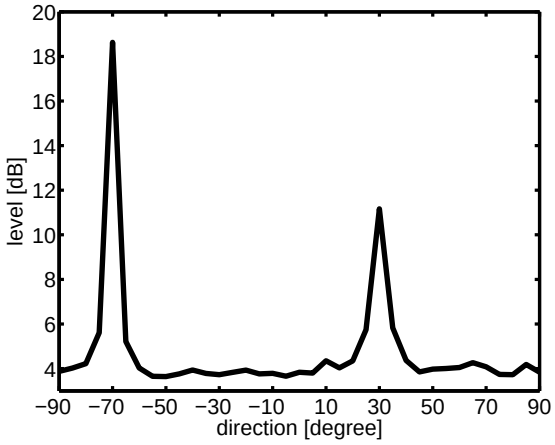

図 3.2: ある特定の帯域における空間スペクトル $P_{MUSIC}(\theta)$ の例

個の固有値に対応した固有ベクトル $\mathbf{E}_n(\omega) = [\mathbf{e}_{N+1}(\omega) \cdots \mathbf{e}_M(\omega)]$ である．ただし，M はマイクロホンの数，N は観測している空間において発音している音源の数である．

図 3.2 は，ある 1 つの帯域における空間スペクトル $P_{MUSIC}(\theta)$ の例である．図において，空間スペクトルのピークが存在する角度が音源の推定位置になる．この例では，-70 度と 30 度の方向に音源が存在すると推定される．広帯域信号に対しては，図 3.3 のように，各帯域ごとに，この空間スペクトルが得られる．ここでは，以下の方法で広帯域信号に対する空間スペクトルを求める [31].

$$\bar{P}_{MUSIC}(\theta) = \sum_{\omega=\omega_l}^{\omega_h} \bar{\lambda}(\omega) P_{MUSIC}(\theta, \omega) \tag{3.4}$$

ここで，ω_l と ω_h は空間スペクトルを合算する帯域を表している．例えば，特に音声がターゲットであれば，音声のエネルギーが集中する帯域を合算する事で推定精度を高める事ができる．また，$\bar{\lambda}(\omega)$ は

$$\bar{\lambda}(\omega) = \sum_{n=1}^{N} \lambda_n(\omega) \tag{3.5}$$

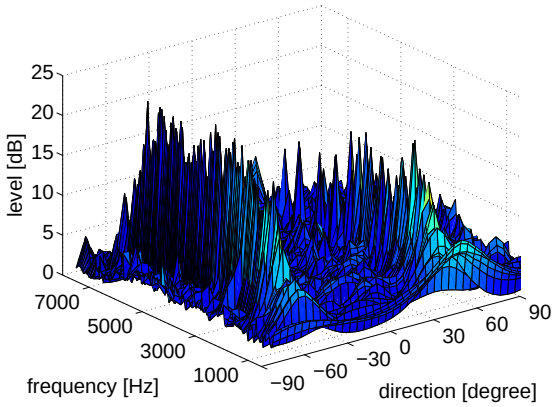

図 3.3: 広帯域での空間スペクトル $P_{MUSIC}(\theta, \omega)$ の例

である．$\lambda_n(\omega)$ は空間相関行列 $\mathbf{R}(\omega)$ を固有値分解して得られる固有値を降順に並べ替えたものである．N 個の音源の直接音のエネルギーは N 個の主要な固有値に集中するため，各帯域において音源数分の固有値にはその帯域における直接音のエネルギーが集中している．そこで，この固有値を重みとして用いる事で，音源の直接音のエネルギーが強い帯域の空間スペクトルを強調する事ができる．このようにして得られた $\bar{P}_{MUSIC}(\theta)$ のピークに対応する方向が推定音源位置となる．

3.3.2　画像情報による人物位置推定

提案手法では，カメラから得られる画像をもとに人物位置の推定を行う．人物位置の推定は，肌色モデルと顔モデルによるテンプレートマッチング [32]，並びにカーネルを用いた追跡 [33] を組み合わせる [32] 事で実現している．この手法では，カメラで観測した範囲内で，肌色が集中している領域を肌色モデルを用いて探し，これを顔が存在している位置の候補とする．図 3.4 はカメラで得られる画像の例である．この図に対して，肌色が存在する領域を検出した結果が図 3.5 である．図 3.5 において黒い点で示された領域が肌色と検出された所である（ただし，10 本の縦線は除く）．図を見ると，人物の顔に対応した領域が黒くなっており，肌色が正

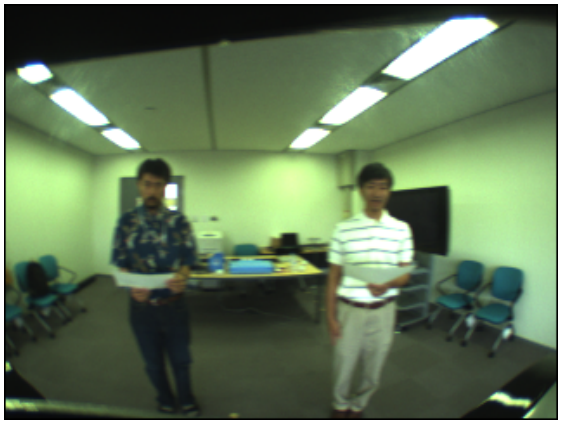

図 3.4: 人物位置推定の例（サンプルデータ）

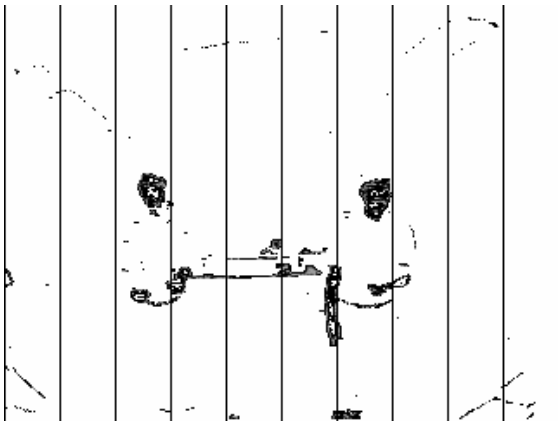

図 3.5: 肌色領域の検出例

しく検出されている．しかし，顔以外にも人物の腕などが肌色の領域として検出されている．このように，肌色の領域をそのまま人物の顔の位置としてしまうと，顔以外で肌色が集中している領域を誤って検出してしまう．そこで，さらに顔の候補となっている領域に対して顔モデルを用いたテンプレートマッチングを行い，顔の位置を検出している．図 3.6 はテンプレートマッチングに用いる顔モデルの例である．この図では，顔の特徴的な形状である目の部分が表現されている．目は，肌色ではないため検出されず，その部分が白く空いている．

図 3.6: 顔モデルの例

3.4　情報統合による目的音源同定法の提案

3.4.1　入力情報の離散化

提案手法では，3.3 節で述べた音源位置及び人物位置の推定結果を，ベイジアンネットワークを用いて統合している．ベイジアンネットワークでは，連続した情報を離散化する事で確率計算を大幅に効率化している．ここでは，まず，音源位置及び人物位置の推定結果の離散化について述べる．

3.3.1 節において，音源位置を推定する手法について述べた．マイクロホンアレイの入力から，最終的には広帯域における空間スペクトル $\bar{P}_{MUSIC}(\theta)$ が得られる．次に，この空間スペクトル $\bar{P}_{MUSIC}(\theta)$ の観測範囲を N_a 個の領域に分割し，各領域において空間スペクトルのピークがあれば，その領域に音源が存在するとみなす．さらに，各領域において音源の存在を表現するために音響仮想センサ $A_1, \ldots, A_{N_a}$ を定義する．A_i は対応する領域において音源が存在すれば $A_i = 1$，そうでなければ $A_i = 0$ の値をとる．こうして，最終的に音源位置に関する情報は，長さが N_a のベクトル（各要素の値は 1 か 0）で表現される．

図 3.7 は実際の空間スペクトルを時系列に並べたものである．図において，色が黒い所の値が大きく，色が薄い所は値が小さい．この図を $A_1, \ldots, A_{N_a}$ に変換したものが図 3.8 である．この図では，A_1 を -90 度方向とし，90 度まで 10 度おきに A_i を定義している．そのため，N_a は 19 である．色が黒い箇所が $A_i = 1$，それ以外は $A_i = 0$ である．主に，図 3.7 のピークが存在する方向に対応した箇所が，図 3.8 では $A_i = 1$ となっている．

3.3.2 節では，人物位置を推定する手法について述べた．カメラからの入力をもとに，人物の位置が入力画像のピクセル値として得られる．次に，カメラ画像の

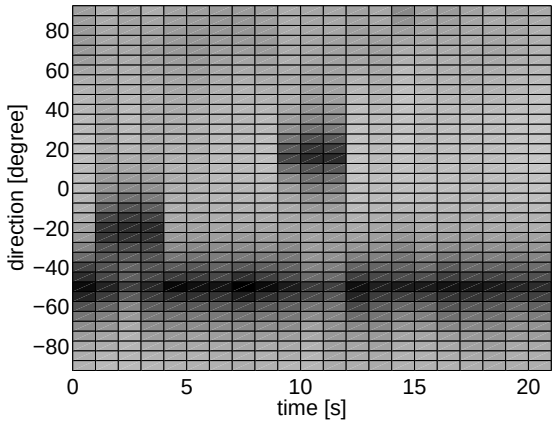

図 3.7: 音源位置推定結果の例

観測範囲も音響の場合と同様に N_v 個の領域に分割する．そして，$V_a, \ldots, V_{N_v}$ の仮想センサを考え，対応する領域内に人物が存在すれば $V_i = 1$ とし，そうでなければ $V_i = 0$ とする．図 3.5 の 10 本の縦線は $N_v = 10$ とした時の V_i の領域を示している．この例では，左から 3 番目と 7 番目の領域に人物が存在するため，V_1 を左端の領域とすると，V_3 と V_7 が 1 となり，残りが 0 となる．

図 3.9 は人物位置の推定結果を時系列に並べたものである．ここでは，横幅が 320 ピクセルの画像を用いており，人物位置の推定結果もこのピクセル値に対応して得られる．図 3.10 はこの図 3.9 を $V_1, \ldots, V_{N_v}$ に変換したものである．

3.4.2 ベイジアンネットワークの構成

ここでは，離散化された音響情報及び画像情報を統合し，目的音源を同定するためのベイジアンネットワークの構成について述べる．目的音源である話者の検出は基本的に 3.2 節で述べた通り，同一方向に音源及び人物が存在するかを調べる事で実現できる．ここで，音源位置の推定結果は 5 度単位で得られ，その情報は最終的に 1/0 のベクトル $A_1, \ldots, A_{N_a}$ として表現される．一方，カメラからの入力をもとに人物位置が推定され，その情報も音響情報の場合と同様に 1/0 のベク

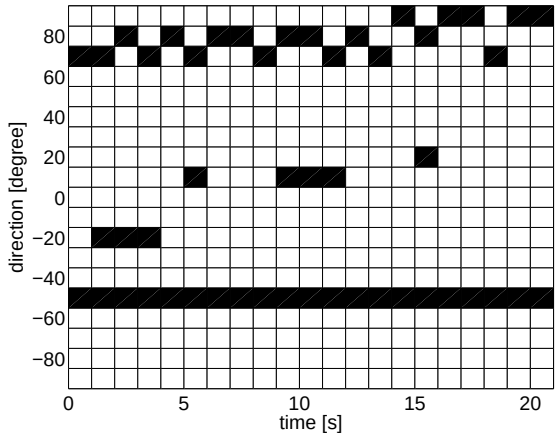

図 3.8: 図 3.7 を $A_1, \ldots, A_{N_a}$ に変換したもの

トル $V_1, \ldots, V_{N_v}$ として表現される．ここで，これらの推定値には曖昧性が存在するため，対応関係を事前に決める事は極めて難しい．また，離散的なベクトル値（$A_1, \ldots, A_{N_a}$ や $V_1, \ldots, V_{N_v}$）では，音響上では別の領域に存在していても，画像上では同一の領域に存在する場合や，その逆の場合がある．このような場合，画像上と音響上の領域を一意に対応付ける事は困難である．

提案手法では，この問題に対処するためにベイジアンネットワーク（例えば [34]）を用いる．ベイジアンネットワークは確率変数間の依存関係をネットワークで表現するものである．図 3.11 は提案手法で用いるベイジアンネットワークである．図では，S と $A_1, \ldots, A_{N_a}, V_i, \ldots, V_{N_v}$ の間に依存関係が存在する事が表現されている．この場合，この依存関係は条件付き確率を用いて $P(S|A_1, \ldots, A_{N_a}, V_1, \ldots, V_{N_v})$ と表せる．ここで，S は発話区間の検出結果を表しており，$S \in \{SS_1, \ldots, SS_{N_s}, SS_{NoEvent}\}$ である．SS_i は話者がある特定の方向から発話している事を表し，$SS_{NoEvent}$ は発話が行われなかった状態を表す．

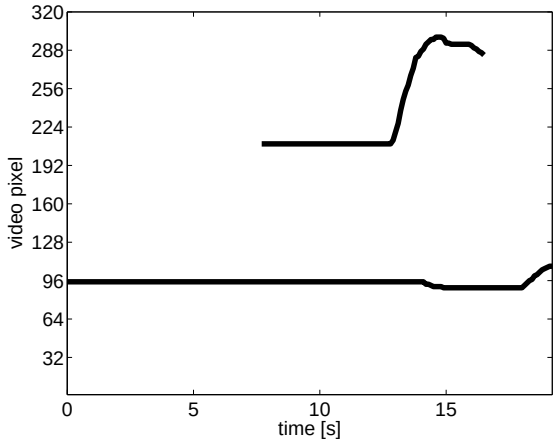

図 3.9: 人物位置推定結果の例

ここで，S は以下のように求める事ができる．

$$\arg\max_S P(S|A_1,\ldots,A_{N_a},V_1,\ldots,V_{N_v}) \tag{3.6}$$

$$= \arg\max_S P(S)\prod_{i=1}^{N_a} P(A_i|S)\prod_{i=1}^{N_v} P(V_i|S) \tag{3.7}$$

式（3.7）中の $P(A_i|S)$，$P(V_i|S)$ は特定の発話状態を条件とした時の各仮想センサの出力の確率となる．これは，あらかじめ発話状態が分かっている観測データから求める事ができる．これらは条件付き確率テーブル（CPT; conditional probability table）と呼ばれる．CPT は，$P(A_i|S)$ や $P(V_i|S)$ から分かる通り，各位置で話者が発話した時にどの A_i や V_i が対応するかを表現したものである．ただし，1 対 1 の対応ではなく，確率値を用いる事で曖昧性を含んだ形での表現が可能である．

3.4.3 ベイジアンネットワークの学習

ベイジアンネットワークを用いて S を求めるためには CPT である $P(A_i|S)$ と $P(V_i|S)$ を求める必要がある．これらは，話者が各方向（$SS_1,\ldots,SS_{N_s}$）で発話している時の A_i と V_i の出力の確率を表している．

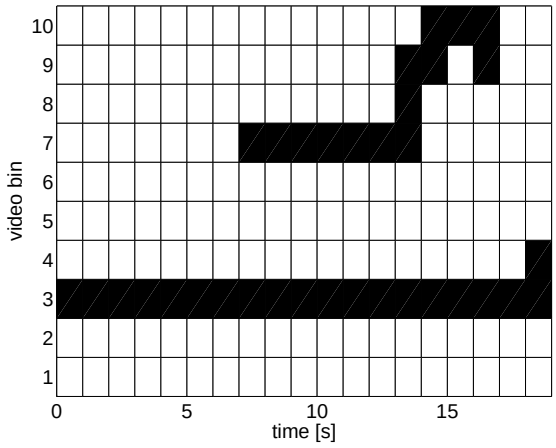

図 3.10: 図 3.9 を $V_1, \ldots, V_{N_v}$ に変換したもの

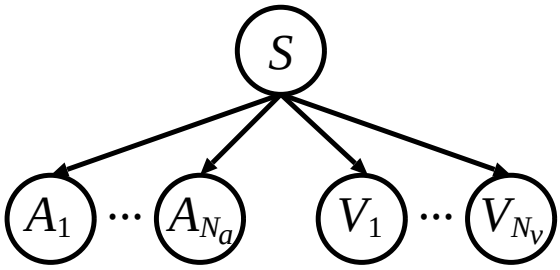

図 3.11: 提案手法で用いるベイジアンネットワーク

CPT は，A_i，V_i，S の値が既知である学習用データを用意し，これらの頻度から，全ての A_i と S，及び V_i と S の組み合わせについて，その確率値を求める事により得られる．例えば，$P(A_i|S_{-10})$ を求める場合は，雑音源が存在しない状況で，-10 度の方向で話者が発話し，この時の A_i の出力を調べる．図 3.14 はこの時の A_i の出力である．この例では，-10 度に音源が存在するため，これに対応する A_{-10} の出力のみが 1 になっている．この事より，$P(A_{-10} = 1|S_{-10}) = 1$ であり，これ以外の A_i に関しては，$P(A_i = 1|S_{-10}) = 0$ となる．この作業を各 S について行う事で，$P(A_i|S)$ や $P(V_i|S)$ を求める事ができる．

図 3.12 は $P(A_i|S)$ の例である．ここでは，S を $S \in \{SS_{-30}, SS_{-20}, SS_{-10}, SS_0, SS_{10}, SS_{20}, SS_{30}, SS$
と定義する．それぞれ SS_{-30} が -30 度方向で発話，SS_{20} が 20 度方向で発話して

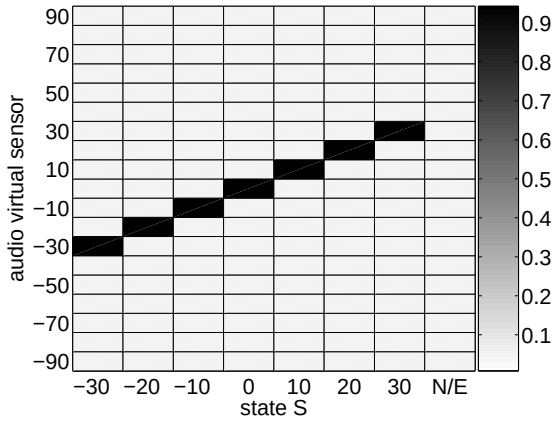

図 3.12: $P(A_i|S)$ の例

いる事を意味する．A_i はそれぞれ音源位置の方位角に対応しているため，SS_i との対応が比較的とりやすい．図においても，$S = SS_{-30}$ の時は $P(A_{-30}|S_{-30})$ の値が一番高くなっているなど，対応関係が明確である．

一方，図 3.13 は $P(V_i|S)$ の例である．この例では，V_i はそれぞれ人物の方位角に対応しているわけではない．例えば，$S = SS_{-30}$ の場合は $P(V_3|SS_{-30})$ の値が一番高くなっている．これは，SS_{-30} で定義した範囲内に人物が存在する場合は，人物位置の推定結果であるピクセル値が V_3 の範囲内に収まる事を表している．また，$S = SS_{-10}$ の場合は V_4 と V_5 の範囲に人物が存在するため，$P(V_4|SS_{-10})$ と $P(V_5|SS_{-10})$ の値が他よりも大きい．このように，実際のデータから CPT を構成する事により，ベイジアンネットワークへの入力と状態の対応関係が明確でないものについても対応関係を表現する事ができる．また，図 3.13 の $P(V_i|SS_{-10})$ のように 1 対 1 の対応関係ではない場合についても，その曖昧性を含んだ形で表現可能である．

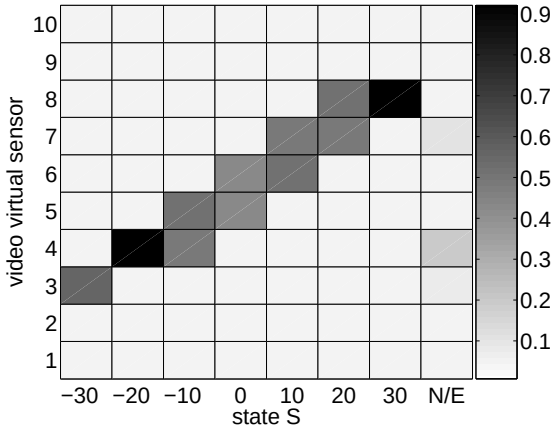

図 3.13: $P(V_i|S)$ の例

3.4.4　ベイジアンネットワークの動作

CPT が得られれば，ベイジアンネットワークへの入力と式（3.7）を用いて状態 S を推定できる．図 3.15 は図 3.8 と図 3.10 をベイジアンネットワークで統合し，S を推定した結果である．

図 3.8 に示す音響情報では，−50 度の方向に雑音源が常に存在している．しかし，図 3.10 に示す画像情報では，この方向に人物は存在しないため，図 3.15 では，この方向で発話イベントが発生しているとは推定されていない．また，図 3.8 の 5 秒過ぎから 1 秒程度，10 度の方向に音源が存在している．しかし，この時点ではこの方向に人物は存在していないため，これは発話として検出されない．一方，図 3.8 では，V_3 の方向に人物が常に存在している．しかし，この人物の方向で発話が検出されるのは，この方向で実際に発話が行われている 1〜4 秒の間のみである．

このように，ベイジアンネットワークを用いる事で，音響情報，画像情報単独では分からない話者の発話を検出する事が可能である．

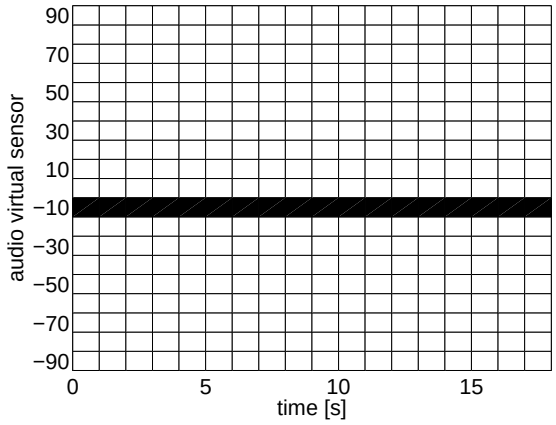

図 3.14: $S = S_{-10}$ の時の A_i の出力

3.5 提案手法の評価

3.5.1 実験条件

本節では，提案手法の性能を評価するために話者の検出実験を行った．ここでは，まずその実験条件について述べる．実験では，3 つの音源を用いた．このうちの 2 つは話者であり，もう 1 つは音楽を雑音源として用いている．図 3.16 は音源の位置を示した図である．実験では，二人の話者がそれぞれ 15 度と −25 度の位置に立ち発話し，雑音源は −90 に位置に存在する．二人の話者は雑音源が発音し続ける中で電総研音声データベース [35, 36] の 492 単語を交互に発話する．雑音源には RWC 研究用音楽データベース [37, 38] の中の一曲（RWC-MDB-J-2001 No.39）を用いた．実験を行った部屋の残響時間は約 0.5 秒であり，話者と雑音源の SN 比は概ね 0 dB となるように事前に雑音源の音量を調整している．また，マイクロホンアレイから各音源までの距離は 1.5 m である．

この実験で用いたマイクロホンアレイは直径が 0.5 m の円形のアレイであり，8 つのマイクロホンが均等に配置されている．カメラ画像の取得には PointGrey Research 社の Digiclops を用いている．図 3.17 はこれらのデバイスを示したもの

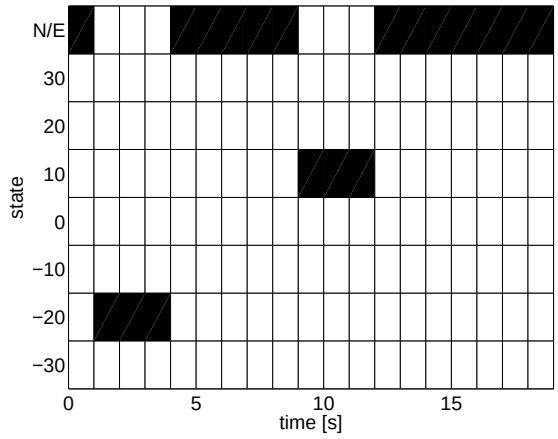

図 3.15: A_i と V_i を統合した結果

である．

マイクロホンアレイの入力から 1 秒ごとに空間相関行列及び空間スペクトルが求められ，音響情報が生成される．カメラからは横 320 ピクセル，縦 240 ピクセルの画像が 10 fps の頻度で得られ，フレームごとに音響情報が生成される．ここで，音響情報と画像情報は生成される頻度が異なる．この同期を取るために，10 フレーム文の画像情報が以下の基準で集約される．

$$V_i = \begin{cases} 1 & \sum_{t=n+1}^{n+10} V_i^t \geq 5 \text{ の時} \\ 0 & \sum_{t=n+1}^{n+10} V_i^t < 5 \text{ の時} \end{cases} \tag{3.8}$$

ここで，V_i は集約後の i 番目の仮想センサの値，V_i^t はフレーム t における i 番目の仮想センサの値，n は対応する音響情報と同期する適当な値である．

提案手法において，音源位置を推定する時に必要となるパラメータである ω_l と ω_h にはそれぞれ 500 Hz と 3000 Hz を用いた．これは，検出対象である音声のパワーが主に存在している帯域から決定している．また，サンプリング周波数は 16000 Hz，分析窓はハミング窓，分析窓長は 32 ms，周期は 10 ms である．

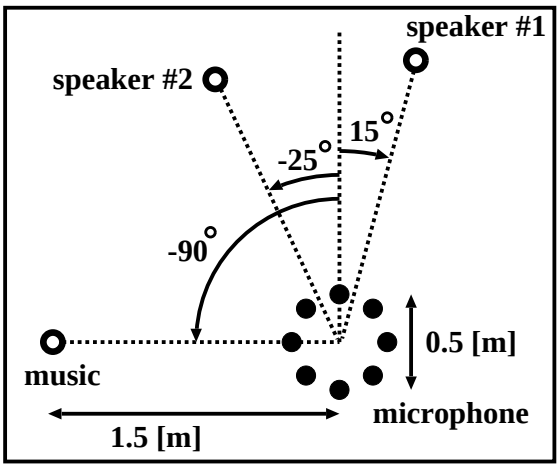

図 3.16: 実験での音源位置

表 3.1: 発話区間検出率（%）

	検出率	適合率	再現率
A	85.7	81.0	82.2
B	74.5	60.3	98.8
C	53.4	44.4	99.8

3.5.2 実験結果

本実験では，提案手法の性能を話者の発話区間の検出率によって評価する．表 3.1 は検出率を示している．表において，検出率，適合率，再現率は以下のように定義される．

$$\text{検出率} = \frac{\text{発話及び非発話を正確に検出できたフレーム数}}{\text{全フレーム数}} \tag{3.9}$$

$$\text{適合率} = \frac{\text{分母の中で実際に正解だったフレーム数}}{\text{システムが検出した発話区間のフレーム数}} \tag{3.10}$$

$$\text{再現率} = \frac{\text{分母の中でシステムが発話区間と検出したフレーム数}}{\text{正解の発話区間のフレーム数}} \tag{3.11}$$

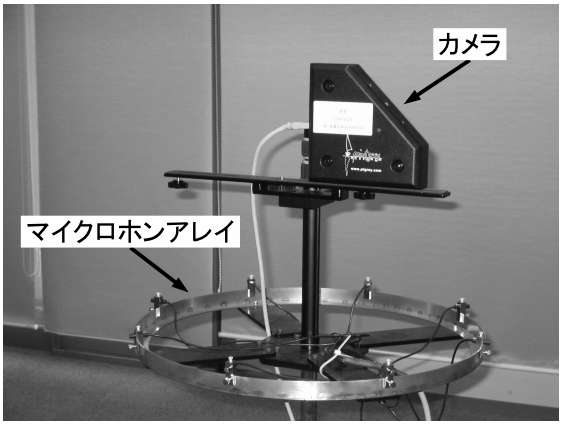

図 3.17: 実験に用いたマイクロホンアレイとカメラ

また，表において A，B，C はそれぞれ，以下のような違いがある．

A： システムが検出した発話区間をそのまま検出率などの集計に用いる．

B： システムが検出した発話区間の前後に 0.5 秒のマージンを加え，この区間全体を検出率などの集計に用いる．

C： システムが検出した発話区間の前後に 1 秒のマージンを加え，この区間全体を検出率などの集計に用いる．

図 3.18 は上記の A，B，C，の違いを示したものである．図の A がシステムが検出した区間そのものであり，B はそれに対して前後に 0.5 秒のマージンを加えている．C はさらにこれを 1 秒まで延ばしたものである．発話の始端や終端は音声のエネルギーが弱いため，音源位置を推定した場合に音源そのものが検出されない可能性がある．そのため，提案手法ではその部分を発話区間として検出できない．そこで，B や C のように前後にマージンを加える事で発話の始端や終端も検出できる事が期待される．

実験結果において A と B の場合を比較してみると，検出率と適合率が低下している一方，再現率が約 16 %上昇している．このうち，再現率の上昇は，マージンを加える事によって発話の始端や終端が検出結果に含まれた結果であると考えら

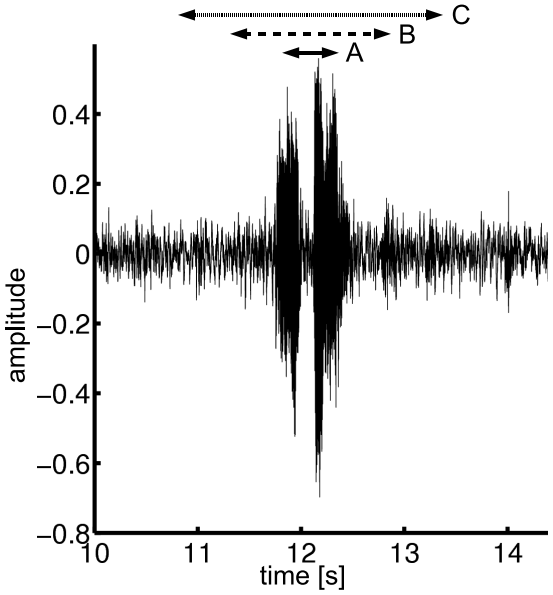

図 3.18: 発話区間の設定方法

れる．ただし，マージン部分には発話の始端や終端だけでなく，その前後に存在する非発話区間も含まれる可能性がある．検出率や適合率が低下したのはこの影響によるものと考えられる．C はマージンをさらに加えているため，検出率と適合率はさらに低下し，再現率は上昇している．

本論文で提案しようとしている音声インタフェースにおいては，話者の発話を検出できなければ，その先の音源分離などの処理が全く行えない．そのため，多少非発話区間を発話区間と誤検出しても，発話区間を網羅するように検出できなければならない．これは，適合率よりも再現率が高い状況である．その点から，A よりも B の方が提案手法にとっては好ましいケースである．また，B と C を比較してみると，C は B よりも再現率が上昇しているが，その差はわずかに 1 %である．その一方で適合率は 20 %以上低下している．これは，B の時点で発話区間はほぼ全て検出できており，C においてさらに付加した 0.5 秒のマージンはほぼ全て非発話区間を発話区間と誤検出していると考えられる．これらの事より，提案手法が検出した発話区間の前後に 0.5 秒のマージンを加えると，特に再現率という点

から高い性能が得られる事が分かる．

3.6　おわりに

本章ではまず，発話区間を検出する際の基本的な考え方について述べた．本章では，音源及び人物が同一の方向に存在する場合，その方向に話者が存在し，その話者が発話しているとみなしている．この際，音源位置推定，及び人物位置検出においては，検出結果に対して曖昧性が存在する．そのため，単純に音響情報と画像情報を 1 対 1 に対応させる事は困難である．この問題を解決するために，提案手法ではベイジアンネットワークを用いている．

次に，音響情報及び画像情報をどのように生成しているかについて述べた．両情報は最終的に適当な長さの 1/0 の要素を持つベクトルとして表現される．さらに，この情報をベイジアンネットワークで統合する方法について述べた．ベイジアンネットワークで統合を行う際には，条件付き確率を計算する過程で，CPT が必要になる．CPT は音響情報と発話の状態，画像情報と発話の状態についての対応関係を表現したものであるが，確率値を用いる事で両情報の曖昧性を含んだ形で表現可能である．

最後に，提案手法の評価実験を行った．実験では，本論文が想定するような，雑音源が発音し続ける中で話者が発話する状況下で，話者の発話区間検出を行った．この際，話者の状態 S が変化するように，二人の話者が交互に発話した．この実験の結果，検出した発話区間の前後に 0.5 秒のマージンを加える事で，発話区間の 98.8 %の検出に成功した．

第4章　音響情報と画像情報を統合したロバスト音声インタフェース

4.1　はじめに

第 1 章でも述べた通り，実環境で動作する音声インタフェースには以下のような技術が必要である．

- 話者検出
- 音源分離
- 音声認識時の音響モデル適応

前章まででは，音声インタフェースにおける話者検出の手法について述べた．まず，第 2 章では，SVM（Support Vector Machines）及び SVR（Support Vector Regression）を用いた音源重畳区間検出法を提案した．提案手法を用いる事で，実環境下において複数の音源が重畳している区間を検出する事が可能である．重畳区間は，話者と雑音源が同時に発音しているのか，話者や雑音源が単独で発音しているのかを推定するための重要な情報となる．また，第 3 章ではマイクロホンアレイを用いた音源位置の推定結果とカメラを用いた人物位置の推定結果をベイジアンネットワークで統合する事により，目的音源を同定し，発話区間を検出する手法を提案した．提案手法を用いる事で，雑音源が発音し続ける中で発話する話者が，どのタイミング・方向で発話しているかを検出する事が可能である．

本章では，前章までに提案した手法に加え，音源分離などの手法を組み合わせる事で，実環境下で動作するロバストな音声インタフェースを提案する．本章で提案する音声インタフェースは，主に 3 つの部分から構成されている．1 つ目は，話者検出であり，第 2 章及び第 3 章で提案した手法を融合し，各音源の発話の状態，すなわち音源が目的音源であるかどうか，及び単独発話なのか，他の音源と重畳しているのかなどを推定する．これらの情報は次の音源分離で用いられる．2 つ目の音源分離では，音源分離の処理を行う．この際，話者検出で得られた話者の音源位置などの情報を事前情報として用いる事で，ML 法などの分離性能の高い手法を用いる事が可能である．最後の音声認識では，音源分離の出力に対して音声認識を行う．音源分離後の信号には部屋の反射の影響などによる残留雑音が存在している．この残留雑音の影響により，認識性能が低下してしまうため，音声認識では音響モデルに対してモデル適応の処理を行う事で認識性能の向上を図っている．

本章では，まず最初に，提案インタフェースの構成について述べ，提案インタフェースを構成する 3 つの部分について詳述する．次に，提案インタフェースの評価実験を行う．実験では，まず，提案インタフェースの話者検出について，様々な状態を含んだデータに対して検出実験を行う．さらに，実際に雑音環境下で話者が発話したデータに対して，提案インタフェースを用いて音声認識を行い，その認識性能から提案インタフェース全体の性能を評価する．

4.2　提案する音声インタフェース

4.2.1　提案する音声インタフェースの構成

図 4.1 は本章で提案する音声インタフェースの構成を示したものである．提案インタフェースは大きく分けて 3 つの部分から構成されている．

(a) 話者検出

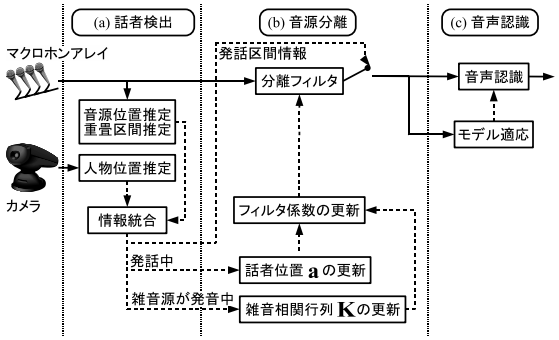

図 4.1: 提案インタフェースの構成

話者検出では，第 2 章及び第 3 章で提案した手法を用いて，話者がどのタイミングで，どの方向で発話しているかを検出する．すなわち，マイクロホンアレイからの入力をもとに音源重畳区間及び音源位置を推定し，カメラからの入力をもとに人物位置を推定する．そして，それらの情報をベイジアンネットワークを用いて統合する事で話者の発話や雑音源の単独発音などの情報を推定している．これらの情報は次の音源分離を行う際に用いられる．

(b) 音源分離

音源分離では，マイクロホンアレイの入力から，目的音である話者の音声のみを分離する．提案インタフェースでは，音源分離の手法に第 1 章で述べた ML 法を用いている．ML 法は，音源分離を行う際に，話者の音源位置と雑音源が単独で発音している区間の情報を必要とする．提案インタフェースでは，話者検出で得られた情報を ML 法の事前情報として用いている．

音源分離では，マイクロホンアレイの全入力に対して音源分離の処理を行う．しかし，音声認識の対象となる音声が含まれているのは，このうちの一部の区間である．そこで，音源分離では話者検出で得られた情報をもとに，話者の発話が含まれている区間のみを音声認識に送る．

(c) 音声認識

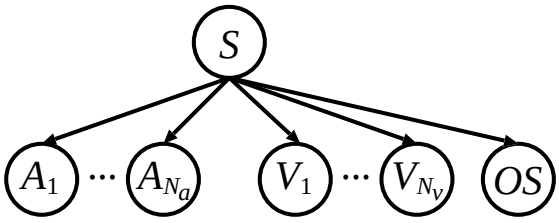

図 4.2: 提案インタフェースで用いるベイジアンネットワーク

音声認識では，音源分離から送られてきた話者の発話に対し，音声認識の処理を行う．本論文では，話者は音声インタフェースに対して「1 チャネル」，「ビデオの再生」といったコマンドを孤立単語として発話する事を想定している．そのため，音声認識で用いる言語制約には，これらのコマンドを単位としたネットワーク文法を用いている．

また，音源分離後の音声には部屋の反射の影響などによる残留雑音が存在している．この残留雑音の影響で，音声認識の音響モデルと入力音声の間でミスマッチが発生し，音声認識の性能が著しく劣化する．この問題に対処するため，ここでは，音響モデルの適応処理を行う事で対処している．

4.2.2　話者検出

話者検出では，マイクロホンアレイとカメラからの入力をもとに，

- 話者が単独で発話している．
- 話者が雑音源と同時に発話している．
- 雑音源のみが単独で発音している．

の 3 つの状態を推定する．さらに，話者が発話している場合は，話者の音源位置も併せて推定する．

提案インタフェースでは，第 3 章で述べた方法で音源位置及び人物位置を推定し，ベイジアンネットワークへの入力を生成している．しかし，これだけでは，話

者が単独で発話しているか否かを推定する事が難しい．そこで，提案インタフェースでは，音源位置と人物位置の情報に加え，第 2 章で提案した重畳区間の情報も用いる事ができるように，ベイジアンネットワークを拡張する．具体的には，変数OSを定義し，該当する区間が重畳区間と推定されれば$OS = 1$とし，そうでなければ$OS = 0$とする．これにより，重畳区間の推定結果も，音源位置などの推定結果と同様に 1/0 のベクトル値で表現され，ベイジアンネットワークの入力とする事が可能である．

図4.2は提案インタフェースで用いているベイジアンネットワークである．$A_1, \ldots, A_{N_a}$と$V_1, \ldots, V_{N_v}$はそれぞれ音源位置及び人物位置を表現したベクトルである．また，Sは$S \in \{SS_1, \ldots, SS_{N_s}, OS_1, \ldots, OS_{N_s}, N\}$である．ここで，$SS_i$は話者がある特定の方向から単独で発話している事を表す．また，OS_iは雑音源が発音している中で話者がある特定の方向から発話している事を表す．SS_iやOS_iは話者の音源位置に対応しているため，状態の推定結果がこれらの場合は，話者の音源位置$\hat{\theta}$を推定する事ができる．Nは話者は発話せず，雑音源のみが発音している状態である．この時，Sは以下のように求める事ができる．

$$\arg\max_S P(S|A_1, \ldots, A_{N_a}, V_1, \ldots, V_{N_v}, OS) \tag{4.1}$$

$$= \arg\max_S P(S)P(OS|S)\prod_{i=1}^{N_a} P(A_i|S)\prod_{i=1}^{N_v} P(V_i|S) \tag{4.2}$$

ここで，$P(OS|S)$，$P(A_i|S)$，$P(V_i|S)$は条件付確率値であり，ベイジアンネットワークの入力である仮想センサの値と 3.4.3 節で述べた条件付確率値テーブル（CPT）から決定される．CPT は 3.4.3 で述べた通り，あらかじめ学習などにより求めておく必要がある．

4.2.3 音源分離

音源分離では，話者検出で推定された情報をもとに，必要に応じて音源分離用のフィルタを更新し，音源分離の処理を行う．提案手法では ML 法を用いて音源

分離を行っており，分離フィルタは以下の式で求められる．

$$\mathbf{w}_{ML}(\omega) = \frac{\mathbf{K}^{-1}(\omega)\hat{\mathbf{a}}(\theta,\omega)}{\hat{\mathbf{a}}^{H}(\theta,\omega)\mathbf{K}^{-1}(\omega)\hat{\mathbf{a}}(\theta,\omega)} \tag{4.3}$$

ここで，$\hat{\mathbf{a}}(\theta,\omega)$ は分離対象となる音源から各マイクロホンへの伝達関数を要素とする位置ベクトルであり，$\mathbf{K}(\omega)$ は雑音のみの空間相関行列である．$\hat{\mathbf{a}}(\theta,\omega)$ は，事前に適当な音源位置 $\theta_1,\ldots,\theta_{N_\theta}$ において $\mathbf{a}(\theta_i,\omega)$ を測定しておき，話者検出において推定された話者の音源位置 $\hat{\theta}$ から最適なものを選択する事で決定できる．$\mathbf{K}(\omega)$ は，雑音源の位置が定常であれば，雑音源が単独で発音している区間から推定される．

このように，ML 法において分離フィルタを求めるためには，以下の情報が必要となる．

- 分離対象である話者の音源位置.
- 雑音源が単独で発音している区間.

これらの情報は，話者検出で推定される S に含まれている．すなわち，S が SS_i や OS_i であれば，話者が発話している事を意味している．また，S が N であれば，雑音源が単独で発音している事を意味している．そこで，話者検出の推定結果をもとに，以下の基準で $\hat{\mathbf{a}}(\theta,\omega)$ と $\mathbf{K}(\omega)$ を更新する．

- $S \in \{SS_1,\ldots,SS_{N_s},OS_1,\ldots,OS_{N_s}\}$ であれば，話者が発話している．この場合は，話者の音源位置に対応する伝達関数を $\hat{\mathbf{a}}(\theta,\omega)$ として更新する．
- $S \in \{N\}$ であれば，雑音源が単独で発音している．この場合は，この区間における空間相関行列を求め，$\mathbf{K}(\omega)$ として更新する．

先にも述べた通り，音声インタフェースに入力される信号のうち，実際に話者が発話している区間はその一部である．話者が発話していない区間を音声認識に送る場合，その部分も認識の対象となり意図しない認識結果が出力され，認識誤

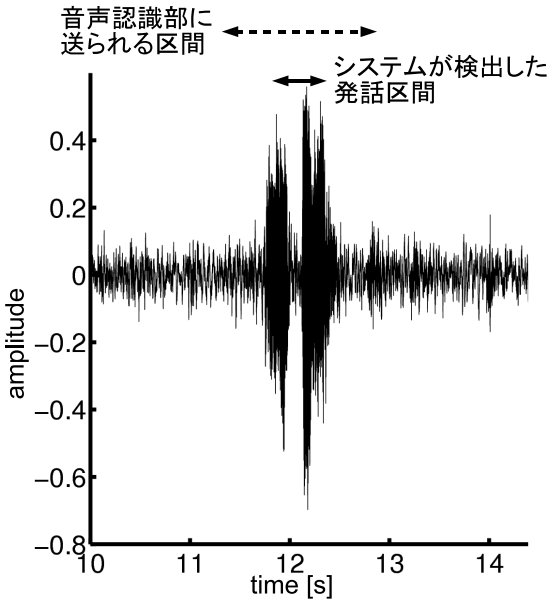

図 4.3: 音声認識へ送る発話区間

りとなる．音源分離ではこの問題に対処するため，分離フィルタを通過した信号のうち，話者が実際に発話している区間のみを音声認識へ送る．この音声認識へ送る区間は，話者検出で推定された状態 S から決定する事ができる．図 4.3 は音源分離が音声認識へ送る発話区間を示した図である．図において，実線で示した区間が話者検出で推定された発話区間である．これに対し，破線は音声認識へ送られる区間を示している．第 3 章での実験結果で述べた通り，発話の始端や終端は音声のエネルギーが弱いため，発話区間として検出されない可能性がある．提案インタフェースでは，話者検出において推定された発話区間の前後に適当なマージンを加える事で，話者の発話全体を音声認識へ送る事が可能である．

4.2.4 音声認識

音声認識では，音源分離から送られてきた話者の発話に対して音声認識の処理を行う．本論文では，話者が家電などの各種機器を制御するコマンドを音声インタフェースに発話する事を想定している．また，音源分離では分離信号を話者の

発話ごとに分割する．そのため，音声認識へは話者の発話ごと，すなわちコマンドごとに分割された音声が入力される．そこで，音声認識ではこの事を利用し，コマンドを単位としたネットワーク文法を用い，孤立単語認識を行う．これにより，連続単語認識を行う場合に比べ，認識時の単語の挿入などの認識誤りが抑えられる事が期待できる．

音源分離後の音声には部屋の反射の影響などによる残留雑音が存在している．この残留雑音の影響で，音声認識の音響モデルと入力音声の間でミスマッチが発生し，音声認識の性能が著しく劣化する．この問題に対し，音声認識では，音響モデルの適応処理を行う事で対処している．ここでは，MLLR [39] によってモデルパラメータの変換を行い，それを事前知識として MAP 推定 [40] を行う MLLR-MAP [41] を用いている．また，この際，提案インタフェースではオンラインで適応処理を行う事を想定している．そのため，適応時に正解ラベルを教師データとして与える事ができない．そこで，ここでは，初期モデルから教師データを自動で生成する教師なし適応 [42] を行っている．

4.3　提案するインタフェースにおける話者検出の評価

4.3.1　実験条件

本節では，提案するインタフェースの総合的な評価を行う前に，話者検出の評価を行う．まず，この実験の実験条件について述べる．

実験は，話者及び雑音源が存在する中程度の大きさの部屋で行った．図 4.4 は話者と雑音源の位置関係を示したものである．実験は，話者が発話により，オーディオ機器を操作する事を想定し行った．表 4.1 は話者の発話と雑音源の状態を示したものである．まず，最初に話者が 2 回発話を行う（表中の区間 1 と区間 2）．この時，雑音源である音楽はまだ発音していない．この後，区間 2 の発話で音楽が再生され，区間 3 から区間 5 の発話は音楽と重畳する．区間 3，区間 4 の発話では音

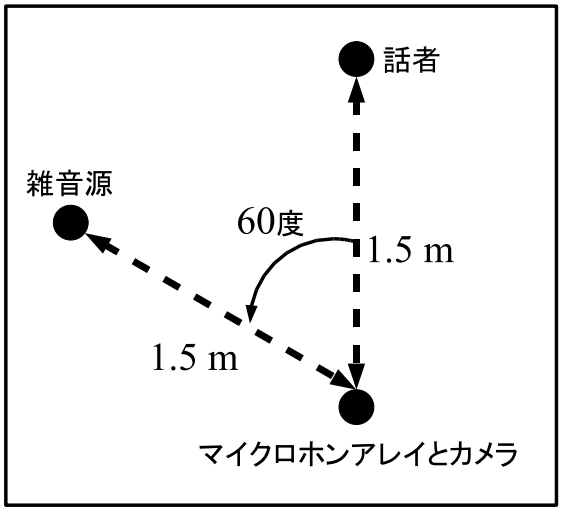

図 4.4: 話者と雑音源の位置関係

楽を切り替える指示を出し，発話ごとに音楽が切り替わる．区間 5 の発話で音楽が停止し，区間 6 と区間 7 は雑音源が発音していない状況での発話となる．

この実験では，

- 雑音源が発音していない状況で，話者が単独で発話．
- 話者が雑音源に重畳して発話．
- 雑音源が単独で発音．

の 3 つの状態が存在する．実験ではこの 3 つの状態を図 4.2 のベイジアンネットワークを用いて検出する．

図 4.5 は実験に用いたマイクロホンアレイとカメラである．本実験では，ヒューマノイドロボット HRP-2 の頭部に搭載したカメラとマイクロホンアレイを用いている．マイクロホンアレイは頭部の周囲に配置された 8 つのマイクロホンで構成されており，カメラは頭部中央に設置されている．

表 4.1: 実験時の話者の発話雑音源の状態

区間	話者の発話	雑音源
1 2	HRP-2 11 号機こんにちは 音楽を再生して下さい	off
3 4 5	次の曲に進んで下さい 次の曲に進んで下さい 音楽を止めて下さい	on
6 7	HRP-2 どうもありがとう さようなら	off

4.3.2　実験結果

図 4.6 はマイクロホンアレイからの入力を用いて得られた音響情報である．図の一番上の段はマイクロホンアレイへの入力信号，中段は入力データの正解，一番下の段は音響情報を示している．音響情報の一番下の行は重畳区間に関する情報（OS）であり，黒で示したフレームが重畳区間と検出されている．さらにその上の行はそれぞれ 90 度から -90 度までの各仮想センサ（A_i）の出力を示したものであり，黒で示したフレームが $A_i = 1$ のフレームである．この図から，話者が最初に雑音のない状況で断続的に発話し，その後発音し始めた雑音源のもとでさらに断続的に発話し，最後に再び話者が単独で発話している様子が分かる．

図 4.7 は人物位置の検出結果である．今回の実験では，観測範囲を 10 等分し，それぞれの領域において仮想センサの出力を求めている．この図から，人物が画面ほぼ中央で静止している様子が分かる．

図 4.8 は提案手法で発話状態を推定した結果と，それに対する正解である．図の実線が正解，破線が推定結果である．図を見ると，ほぼ全てのフレームについて正確に状態を推定できている事が分かる．この実験における検出率は 96 %である．また，この図は OS，SN，SS の違いのみを示しているが，音源方向に関しても，話者が発話している方向を正しく推定している．

図 4.5: 実験に用いたマイクロホンアレイとカメラ

表 4.2: 実験で用いた音声認識器のパラメータ

特徴量	MFCC 26 次元
フレーム長	25 ms
フレーム周期	10 ms
音素数	43

4.4 提案するインタフェースの評価

4.4.1 実験条件

本節では，提案インタフェースの全体の評価実験を行う．実験では，本論文が想定する環境下で話者が提案インタフェースに対して発話する．提案インタフェースは，話者の発話を検出し，音源分離を行い，音声認識を行う．音声インタフェースの性能は，最終的な音声認識率で評価した．ここでは，この実験の実験条件について述べる．

実験では，第 3 章で行った実験で収録されたデータを用いた．音源位置や音源の種類などの実験条件は 3.5.1 節で述べた通りである．ただし，話者・雑音源とマ

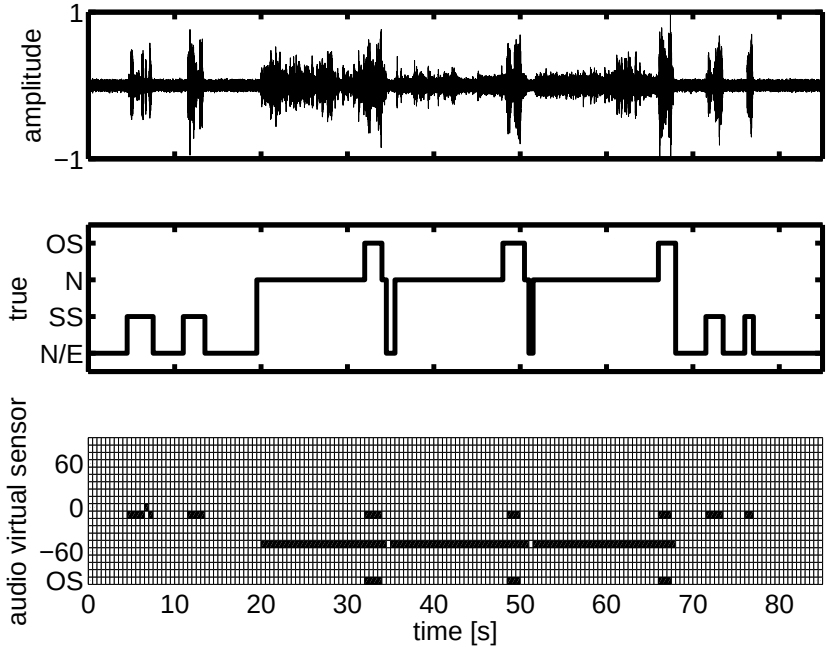

図 4.6: 音響情報（OS，$A_1, \ldots, A_{N_a}$）

イクロホンアレイの距離は 1.5 m に加え 3.0 m についても実験を行っている．提案インタフェースは，このデータに対し話者の発話を検出し，その結果をもとに音源分離・音声認識を行う．音声認識器には HTK Ver. 3.2 [43] を用いた．音声認識器で用いる音響モデルには，連続音声認識コンソーシアム 2002 年度版ソフトウェア（ベータ版）[44] で提供されている monophone モデル（男女不特定話者）を用いた．表 4.2 は音声認識器の各種パラメータを示したものである．

提案インタフェースでは，検出した発話区間に応じて入力信号を区切り音声認識へ送っている．この実験では，話者の発話が各単語に区切られ，音声認識へ送られる．音声認識ではこの入力に対して，孤立単語認識を行う．このような方法を用いる事で，連続単語認識を行うより，認識時における単語の挿入や削除を大幅に抑える事ができる．

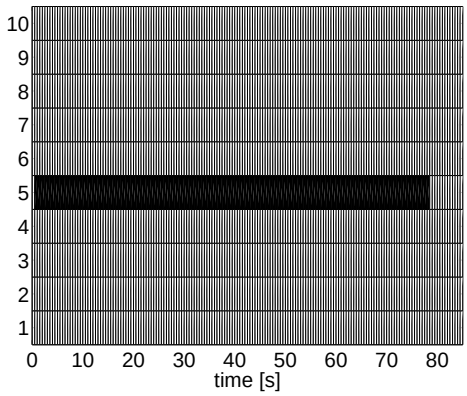

図 4.7: 画像情報（$V_1, \ldots, V_{N_v}$）

表 4.3: 単語正解精度

条件	音源分離	発話区間分割	モデル適応	単語正解精度	
				1.5 m	3.0 m
A	on			26.0 %	−36.0 %
B		on		29.7 %	9.6 %
C	on	on		79.7 %	54.3 %
D	on	on	on	91.1 %	80.1 %

4.4.2 実験結果

表 4.3 は実験で得られた結果を示している．表において，条件 A，B，C，D はそれぞれ 3 つの処理である音源分離，発話区間分割，モデル適応の処理の組み合わせが異なる．各条件において，on と示されている処理が有効になっている．ここで，音源分離は ML 法を用いた音源分離の処理を，発話区間分割は話者検出において推定された発話区間に基づいて図 4.3 で示された発話区間に区切る処理を，モデル適応は音声認識におけるモデル適応の処理を意味している．

本実験において，提案インタフェースの性能は単語正解精度によって評価して

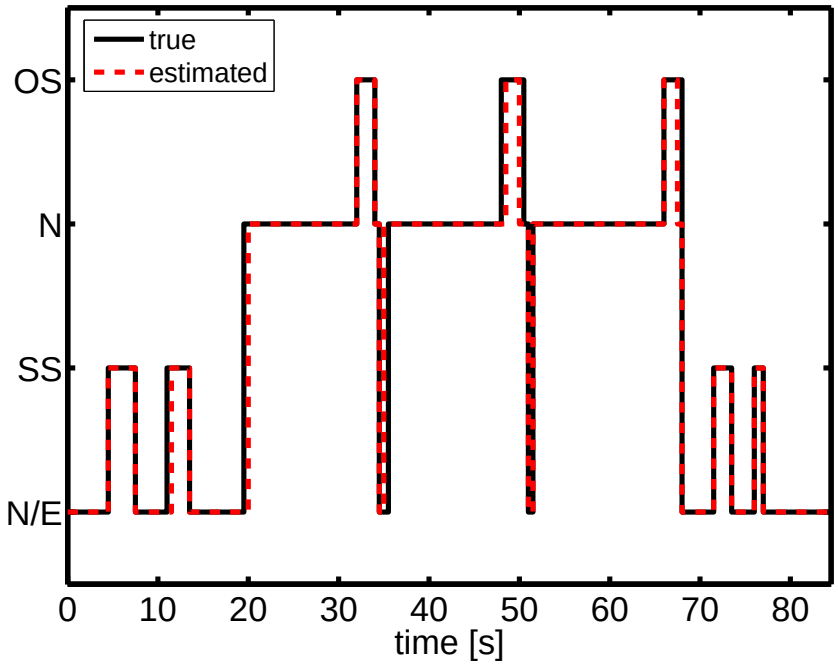

図 4.8: 正解と検出結果

いる．単語正解精度の定義式は以下の通りである．

$$\frac{N_w - \text{Subs.} - \text{Dels.} - \text{Ins.}}{N_w} \times 100 \tag{4.4}$$

ここで，N_w は認識単語数であり，この実験では $N_w = 492$ である．Subs. は置換誤りとなった単語の数，Dels. は削除誤りとなった単語の数，Ins. は挿入誤りとなった単語の数である．本実験では基本的に孤立単語認識を用いているため，この段階では削除誤り及び挿入誤りは発生しない．しかし，最終的には，話者が発話した単語の系列を正解とし，これに対して音声認識が出力した認識結果の系列の単語正解精度を算出している．図 4.9 は本実験で起こり得る置換誤りの例である．この図は，話者が「あんぎゃ」,「あやうい」,「うるおう」と発話し，話者検出ではその発話を全て正しく検出できた事を表している．しかし，最後の「うるおう」の発話を「ぶんうん」と誤って認識したため，これが置換誤りとなっている．これは，主に音源分離や音声認識の性能に依存する．図 4.10 は削除誤りの例である．この

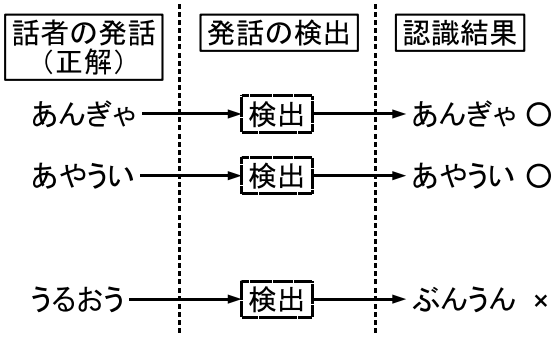

図 4.9: 置換誤りの例

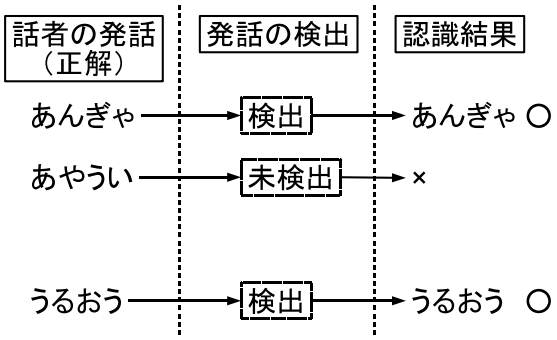

図 4.10: 削除誤りの例

図は，話者の発話のうち「あやうい」を話者検出で検出できなかった事を表している．そのため，この発話に対応する認識結果が得られず，これが削除誤りとなっている．図 4.11 は挿入誤りの例である．この図は,「あやうい」と「うるおう」の間に，実際には存在しない発話を誤って検出した事を表している．このため，この誤検出した発話区間が音声認識に送られ，対応する認識結果が挿入誤りとなっている．この削除誤りと挿入誤りは話者検出の性能に依存する．このように，単語正解精度を評価値として用いる事で，話者検出・音源分離・音声認識の性能を総合的に評価する事ができる．

表 4.3 において条件 A は，音源分離の処理のみを行った結果である．この場合，分離結果は発話区間に分割されないため，音声認識において孤立単語認識が行え

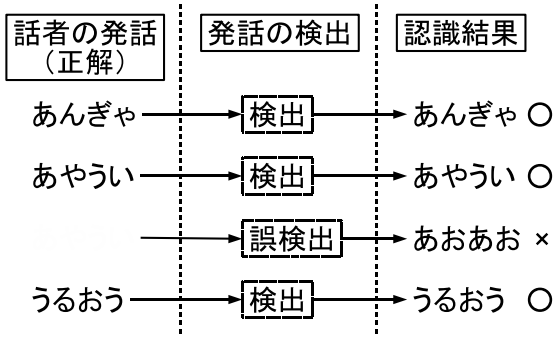

図 4.11: 挿入誤りの例

ない．そのため，この条件に限って音声認識では連続単語認識を行っている．また，条件 B では，音源分離を行わずに発話区間分割のみを行っている．条件 A の場合は連続音声認識を行うため，音声認識に依存した挿入誤りが発生する．このため，条件 A の距離が 3.0 m の場合は正解精度が極めて低い値になっている．これに対し，条件 B の距離が 3.0 m では，正解精度が大幅に回復しており，発話区間に分割する処理が有効に機能している事が分かる．一方，条件 A と条件 B の距離が 1.5 m の場合は，結果に大きな差が見られない．これは，条件 A と条件 B では 2 つの機能がそれぞれ 1 つずつ有効になっており，その効果が相乗的に表れないためである．

条件 C では，音源分離と発話区間分割の処理が同時に有効となっている．この時，1.5 m と 3.0 m のどちらとも，条件 A，B より高い正解精度が得られている．この事より，音源分離と発話区間分割の処理が相乗的に高い効果を発揮している事が分かる．また，条件 D では 3 つの処理が全て有効になっている．この時には，1.5 m と 3.0 m のどちらも全ての条件の中で最も高い正解精度となっている．これらの事より，提案インタフェースで行われている音源分離，発話区間分割，モデル適応の処理はいずれも有効に機能している事が分かる．また，これらの機能を同時に用いる事で，相乗的に高い性能が得られた．

4.5 おわりに

本章では，実環境下で動作するロバスト音声インタフェースを提案した．提案インタフェースは主に 3 つの部分から構成されている．話者検出では，第 3 章で提案した手法をもとに，話者の発話を検出する．この際，第 2 章で提案した重畳区間情報を用いる事で，話者が雑音源に重畳して発話しているのか，雑音源が単独で発音しているかなどの状態を推定する事が可能である．これらの推定結果は，後段の音源分離で用いられる．

音源分離では，ML 法を用いて音源分離の処理を行っている．ML 法では，分離フィルタを求める際に，話者の音源位置や雑音源が単独で発音している区間に関する情報を必要とする．これらの情報は話者検出で推定可能な情報であり，これらの情報を事前情報として用いる事で ML 法のような分離性能の高い手法を用いる事が可能である．また，音源分離では，話者検出で推定された発話区間に基づいて，分離信号を発話ごとに分割する．これにより，音声認識に送られる分離結果は発話ごとに区切られる．

音声認識では，音源分離から送られてきた発話に対して音声認識を行う．音声認識への入力は発話ごとに区切られているため，各発話を単位とした孤立単語認識を行う事が可能である．これにより，連続単語認識よりも挿入誤りが少ない事が期待される．また，音源分離後の信号には，部屋の反射の影響などによる残留雑音が存在している．そこで，音声認識では音響モデルに対してモデル適応の処理を行う事でさらなる認識性能の改善を図っている．

続いて，本章では提案インタフェースの評価実験を行った．まず，話者と雑音源がともに断続的に発話する状況下で，話者検出の評価を行った．この実験の場合，話者が単独で発話する，話者が雑音源に重畳して発話する，雑音源が単独で発音する，の 3 つの状態が存在する．実験では，話者検出はこれらの状態を正確に推定できた．

最後に，音声認識まで含めた提案インタフェース全体の性能を評価した．実験

では第 3 章と同じ実験条件において，話者の発話を認識し，その単語正解精度を算出した．この際，提案インタフェースの 3 つの機能である音源分離，発話区間分割，モデル適応の 3 つの処理を有効・無効にし，結果を比較した．実験の結果，音源分離と発話区間分割のそれぞれを単独で有効にした場合はその影響が限定的であった．しかし，この 2 つを同時に有効にする事で相乗的に高い性能が得られた．また，さらにモデル適応の処理も有効にする事で，音源距離が 1.5 m の場合で 91.1 %，3.0 m の場合で 80.1 %という高い単語正解精度が得られた．これらの結果より，提案インタフェースは実環境下で有効に機能する事が示された．また，提案インタフェース中の各処理が相互に補完する事で，高い性能を達成している事も示された．

第5章　提案インタフェースのリアルタイムシステム

5.1　はじめに

第 4 章では実環境下でロバストに動作する音声インタフェースを提案した．提案インタフェースは，音響情報と画像情報をベイジアンネットワークで統合する事により話者の発話などを検出する．さらに，この検出結果に基づいて音源分離・音声認識を行う．評価実験の結果，提案インタフェースが実環境下で有効に機能する事が示された．これらの実験はあらかじめマイクロホンアレイとカメラからの入力を収録し，オフラインで行ったものである．しかし，提案インタフェースを実際に実環境下で用いる際には，マイクロホンアレイとカメラからの入力を逐次処理し，リアルタイムで動作する事が必要である．

一方で，近年，家庭やオフィスなどの環境でサービスを提供する事を目標としたロボットが多数開発されている．ロボットが提供するサービスとしては，以下のようなものが考えられる．

1. ペットロボットなどに代表されるようなエンターテイメント．
2. 物を取ってくるなどの動作支援や介助．
3. 情報家電機器や情報検索などネットワーク上のリソースのためのインタフェース．

ロボットがこうした環境でサービスを行うためには，音声を用いて人間とコミュニケーションをとる事が必要不可欠である．しかし，家庭やオフィスを含む実環

図 5.1: ヒューマノイドロボット HRP-2

境には様々な雑音が存在する．このため，音声認識システムを単純にロボットに搭載するだけでは，ロボットと人間がコミュニケーションをとる事が難しい．そこで，このような環境下でも十分な音声認識の性能が得られる，ロバスト音声インタフェースが重要となる．本章では産業技術総合研究所と川田工業（株）が共同で開発したヒューマノイドロボット HRP-2（図 5.1）に提案インタフェースをリアルタイムシステムとして実装する．また，特に先の 2. 及び 3. のケースを想定し，実環境下でリアルタイムシステムの評価実験を行う．

ロボットにシステムの実装を行う場合，全ての処理はロボット内部の計算資源，及びロボットがネットワークを介して利用できる計算資源を用い，リアルタイムで実行されなければならない．ネットワーク上の計算資源を利用する場合も，ロボットとネットワークは無線 LAN で接続されているため，高帯域な有線 LAN の場合のように大量のデータを短時間で転送する事はできない．また，ロボット内部の計算資源は，スペースや電力などの点で著しい制約を受けている．一方，提案インタフェースは，マイクロホンアレイ及びカメラからの大量の情報（音響データ：2 Mbps，画像データ：28 Mbps）の処理を必要とする．このため，音響信号処理

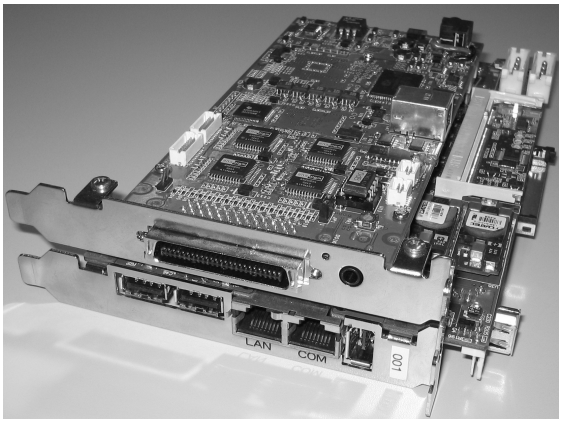

図 5.2: RASP-2 の外観

及び画像処理は，ロボット内部の計算資源を用いて行わなければならない．以上から，音響信号処理及び画像処理を行うハードウェア RASP-2 を開発した．本システムは，この RASP-2，ロボット内部の他の計算資源，及びネットワーク上の利用可能な計算資源を用いて，効率よくリアルタイムに分散処理されるよう実装ている．さらに，実環境での評価・改良を行うため，音声インタフェースを利用したロボットの制御，及びロボットとネットワークを介して接続された情報家電機器の制御を行う簡単なアプリケーションを開発し，これを用いて評価実験を行った．

以下では，まず，アレイ信号処理をリアルタイムで行うために設計，製作した RASP-2 について述べる．続いて，この RASP-2 及び他のデバイスを用いて構築したリアルタイム音声インタフェースシステムについて詳述する．最後に，本システムの評価実験及びその実験結果について述べる．

5.2 リアルタイム音響信号処理装置 RASP-2

本節では，第 4 章で述べた音声インタフェースのうち，アレイ信号処理部分をリアルタイムで処理するために設計，製作したハードウェア RASP-2 について述べる．HRP-2 ではスペースに制約があるため，RASP-2 は PCI ハーフサイズの基盤上に実装されており，2 スロット分のスペースに収まるように設計されている．

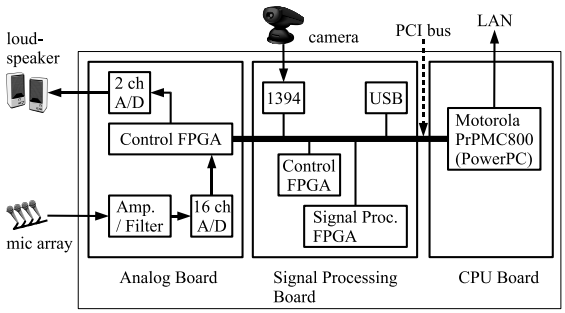

図 5.3: RASP-2 の構成図

さらに，将来的には画像処理も行えるよう IEEE 1394 のインタフェースなどが実装されている．図 5.2 は RASP-2 の外観である．

図 5.3 は RASP-2 の構成図である．RASP-2 は以下の 3 つのボードから構成されている．

- アナログボード

  16 チャネルの A/D コンバータ，2 チャネルの D/A コンバータ及びアンチエリアシング用のフィルタから構成されており，本システムでは 16 チャネルのうち，8 チャネル分を使用している．D/A コンバータはマイクロホンアレイのキャリブレーション時に使用される．

- CPU ボード

  汎用の CPU ボード（PowerPC 450 MHz を搭載した Motorola PrPMC800, メモリは 192 MB）を用いている．汎用の CPU ボードを使用する事で，より高度な CPU ボードが入手可能となった時点で，システムを容易にアップグレードできる利点がある．オペレーティングシステム（OS）としては Monta Vista Linux が稼働しており，音源位置推定や分離フィルタの計算などの複雑な信号処理用のソフトウェアを C 言語で作成する事ができる．

- 信号処理ボード

表 5.1: 0.5 秒のデータに対する RASP-2 の CPU ボードでの演算時間

	時間（秒）	負荷率	FPGA
FFT	0.081	0.16	△
空間相関行列の計算	0.084	0.17	△
FIR フィルタ	0.094	0.19	○
音源位置推定	0.027	0.05	
分離フィルタの計算	0.034	0.07	

RASP-2 のマザーボードとしての機能を持ち，このボードから，HRP-2 の PCI バスに接続する．また，アナログボード及び CPU ボードとは，PMC（PCI Mezzanine Card）接続ポートを介して接続されており，ロボットの PCI バスとは独立なバスを持つ事により， ロボットの PCI バスの負荷に関係なく，大量の音響・画像データのやり取りが行えるようになっている．

このボード上には FPGA（Xilinx Virtex II）が搭載されており，単純ではあるがリアルタイム性の要求される計算量の多い処理を行う事ができる．現時点では FIR フィルタ（1024 タップ ×16 チャネル）が実装されており，現在 FFT，空間相関行列の計算機能を実装している．

RASP-2 の最大の特徴は汎用の CPU と FPGA を組み合わせたハイブリットな構成になっているという点である．FIR フィルタや FFT などは，単純な積和演算で構成され，FPGA を用いて実装する事により比較的容易に並列化する事ができ，処理の高速化が可能である．また，これらは，音源分離のアルゴリズムが多少変化してもアーキテクチャが変化する事はあまりないので，プログラミングのフレキシビリティは少なくてもよい．一方，音源定位や音源分離などでは，固有値演算や逆行列演算を多用するため，64 ビット浮動小数点演算が必要とされ，C 言語で書かれた汎用ライブラリなどを使用する必要がある．また，アルゴリズムなどの改良のため，プログラミングのフレキシビリティも必要である．以上の事を踏まえ，RASP-2 では，FPGA と汎用 CPU のハイブリッド構造になっている．この

事により，フィルタなどの単純な演算はFPGAに，音源定位や音源分離などの複雑な演算は汎用CPUに分散して処理を行う事が可能になり，制約された計算資源を用いてリアルタイム処理を達成できる．表5.1の1列目は0.5秒分のデータに対し，RASP-2のCPUボードで音響信号処理の各部分の演算を行った場合の演算時間を示したものである．また，0.5秒に対する各行の割合を示したものであり，3列目は各音響信号処理のうちFPGAへの実装状況を示したものである．3列目の○は，現時点でFPGAに実際に実装されているもの，△は将来的に実装が予定されているものである．現時点ではこの表のうち，一番負荷の重いFIRフィルタがFPGAで処理されているため，表5.1の全音響信号処理をCPUボードで処理するのに比べ，ある程度負荷が低下している．現在はFIRフィルタのみがFPGAに実装されているが，さらに，FFTや相関の計算もFPGAに移行する事により，CPUボード上にあるPowerPCの負荷を下げる事ができ，より安定なリアルタイム性を確保するともに，新たな演算を行う事も可能になる．

RASP-2はこの他にもインタフェースとして，CFスロット，USB，100 Mbpsの有線LANコネクタ，IEEE 1394を搭載している．Linuxの起動はCFカードかUSBメモリを用いて行う．また，IEEE 1394を搭載しているため，将来的にはカメラを接続し画像データを取得する事も可能である．

5.3 リアルタイムシステムの構築

5.3.1 システムの全体構成

本節では，リアルタイムシステム全体の構成について述べる．図5.4はリアルタイムシステムにおけるハードウェアの構成図である．入力デバイスとしてはディジタルカメラとマイクロホンアレイがあり，それぞれ画像処理用CPUボード及び音響処理用ハードウェアRASP-2に接続されている．画像処理用CPUボード及びRASP-2は，ロボットの胸部にあるPCIスロットに搭載されている．RASP-2と画像処理用CPUボードは，100Mbpsの有線LANにより接続されている．また，

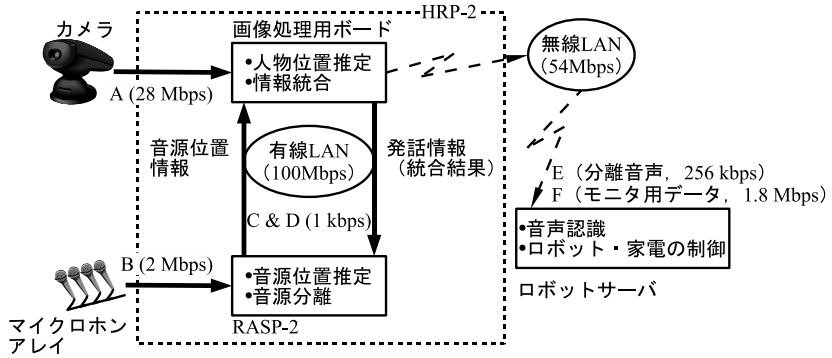

図 5.4: リアルタイムシステムの処理の流れ

HRP-2 の外部には，無線 LAN（IEEE 802.11g）により接続されたロボットサーバ（Pentium 4 3.20 GHz を搭載したパーソナルコンピュータ）があり，ロボットの制御を行う．音声認識は，このロボットサーバに実装されている．

続いて，図 5.4 を用いて，処理の流れを説明する．画像処理用 CPU ボードでは，カメラから取得したデータをもとに，人物位置の検出が行われる．これと並行して RASP-2 では，0.5 秒ごとに音源位置の推定が行われる．音源位置の情報は画像処理用 CPU ボードに送られ，人物位置の情報と共に，ベイジアンネットワークを用いた情報統合に入力され，発話区間及び話者位置が推定される．その後，発話区間及び話者位置の情報は，再び RASP-2 に送られる．RASP-2 では，この情報をもとに音源分離フィルタが更新がされ，更新されたフィルタを用いて音源分離が行われる．音源分離の出力は，さらに，発話区間情報をもとに発話部分だけが切り出され，無線 LAN を介してロボットサーバへ送られる．ロボットサーバでは，音声認識が行われ，その結果に基づいて，ロボット用の制御コマンドや，後述する情報家電制御用のコマンドが発行される．

これらリアルタイムシステムの状態のうち，主要な部分についてはモニターに表示する事ができる．図 5.5 は，システムのモニター画面である．図の (A) はサブスペース法により求められた空間スペクトルである．この例では 0 度方向にピークが存在しており，この方向に音源があると推定される．また，空間スペクトルの

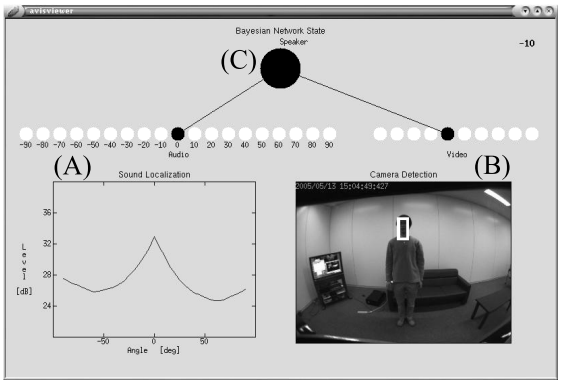

図 5.5: リアルタイムシステムの状態出力のモニター図

上にある 19 個の円はそれぞれの領域に対応した仮想センサの状態を表したものである．円の色が白であれば対応する領域に音源が存在しない事を示し，黒であれば音源が存在する事を示している．先に述べたように，この例では 0 度方向に音源が存在しており，それに対応する仮想センサの状態がアクティブとなっている．

図の (B) は顔の位置の検出結果である．この図において，白い長方形で囲まれた領域が顔が存在していると推定された領域である．また，顔位置検出結果の上にある 10 個の円が仮想センサの状態を表している．この例では，画面の中央に人物が存在し，それに対応した仮想センサの状態がアクティブとなっている．

図の (C) は本システムで用いてるベイジアンネットワークである．図の例では，音源位置と人物位置とが一致しているため，ベイジアンネットワークの出力ノードがアクティブになっている．

図 5.4 には，上述の各モジュール間でのデータ通信量が示されている．入力デバイス直後の通信路である A 及び B でのデータ量は，28 Mbps 及び 2 Mbps と大きいが，分離音声をロボットサーバに送信する際のデータ量は 256 kbps となり，入力データが 1/117 まで削減されているのが分かる．これは，HRP-2 とロボットサーバ間の E では分離音声を 1 チャネル分送信すればよいからである．さらに，画像処理用 CPU ボードと RASP-2 の間の通信量（C 及び D）は 1 kbps と比較的少ない．また，音響情報，画像情報及び情報統合結果をモニターするために，画像処

図 5.6: HRP-2 頭部の外観

理用 CPU ボードから無線 LAN を介して，外部のモニター用 PC へ情報を送れるようになっている．

上述の C，D でのデータ送受信には，RMCP [45] を用いた．RMCP は UDP/IP を用いたブロードキャストによる通信であるため，本論文で示すような複数のハードウェアやソフトウェアモジュールがネットワーク上に分散している場合に効率良く情報を共有する事ができる．

5.3.2 入力機器

本システムには画像データを取得するためのカメラと音響データを取得するためのマイクロホンアレイが存在する．カメラには Point Grey Research 社 Flea を，マイクロホンアレイには携帯電話用の小型な無指向性のコンデンサマイクロホン（プリモ社 EM147TN）を用いた．カメラは HRP-2 頭部の中央部に 1 つ配置されている．マイクロホンは，図 5.6 及び図 5.7 に示すように，HRP-2 頭部側面及び前面に 8 個がコの字型に配置されている．

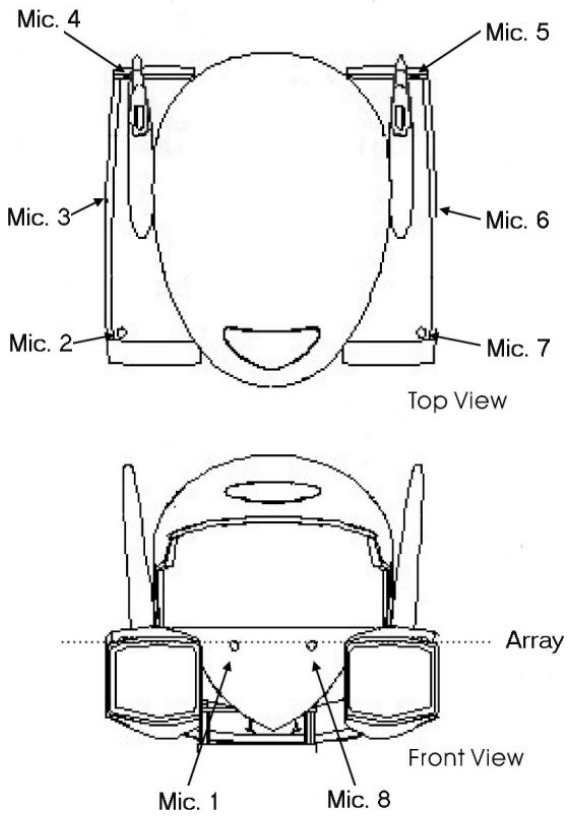

図 5.7: HRP-2 でのマイクロホンの配置図

5.3.3　画像処理用 CPU ボード

5.2 節で述べたように，将来的には人物位置検出などの画像処理は RASP-2 において実現可能であるが，現時点では画像処理用 CPU ボードを用いている．画像処理用 CPU ボードは PCI ハーフサイズの SBC（Single Board Computer）であり，CPU は Pentium III-S 1.40 GHz，メモリは 256 MB である．カメラで観測されたデータは IEEE 1394 を経由して縦 240 ピクセル，横 320 ピクセルのカラー画像として画像処理用 PC に取り込まれる．

カメラから取得された画像はこの PC 内で人物位置検出の処理が行われ，人物位置の特徴ベクトル（仮想センサの出力）が生成される．生成された特徴ベクトルは RMCP を用いて情報統合モジュールへ送信される．処理速度は PC の能力に

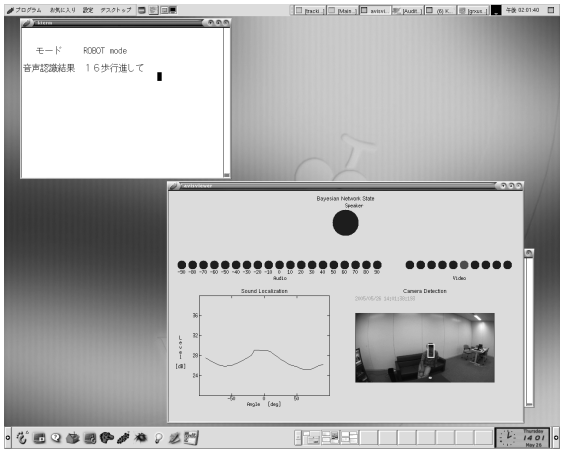

図 5.8: 音声認識結果と状態の表示画面

依存するが，現在の実装では 12〜15 fps である．

5.3.4 ロボットサーバ

ロボットサーバは HRP-2 の外部の計算資源であり，HRP-2 とは無線 LAN で接続されている．ハードウェアは，Pentium 4 3.20 GHz と 1 GB のメモリを搭載したパーソナルコンピュータ（PC）である．ロボットサーバでは，ロボットの制御用コマンドの発行を行う．また，音声認識もこのロボットサーバに実装されている．図 5.8 は音声認識の結果及びシステムの状態表示の画面である．

この他，本システムでは，HRP-2 を情報家電のインタフェースとして用いるため，情報家電を制御するためのモジュールも，実装されている．図 5.9 は，ネットワークリソースの概要を示したものである．ロボットサーバは，有線 LAN を経由してホームサーバ（PC）に接続されている．音声認識モジュールで，情報家電制御用のコマンドが認識された場合，ロボットサーバ上の情報家電制御モジュールに渡され，ここで，情報家電制御用のコマンドが発行される．ホームサーバは，このコマンドを受け取り，これに接続された情報家電機器を制御する．現時点では，液晶テレビ（SHARP AQUOS LC-37GD3）が接続されており，ビデオ機能はホー

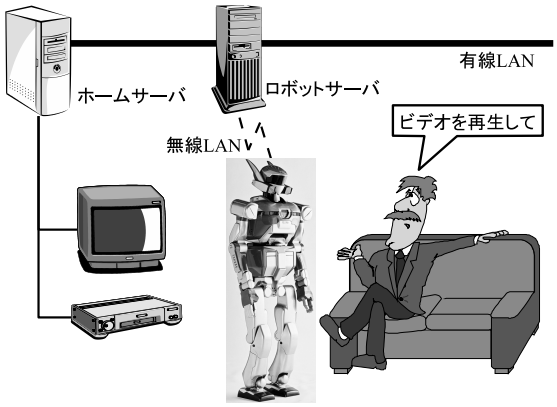

図 5.9: 各種ネットワークリソース

ムサーバ上において Windows Media Player を制御する事で実現した.

本システムではロボットやロボットとネットワークで接続された情報家電などのリソースを制御するためのコマンドを，孤立単語として発話する事を想定している．そのため，音声認識システムにおける言語制約としては，家電制御コマンドを単位としたネットワーク文法を用いている．また，本システムではロボット，ネットワーク接続されたテレビ，ネットワーク接続されたビデオの 3 つの機器を操作できる．音声認識では，操作する機器ごとに認識のモードを分けており，各モードごとに，該当する機器の操作に関する内容を記述したネットワーク文法が用意される．このように，個別の文法と語彙を持つ事で，語彙数を制限し，高雑音下での音声認識精度を上げる工夫がなされている．表 5.2 は，それぞれのモードでのコマンドの例（表 5.2 で列挙しているコマンドが全てではない）である．また，モード名の下の括弧内に，各モードで受理可能なコマンドの総数を示す．それぞれのモードへ移行する場合は,「ロボットの操作」,「テレビの操作」,「ビデオの操作」というキーワードを発話する事により行う.

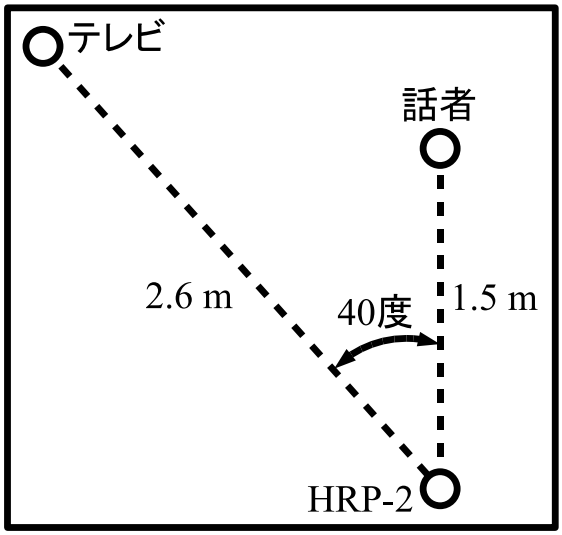

図 5.10: 実験条件

5.4 リアルタイムシステムの評価

5.4.1 実験条件

本節では本論文が想定する環境下において，テレビ，ビデオ，ロボットを制御するコマンドを発話し，実験を行った．実験結果では，発話したコマンドに対するタスクの達成率を用いる事で本システムの雑音環境下でのトータルの性能を評価する．ここでは，まず，実験条件について述べる．

実験は，3 人の成人男性が雑音環境下で，リアルタイムシステムを実装したロボットに対してコマンドを発話する事で行った．図 5.10 は話者，ロボット，テレビの位置関係を示したものである．部屋の残響時間は約 0.5 秒である．また，実際の実験風景は図 5.11 の通りである．

実験に用いたコマンドは表 5.2 の 45 個である．最初の 18 個がテレビモードにおけるコマンドで，テレビのチャネルや音量を変更する．次の 16 個がロボットモードにおけるコマンドで，ロボットに話しかけたり各種動作を指示する．最後の 11 個がビデオモードにおけるコマンドである．

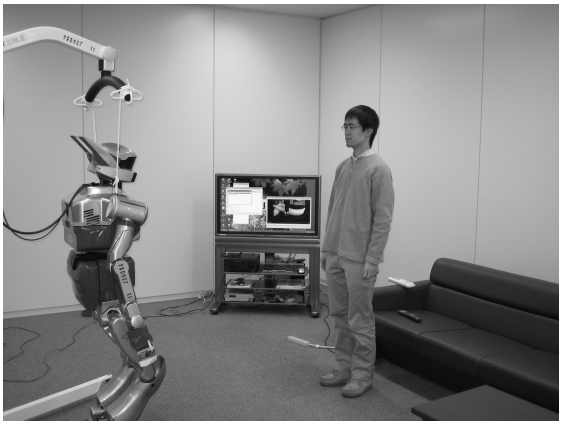

図 5.11: 実験風景

被験者は各コマンドを順番に発話していく．正しく認識されず意図したタスクが実行されなかった場合は，正しく認識されるまでそのコマンドを繰り返し発話する．ただし，同じ内容のコマンドとして認識されれば正しく認識されたものとする．例えば，発話内容が「挨拶をして」である場合に認識結果が「挨拶して」というケースが該当する．

本実験においては，実際のテレビ音声を雑音として用いた．ただし，話者の発話によりテレビの電源を入れるまで（テレビモードの 1 番と 2 番）は，雑音が存在しない．テレビモードの 3 番以降はロボットモードも含めてテレビ音声が雑音として存在している．ビデオモードに入るとテレビの画面が切り替わるため，テレビからの音が一旦途切れる．その後ビデオを再生中はビデオの音声が雑音として存在している．今回の実験では SN 比が概ね 0 dB となるように，事前にテレビのボリュームを調整した．

音声認識システムに用いる音響モデルとしては，連続音声認識コンソーシアムソフトウェア 2003 年度版の PTM（Phonetic Tied Mixture）型 triphone モデル [46] を用いた．実験では，上記の音響モデルに対し，教師ありの事前適応を行った．適応用のデータは本システムで高い頻度で用いる可能性のある 60 個のコマンドを 3 人の話者にそれぞれテレビ音声を雑音とした環境下で発話させ，本システムで音

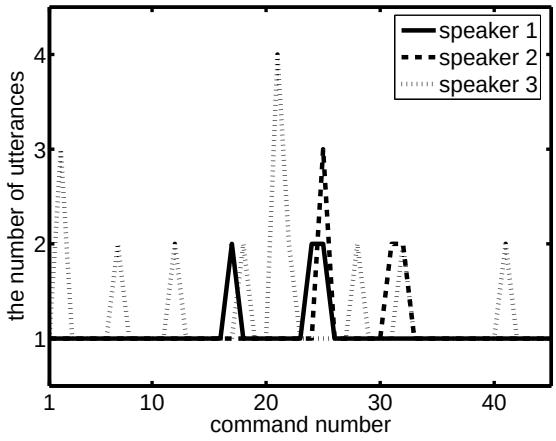

図 5.12: 実験結果

源分離を行い収録した．実験で用いたコマンドは，発話によって制御可能な機能が一通り動作するように選ばれており，テレビモードのコマンドの 89 %が，ロボットモードの 50 %が，ビデオモードの 91 %が適応用のデータに含まれている．

5.4.2 実験結果

実験は各コマンドを何回の発話で実行できたかで評価した．図 5.12 は実験結果を示したものである．図の横軸は各コマンドの通し番号で，横軸の 1 番目から 18 番目が表 5.2 のテレビモードの 1 番から 18 番に対応し，19 番目から 34 番目までがロボットモードの 1 番から 16 番に対応し，35 番目から 45 番目までがビデオモードの 1 番から 11 番に対応する．また，縦軸はそれぞれのコマンドを実行するのに要した発話回数である．結果は 3 人の話者ごとに示している．

複数の話者において認識に失敗したものとしては，ロボットモードの 5，6，7 番において「上げて」を「下げて」と認識するケースや，その逆のケースがある．また，ロボットモードの 13 番や 14 番において数字の認識に失敗したものもある．同様に，テレビモードにおいて数字の認識を誤ったケースも見られた．

また，話者 3 の 21 番目のタスクのように，4 発話もかかっている例もある．こ

れは,「あっちをみて」という発話内容に対し,「終了」と認識してしまい，ロボット制御モードを抜けてしまったため，再度このモードに復帰するのに余分な発話を要した例である．この例からも，対話における誤認識とそのリカバリーなどについては，将来的に改善する必要があると考えられる．

表5.3はタスクの成功率を話者別に示したものである．表において，総発話数は45個のコマンドを実行するのに要した発話回数である．また，タスク成功率1は45個のコマンドのうち，1回の発話で，すなわち言い直しがなしで実行されたものの割合である．タスク成功率2は同様に，1回もしくは2回の発話で実行されたものの割合である．この表より，8割から9割程度のタスクは1回の発話で実行に成功しており，誤認識した場合でも高々1回の言い直しでほぼ完全に実行されているのが分かる．

本実験は実際の環境下で行われたものであるため，S/Nがどの程度改善したかを調べる事は難しい．ただし，今回実験で使用したマイクロホンアレイで測定したインパルス応答を用いて計算機上で測定を行った結果，SN比が0 dBであったものが20 dB程度まで改善した．この事より実際の環境下でもこの程度までの改善が見込まれる．

これらの実験結果より，本システムが想定している環境下で有効に機能している事が示された．

5.5 おわりに

本章では，第4章で提案した音声インタフェースをヒューマノイドロボットHRP-2にリアルタイムシステムとして実装した．この際，新たにリアルタイム音響信号処理装置RASP-2を開発する事で，話者検出及び音源分離までの全ての処理をHRP-2の内部で行う事が可能になった．これにより，マイクロホンアレイ及びカメラからの大量の情報はHRP-2の内部で処理され，無線LANを介してHRP-2の外部に送られるデータが必要最小限に抑えられた．また，本システムを用いる事

で話者の検出及び音源分離の処理が自動化され，システムが動作中はロボットの視野内で発話するだけで，話者の発話を認識可能である．

さらに，本章では構築したリアルタイムシステムを用いて評価実験を行った．実験ではコマンドをロボットに発話して実際にテレビなどの家電を操作し，そのタスクの達成率を調べた．この結果，1 回目の発話で 8 割以上のコマンドが認識された．また，1 回目の発話で認識に失敗したものについても，再度言い直す事でほぼ完全に認識された．この事から，リアルタイムシステムが想定した一般家庭やオフィスなどの環境において，有効に動作する事が検証された．

表 5.2: 各モードでのコマンドの例

モード	番号	発話内容	モード	番号	発話内容
テレビモード（41）	1	テレビの操作	ビデオモード（41）	1	ビデオの操作
	2	電源を入れて		2	３番目を再生
	3	ＮＨＫ総合		3	一時停止
	4	ＮＨＫ教育		4	１番目を再生
	5	日本テレビ		5	ボリュームアップ
	6	ＴＢＳ		6	ボリュームダウン
	7	フジテレビ		7	早送り
	8	テレビ朝日		8	再生
	9	テレビ東京		9	巻き戻し
	10	ボリュームアップ		10	再生
	11	ボリュームダウン		11	停止
	12	１チャンネル			
	13	３チャンネル			
	14	４チャンネル			
	15	６チャンネル			
	16	８チャンネル			
	17	１０チャンネル			
	18	１２チャンネル			
ロボットモード（1056）	1	ロボットの操作			
	2	こんにちは			
	3	あっちを見て			
	4	こっちを向いて			
	5	右手を上げて			
	6	左手を上げて			
	7	左手を下げて			
	8	比留川さんを探して			
	9	金広さんにこれを届けて			
	10	井上先生のところにいって			
	11	ジュースを持ってきて			
	12	テレビのところにいって			
	13	３歩前へ進んで			
	14	４歩後ろへ下がって			
	15	挨拶をして			
	16	さようなら			

表 5.3: タスクの成功率

	総発話数	タスク成功率 1	タスク成功率 2
話者 1	48	0.93	1.0
話者 2	49	0.93	0.98
話者 3	57	0.80	0.96

第6章　結論

本論文は，実環境に対してロバストな音声インタフェースを実現する事を最終目的としている．具体的なシナリオとしては，音声インタフェースに対し話者が発話し，テレビなどの機器を操作する事を想定している．この場合，以下のような条件が考えられる．

- 話者の発話は接話マイクロホンではなく，話者からある程度離れた位置にあるマイクロホンで観測する．
- 話者の他にもテレビなどの雑音源が存在する可能性がある．

このような状況下では特に以下の 3 つの点が問題となる．

- 話者と雑音源が存在する場合，雑音源と区別して話者の発話を検出する事．
- 話者の発話が雑音源に重畳する場合，雑音源から話者の発話を分離する事．
- 音源分離後の音声に対して，頑健に音声認識を行う事．

上述の 3 つの問題を解決し得る既存の技術には，それぞれ，VAD などの話者検出，BSS や ABF などの音源分離，音響モデルに対するモデル適応が存在する．しかし，これらの技術をそれぞれ単独で用いる場合には，様々な問題が存在する．第 1 章では，これらの問題について述べた．また，この問題を克服する手法として，様々な情報や手法を統合する事により，ロバストな音声インタフェースを実現するという，本論文の方向性について述べた．さらに，上述の問題を解決するための具体的な目標として以下の 3 つを設定した．

1. 実環境下における重畳区間の検出.
2. 雑音源が存在する環境下での，目的音源の検出.
3. 上記の手法及びその他の手法を組み合わせた，ロバストな音声インタフェースの実現.

第 2 章では，本論文の第 1 の目標である目的音源と雑音源の重畳区間を検出する手法を提案した．重畳区間の情報は，提案インタフェースの話者検出において，重要な情報である．本論文で想定している環境には，部屋の反射や残響が存在する．既存の手法は，このような環境を想定していないため，うまく重畳区間を検出できない．これに対し，本論文では，実環境下で得られる固有値分布と SVM, SVR を用いて重畳区間を検出する手法を提案した.

実環境下で得られる固有値分布には，音源数に関する情報がある程度反映している．そこで，提案手法では，SVM や SVR を用いて固有値分布の音源数を推定し，重畳区間を検出する．SVM を用いた手法では，固有値分布を 2 つのクラスに分類する事で重畳区間の検出を実現している．また，固有値分布は単に音源数だけではなく，音源間のパワー差によってもその形状が変化する．SVR を用いた手法では，固有値分布から音源間のパワー差を推定する事で，より正確に重畳区間を検出する事を目指している.

実験では，実環境下で収録されたグループインタビューのデータに対して，重畳区間検出を行った．実験結果より，SVM，SVR を用いた手法とも，既存の手法である閾値法に比べ高い検出性能を示した.

第 3 章では，本論文の第 2 の目標である，音響情報と画像情報の統合に基づいた目的音源検出法を提案した．本論文で想定する環境下では，SN 比が 0 dB 程度の悪条件であったり，雑音源に音声が含まれる場合がある．そこで，提案手法では，人物が存在する方向と音源が存在する方向が一致する場合，その方向に話者が存在するという規範で話者の検出を行った．このため，音響情報と画像情報の統合

が必要不可欠であるが，例えば音源の方向は角度で得られ，人物の方向は画像のピクセル値で得られるというように，両者の座標系は直ちに一致しない．また，1対1で対応関係を定義する事は，煩雑であると同時に，これらの情報に曖昧性が存在するという点からも困難である．提案手法では，この問題を解決するために，ベイジアンネットワークを用いて両者の情報を統合している．ベイジアンネットワークを用いる事で，曖昧性を含んだ形で両者の対応関係を表現可能である．

実験では，雑音源が発音し続ける中で二人の話者が交互に発話する様子を提案手法で検出した．実験の結果より，全発話の 98.8 %の区間が提案手法で検出された．この結果より，提案手法が有効に機能している事が示された．

第 4 章では，第 2 章，第 3 章で提案した手法やその他の技術を組み合わせる事で，本論文の第 3 の目標である，実環境に対してロバストな音声インタフェースを提案した．提案インタフェースは話者検出において，音源位置の情報，人物位置の情報，重畳区間の情報をベイジアンネットワークで統合する事により，話者が単独で発話しているか，雑音源に重畳して発話しているか，雑音源が単独で発音しているかを推定する．また，話者検出での推定結果を音源分離における事前情報として用い，これを ML 法に適用する事で高精度の音源分離を実現した．音声認識では，音源分離後の音声を用いてモデル適応を行う事で，さらに認識性能の向上を図った．

実験では，雑音源が存在する実環境下で，雑音源に重畳した話者の発話を提案インタフェースで認識した．実験の結果より，各機能単独では性能が実現されなかった高い認識性能が，複数の機能を組み合わせる事で達成された．これは，各機能の効果が相乗的に表れたものであると考えられる．

第 5 章では，本論文の第 3 の目標及び最終目的の達成度を評価するため，第 4 章で提案したインタフェースをリアルタイムシステムとしてヒューマノイドロボット HRP-2 に実装した．ロボットにこのようなインタフェースを実装する場合，ロボットの外部との通信は帯域幅などに制約があるため，なるべくロボットの内部

で処理を行う事が望まれる．一方で，ロボットの内部は，実装するハードウェアに対し，スペースや電力などの点で著しい制約が存在する．これらの問題に対処するために，新たに音響信号処理用ハードウェアである RASP-2 を開発し，音声認識以外の全ての処理がロボットの内部で行えるよう，システムを構築した．

実験では，実際に話者が発話によってテレビなどを操作し，そのタスク達成率でシステムの評価を行った．実験の結果より，2 回以内の発話によりほぼ全てのコマンドが実行可能であった．これより，本システムは実環境下で有効に機能している事が示された．

一方で，提案インタフェースをさらに人間にとって使いやすいものにしていくための課題も存在する．提案インタフェースは，音源と人物が同一方向に存在する場合，それを話者として検出する．そのため，人物と同一方向に話者とは関係ない音源が存在する場合にも，それを発話として検出する．話者が発話するのではなく手を叩くのはその一例である．これは，話者に発話以外の物音を立ててはならないという制約を与える事になり，不便である．この問題に対しては，例えば，VAD [5] によってさらに音響的な特徴を加味する事で解決が可能であると考えられる．

また，話者の発話であっても，音声インタフェース以外に対して発話を行う場合が考えられる．この場合，音声インタフェースの方に顔を向けずに発話を行っても，人物がインタフェースの視野内に存在すれば，それを発話として検出する．これは，話者に音声インタフェース以外への発話を行ってはならないという制約を与える事になる．この問題に対しては，人物の視線や顔の向きを検出する手法 [19, 47] でこれらの情報を検出し，ベイジアンネットワークの入力に加える事で対処が可能であると考えられる．

提案インタフェースでは，話者が発話中は話者，雑音源，マイクロホンアレイの位置が変化しない事を仮定している．音声インタフェースが備え付けの機器や壁などに設置されている場合は，この仮定が成立する事が期待される．しかし，ロ

ボットに音声インタフェースが実装される場合は，この限りではない．例えば，ロボットが歩行中にロボットに対して何らかの指示を出したい，という事が考えられるからである．この場合，先の仮定は成立しない．この問題を解決する手法として，音響・画像情報を用いてパーティクルフィルタで話者を追跡する手法 [48, 49] や，EM アルゴリズムに基づく，マイクロホンアレイを用いた移動音源の追跡と分離 [50] などの研究が進められている．

以上より，本研究では情報統合の考え方に基づいて，実環境でロバストに動作する音声インタフェースの構築を目標として，研究を行った．上述のように，この目標はほぼ達成されたと考えているが，提案した手法が日常生活の中で使われるレベルになるまでには，以前としてまだ多くの課題が存在する事も明らかとなった．今後の研究により，これらの問題が一つ一つ解決され，音声インタフェースが真に実社会に貢献できる日が訪れる事に期待したい．この過程で，本研究の成果がその一助となれば幸いである．

謝辞

本研究を進めていく上で，また本論文を執筆する上で数多くの方々に御指導，御教授を賜りました．ここに，特にお世話になった方々を以下に列記させて頂き，深い感謝の意を表します．

本学大学院システム情報工学研究科 北脇信彦 教授には指導教官として研究全般に渡り貴重な御意見や数多くの御配慮を賜りました．産業技術総合研究所情報技術研究部門及び本学システム情報工学研究科連携大学院助教授 浅野太 博士には，本研究全般に渡って熱心な御指導，御鞭撻を頂き，また，本研究の核となる部分について数多く示唆して頂きました．本学大学院システム情報工学研究科 山田武志 助教授には，本論文をまとめるにあたり温かい激励と御配慮を頂きました．本学大学院システム情報工学研究科 椎名毅 教授，並びに本学大学院システム情報工学研究科 水谷孝一 教授には，本研究に関しまして的確な御助言を頂きました．

産業技術総合研究所情報技術研究部門 麻生英樹 氏には，本研究における重要な要素である情報統合やベイジアンネットワークに関して数多くの御教授を頂きました．産業技術総合研究所情報技術研究部門 原功 博士には，ヒューマノイドロボット HRP-2 にリアルタイムシステムを実装する際に，様々な御支援を頂きました．産業技術総合研究所情報技術研究部門 緒方淳 博士には，音声インタフェースの音声認識の部分に関して，数多くの御支援を頂きました．産業技術総合研究所情報技術研究部門 後藤真孝 博士は，リアルタイムシステムのデータ送受信プロトコルとして用いた RMCP の開発者であり，RMCP を用いてシステムを実装する際に御支援を頂きました．産業技術総合研究所情報技術研究部門 吉村隆 氏には，音声インタフェースの評価実験に際して，多大な御協力を頂きました．また，産

業技術総合研究所情報技術研究部門メディアインタラクショングループの皆様には公私に渡り様々な温かい御支援を頂きました.

本学システム情報工学研究科マルチメディア研究室の皆様には，研究を進めるにあたって様々な御支援を頂きました．また，博士課程の在籍に御理解を頂いた両親，全国に散らばる友人から様々な激励，御支援を頂きました．本研究はこのような方々の温かい御支援，激励なくしては，成し得なかったものです.

末筆ながら，学会，研究会などを通じて数多くの先輩諸氏の方々に御意見，激励を頂きました事に，厚く御礼申し上げます.

参考文献

[1] Steven F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction”, IEEE Transactions on Acoustics, Speech, and Signal Processing, **27**, 2, pp. 113–120 (1979).

[2] T. W. Lee, Independent Component Analysis, Kluwer Academic Publisher (1998).

[3] Hiroshi Sawada, Ryo Mukai, Shoko Araki and Shoji Makino, “Real-time blind extraction of dominant target sources from many background interferences”, Porceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC 2005), pp. 73–76 (2005).

[4] Don H. Johnson and Dan E. Dudgeon, Array Signal Processing, Prentice Hall (1993).

[5] Virginie Gilg, Christophe Beaugeant, Martin Schönle and Bernt Andrassy, “Methodology for the design of a robust voice activity detectory for speech enhancement”, Porceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC 2003), pp. 131–134 (2003).

[6] K. Srinivasan and A. Gersho, “Voice activity detection for cellular networks”, Proceedings of IEEE Speech Coding Workshop, pp. 85–86 (1993).

[7] 新美康永，音声認識, 共立出版株式会社 (1979).

[8] J. D. Hoyt and H. Wechsler, “Detection of human speech in structured noise”, Proceedings of 1994 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 1994), Vol. 2, pp. 237–240 (1994).

[9] S. E. Bou-Ghazale and K. Assaleh, “A robust endpoint detection of speech for noisy environments with application to automatic speech recognition”, Proceedings of 2002 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2002), Vol. 4, pp. 3808–3811 (2002).

[10] 吉村隆, 浅野太, 麻生英樹, 北脇信彦, “高次統計量の分布モデルを用いた音声・環境音識別法の検討”, 音声言語情報処理研究会, 情報処理学会 (2005).

[11] R. O. Schmidt, “Multiple emitter location and signal parameter estimation”, RADC Spectral Estimation Workshop, pp. 243–258 (1979).

[12] Osamu Hoshuyama, Akihiko Sugiyama and Akihiro Hirano, “A Robust Adaptive Beamformer for Microphone Arrays with a Blocking Matrix Using Constrained Adaptive Filters”, IEEE Transactions on Signal Processing, **47**, 10, pp. 2677–2684 (1999).

[13] 近藤啓介, 長井隆行, 金子正秀, 樽松明, “マイクロホンアレーを用いた話者位置推定による車載音声認識”, 電子情報通信学会論文誌, **J85-D-II**, 7, pp. 1176–1187 (2002).

[14] 神沼充伸, 斉藤大介, 猿渡洋, 西川剛樹, 李晃伸, “車室内音声入力系における雑音除去の検討”, 日本音響学会春季講演論文集, pp. 819–822 (2005).

[15] http://www.honda.co.jp/ASIMO/

[16] http://www.kawada.co.jp/ams/promet/

[17] http://www.incx.nec.co.jp/robot/

[18] 原直, 西野隆典, 伊藤克亘, 宮島千代美, 武田一哉, “コミュニケーションロボット・DAGANE”, 第 22 回 AI チャレンジ研究会, pp. 47–52 (2005).

[19] 松坂要佐, 東條剛史, 小林哲則, “グループ会話に参与する対話ロボットの構築”, 電子情報通信学会論文誌, **J84-D-II**, 6, pp. 898–908 (2001).

[20] 佐藤幹, 杉山昭彦, 大中慎一, “パーソナルロボット PaPeRo における近接話者方向推定と 2 マイク音声強調”, 第 22 回 AI チャレンジ研究会, pp. 41–46 (2005).

[21] 鈴木薫, 古賀敏之, 廣川潤子, 小川秀樹, 松日楽信人, “ハフ変換を用いた音源音のクラスタリングとロボット用聴覚への応用”, 第 22 回 AI チャレンジ研究会, pp. 53–58 (2005).

[22] 戸上真人, 天野明雄, 新庄広, 鴨志田亮太, 玉本淳一, 柄川索, “人間共生ロボット “EMIEW” の聴覚機能”, 第 22 回 AI チャレンジ研究会, pp. 59–64 (2005).

[23] Mati Wax and Thomas Kailath, “Detection of signals by information theoretic criteria”, IEEE Transactions on Acoustics, Speech, and Signal Processing, **33**, pp. 387–392 (1985).

[24] Richard Roy and Thomas Kailath, “ESPRIT - Estimation of signal parameters via rotational invariance techniques”, IEEE Transactions on Acoustics, Speech, and Signal Processing, **37**, 7, pp. 984–995 (1989).

[25] Futoshi Asano, Shiro Ikeda, Michiaki Ogawa, Hideki Asoh and Nobuhiko Kitawaki, “A Combined Approach of Array Processing and Independent Component Analysis for Blind Separation of Acoustic Signals”, Proceedings of ICASSP 2001, MULT-P2 (2001).

[26] 山下浩, 田中茂, “サポートベクターマシンとその応用”, 第 13 回日本 OR 学会 RAMP シンポジウム論文集 (2001).

[27] Bernhard Scholkopf, Christopher J. C. Burges and Alexander J. Smola, Advances in Kernel Methods: Support Vector Learning, MIT Press (1999).

[28] Nello Christianini and John Shawe-Taylor, An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press (2000).

[29] C. J. van Rijsbergen, Information retrieval (second edition), Butterworths (1979).

[30] T. Chaodhury, J. M. Rehg, V. Pavlovic and A. Pentland, "Boosted learning in dynamic Bayesian networks for multimodal detection", Proceedings of the Fifth International Conference on Information Fusion (Fusion 2002), Vol. 1, pp. 550–556 (2002).

[31] Futoshi Asano, Youichi Motomura, Hideki Asoh, Takashi Yoshimura, Naoyuki Ichimura and Satoshi Nakamura, "Fusion of Audio and Video Information for Detecting Speech Event", Proceedings of the 6th International Conference on Information Fusion (Fusion 2003), pp. 386–393 (2003).

[32] Isao Hara, Futoshi Asano, Hideki Asoh, Jun Ogata, Naoyuki Ichimura, Yoshihiro Kawai, Fumio Kanehiro, Hirohisa Hirukawa and Kiyoshi Yamamoto, "Robust Speech Interface Based on Audio and Video Information Fusion for Humanoid HRP-2", Proceedings of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004), pp. 2404–2410 (2004).

[33] Dorin Comaniciu, Visvanathan Ramesh and Peter Meer, "Kernel-based object tracking", IEEE Transactions on Pattern Analysis Machine Intelligence, **25**, 5, pp. 564–575 (2003).

[34] Finn V. Jensen, Bayesian Networks and Decision Graphs, Springer (2001).

[35] Kazuyo Tanaka, Satoru Hayamizu and Kozo Ohta, “The ETL Speech Database for Speech Analysis and Recognition Research”, Proceedings of First International Conference on Spoken Language Processing (ICSLP 90), pp. 1101–1104 (1990).

[36] http://www.aist.go.jp/RIODB/db066/

[37] 後藤真孝, 橋口博樹, 西村拓一, 岡隆一, “RWC 研究用音楽データベース: クラシック音楽データベースとジャズ音楽データベース”, 2002-MUS-44-5, 情報処理学会音楽情報科学研究会研究報告 (2002).

[38] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura and Ryuichi Oka, “RWC Music Database: Popular, Classical, and Jazz Music Databases”, Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002), pp. 287–288 (2002).

[39] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models”, Computer Speech and Language, **9**, 2, pp. 171–185 (1995).

[40] J. L. Gauvain and Chin Hui Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains”, IEEE Transactions on Speech and Audio Processing, **2**, 2, pp. 291–298 (1994).

[41] E. Thelen, X. Aubert and P. Beyerlein, “Speaker Adaptation in the Philips System for Large Vocabulary Continuous Speech Recognition”, Proceedings of ICASSP ’97, pp. 1035–1038 (1997).

[42] Jun Ogata and Yasuo Ariki, “Unsupervised Acoustic Model Adaptation Based on Phoneme Error Minimization”, Proceedings of ICSLP 2002, Vol. II, pp. 1429–1432 (2002).

[43] Steve Young, Gunnar Evermann, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev and Phil Woodland, The HTK Book, Version 3.2, Cambridge University Engineering Department (2002).

[44] 河原達也, 住吉貴志, 李晃伸, 坂野秀樹, 武田一哉, 三村正人, 伊藤, 克亘, 伊藤彰則, 鹿野清宏, “連続音声認識コンソーシアム 2002 年度版ソフトウエアの概要”, SLP-48-1, 情報処理学会音声言語情報処理研究会報告 (2002).

[45] M. Goto, R. Neyama and Y. Muraoka, “RMCP: Remote Music Control Protocol — Design and Applications —”, Proceedings of the 1997 International Computer Music Conference, pp. 446–449 (1997).

[46] 河原達也, 武田一哉, 伊藤克亘, 李晃伸, 鹿野清宏, 山田篤, “連続音声認識コンソーシアムの活動報告及び最終版ソフトウェアの概要”, SP2003-169, NLC2003-106 (SLP-49-57), 電子情報通信学会技術研究報告 (2003).

[47] 松本吉央, 小笠原司, Alexander Zelinsky, “リアルタイム視線検出・動作認識システムの開発”, PRMU99-151, 電子情報通信学会技術研究報告 (1999).

[48] 浅野太, 麻生英樹, “マイクロホンアレイを用いた移動音源の追跡と分離について”, AI チャレンジ研究会, 人工知能学会研究会 (2004).

[49] Hideki Asoh, Isao Hara, Futoshi Asano and Kiyoshi Yamamoto, “Tracking Human Speech Events Using a Particle Filter”, Proceedings of 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), MSP-P2 (2005).

[50] Futoshi Asano and Hideki Asoh, “Sound Source Localization and Separation Based on the EM Algorithm”, Proceedings of 2004 ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA 2004) (2004).

付 録A　本論文中で用いたマイクロホンアレイ

表 A.1 は本論文の各章の実験で用いたマイクロホンアレイを示したものである．表中のタイプ A のマイクロホンアレイは，直径が 0.2 m の円形アレイであり，マイクロホンが円周上に均等に配置されている．図 A.1 はタイプ A のマイクロホンの配置図であり，その外観は図 2.45 の通りである．タイプ B のマイクロホンアレイは，直径が 0.5 m の円形アレイであり，タイプ A と同様にマイクロホンが円周上に均等に配置されている．図 A.2 はタイプ B のマイクロホンの配置図であり，その外観は図 3.17 の通りである．また，タイプ C のマイクロホンアレイは，ヒューマノイドロボット HRP-2 の頭部にマイクロホンを配置したものである．図 A.3 はタイプ B のマイクロホンの配置図であり，その外観は図 4.5 の通りである．ただし，図 A.3 で示されているマイク間などの寸法は実測値である．

表 A.1: 本論文で用いたマイクロホンアレイのサイズ

	マイクロホンアレイの種類（直径）
第 2 章	タイプ A（0.2 m）
第 3 章	タイプ B（0.5 m）
第 4 章（4.3）	タイプ C
第 4 章（4.4）	タイプ B（0.5 m）
第 5 章	タイプ C

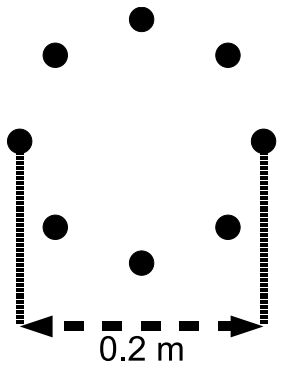

図 A.1: タイプ A のマイクロホン配置

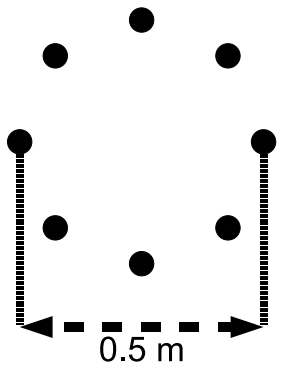

図 A.2: タイプ B のマイクロホン配置

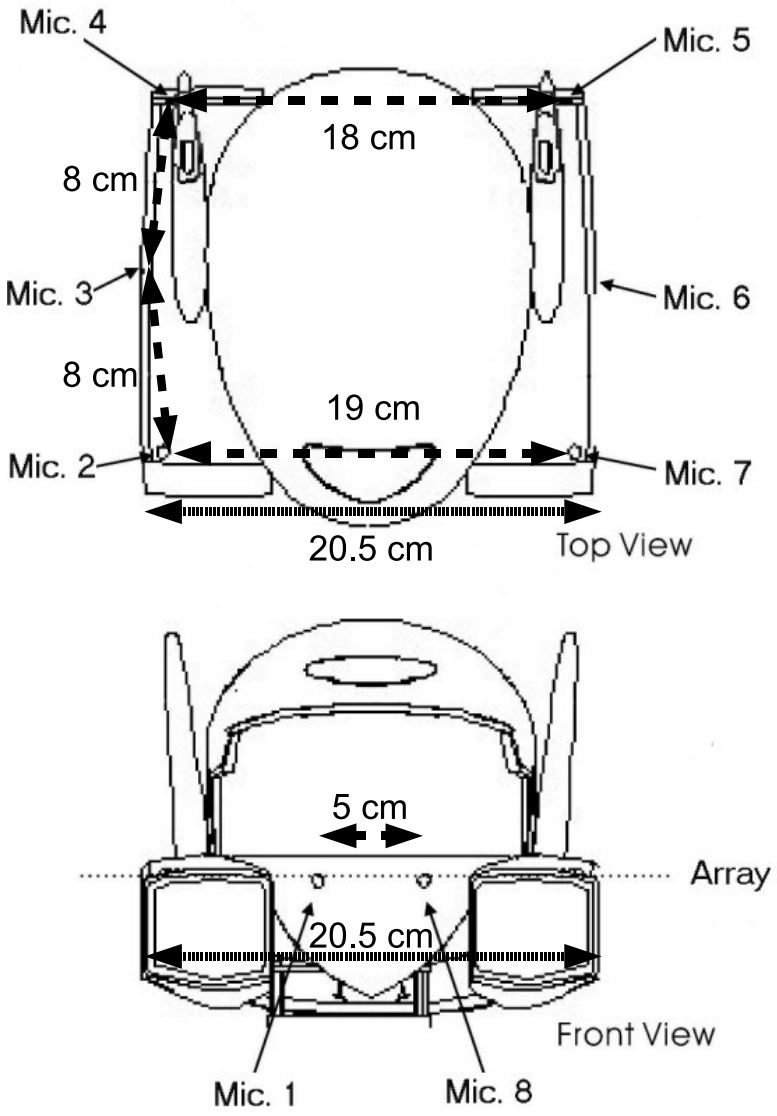

図 A.3: タイプ C のマイクロホン配置

付 録B　研究論文リスト

★は本論文の内容に関するものである.

査読のある学術雑誌に掲載のフルペーパー

★ Futoshi Asano, Kiyoshi Yamamoto, Isao Hara, Jun Ogata, Takashi Yoshimura, Yoichi Motomura, Naoyuki Ichimura, Hideki Asoh, “Detection and Separation of Speech Event Using Audio and Video Information Fusion and Its Application to Robust Speech Interface”, Eurasip Journal on Applied Signal Processing (JASP), Vol. 2004, pp.1727–1738, 2004.

★ 山本潔，浅野太，原功，緒方淳，麻生英樹，山田武志，北脇信彦, “ヒューマノイドロボット HRP-2 における音響情報と画像情報を統合したリアルタイム音声インタフェース”, 日本音響学会論文誌, Vol. 62, No. 3, pp.161–172, 2006 年 3 月.

★ Kiyoshi YAMAMOTO, Futoshi ASANO, Takeshi YAMADA, Nobuhiko KITAWAKI, “Detection Of Overlapping Speech In Meetings Using Support Vector Machines And Support Vector”, IEICE Transactions on Fundamentals, accepted for publication.

査読のある国際会議録に掲載のフルペーパー

★ Kiyoshi YAMAMOTO, Futoshi ASANO, W.F.G. van ROOIJEN, E.Y. LING, Takeshi YAMADA, and Nobuhiko KITAWAKI, “ESTIMATION OF THE NUMBER OF SOUND SOURCES AND ITS APPLICATION TO SOUND SOURCE SEPARATION”, 2003 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2003), pp.485–488, 2003.

★ Futoshi ASANO, Yoichi MOTOMURA, Hideki ASOH, Takashi YOSHIMURA, Naoyuki ICHIMURA, Kiyoshi YAMAMOTO, Nobuhiko KITAWAKI, and Satoshi NAKAMURA, “Detection and Separation of Speech Segment using Audio and Video Information Fusion”, 8th European Conference on Speech Communication and Technology (Eurospeech 2003), pp.2257–2260, 2003.

★ Takashi YOSHIMURA, Futoshi ASANO, Youichi MOTOMURA, Hideki ASOH, Naoyuki ICHIMURA, Kiyoshi YAMAMOTO, Satoshi NAKAMURA, “Detection of Speech Events in Real Environments Through Fusion of Audio and Video Information using Bayesian Networks”, 2003 International Workshop on Acoustic Echo and Noise Control (IWAENC 2003), pp.319–322, 2003.

★ Futoshi Asano, Kiyoshi Yamamoto, Hideki Asoh, Takashi Yoshimura, Isao Hara, Yoichi Motomura, Naoyuki Ichimura, Jun Ogata, “INFORMATION FUSION OF MICROPHONE ARRAY AND CAMERA ARRAY FOR ROBUST SPEECH INTERFACE”, 18th International Congress on Acoustics (ICA 2004), pp.2747–2750, 2004.

Hideki Asoh, Futoshi Asano, Kiyoshi Yamamoto, Takashi Yoshimura, Youichi Motomura, Naoyuki Ichimura, Isao Hara, Jun Ogata, “An application of a particle filter to Bayesian multiple sound source tracking with audio and

video information fusion", 7th International Conference on Information Fusion (Fusion 2004), pp.805–812, 2004.

★ Kiyoshi Yamamoto, Futoshi Asano, Isao Hara, Jun Ogata, Masataka Goto, Hiromitsu Furukawa, Tsutomu Kamashima, Nobuhiko Kitawaki, "Real-time Implementation and Evaluation of Speech Event Detection and Separation Based on the Fusion of Audio and Video Information", 2004 Global Signal Processing Expo. (GSPx 2004), in CD-ROM, 2004.

★ Isao Hara, Futoshi Asano, Hideki Asoh, Jun Ogata, Naoyuki Ichimura, Yoshihiro Kawai, Fumio Kanehiro, Hirohisa Hirukawa, Kiyoshi Yamamoto, "Robust Speech Interface Based on Audio and Video Information Fusion for Humanoid HRP-2", 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004), pp.2404–2410, 2004.

Hideki Asoh, Isao Hara, Futoshi Asano, Kiyoshi Yamamoto, "Tracking Human Speech Events Using a Particle Filter", 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), pp.1153–1156, 2005.

Michiaki KATOH, Kiyoshi YAMAMOTO, Jun OGATA, Takashi YOSHIMURA, Futoshi ASANO, Hideki ASOH, Nobuhiko KITAWAKI, "State Estimation of Meetings by Information Fusion using Bayesian Network", 9th European Conference on Speech Communication and Technology (Eurospeech 2005), pp.113–116, 2005.

★ Kiyoshi Yamamoto, Futoshi Asano, Takeshi Yamada, Nobuhiko Kitawaki, "Detection of Overlapping Speech in Meetings Using Support Vector Regression", 2005 International Workshop on Acoustic Echo and Noise Control (IWAENC 2005), pp.37–40, 2005.

その他

研究会報告

山本潔，浅野太，山田武志，北脇信彦, “音源分離における SVM を用いた音源数推定法について”, 電子情報通信学会技術研究報告, pp.19–24, 2002.

W.F.G. van Rooijen，E.Y. Ling，浅野太，山本潔，北脇信彦, “SVM を用いた音源数推定の音源分離システムへの応用”, 電子情報通信学会技術研究報告, pp.25–30, 2002.

吉村隆，浅野太，本村陽一，麻生英樹，市村直幸，山本潔，中村哲, “実環境における発話区間検出のための音響情報と画像情報の統合”, 電子情報通信学会技術研究報告, pp.13–18, 2003.

浅野太，麻生英樹，原功，吉村隆，緒方淳，市村直幸，本村陽一，後藤真孝，山本潔, “音響と画像の情報統合を用いた話者追跡と音源分離”, 人工知能学会第 18 回 AI チャレンジ研究会, pp.19–26, 2003.

麻生英樹，本村陽一，吉村隆，山本潔，市村直幸，原功，浅野太, “パーティクルフィルタを用いた複数話者の位置と発話状態の追跡”, 2003 年ベイジアンネットワークセミナー（BN 2003）, pp.93–100, 2003.

原功，浅野太，麻生英樹，緒方淳，比留川博久，金広文男，山本潔, “ヒューマノイドロボット HRP-2 におけるロバスト音声インターフェース”, 人工知能学会第 22 回 AI チャレンジ研究会, 2005, 発表予定.

学会講演

山本潔，浅野太，山田武志，北脇信彦, “ブラインド信号分離における音源数推定法について”, 日本音響学会 2002 年春季研究発表会, 3-4-11, pp.623–624, 2002.

山本潔，W.F.G. van Rooijen，E.Y. Ling，浅野太，山田武志，北脇信彦, “SVM を用いた音源数推定法の音源分離システムへの応用”, 日本音響学会 2002 年秋季研究発表会, 2-5-10, pp.537–538, 2002.

山本潔，浅野太，吉村隆，本村陽一，麻生英樹，原功，市村直幸，緒方淳，北脇信彦, “音響情報と画像情報の統合による発話区間検出・分離システムの評価”, 日本音響学会 2003 年秋季研究発表会, 3-6-10, pp.121–122, 2003.

麻生英樹，浅野太，山本潔, “パーティクルフィルタを用いた人発話の追跡”, 第 5 回計測自動制御学会システムインテグレーション部門講演会（SI 2004）, pp.338–339, 2004.

原功，浅野太，麻生英樹，緒方淳，比留川博久，金広文男，山本潔, “ロバスト音声インターフェースを用いたヒューマノイドロボット HRP-2 の聴覚機能の実現”, 第 23 回日本ロボット学会学術講演会, in CD-ROM, 2005.

山本潔，浅野太，原功，緒方淳，麻生英樹，山田武志，北脇信彦, “音響・画像情報の統合によるヒューマノイドロボット HRP-2 の音声インタフェース”, 日本音響学会 2005 年秋季研究発表会, 1-7-22, pp.35–36, 2005.

山本潔，浅野太，山田武志，北脇信彦, “Support Vector Regression を用いた会議音声における発話の重畳区間の検出”, 日本音響学会 2005 年秋季研究発表会, 3-Q-22, pp.691–692, 2005.