音響伝達特性の識別に基づくシングルチャネル音源位置推定の研究

髙島, 遼一

(Degree)
博士（工学）

(Date of Degree)
2013-03-25

(Date of Publication)
2013-05-07

(Resource Type)
doctoral thesis

(Report Number)
甲5786

(URL)
https://hdl.handle.net/20.500.14094/D1005786

博 士 論 文

音響伝達特性の識別に基づく
シングルチャネル音源位置推定の研究

平成 25 年 1 月

神戸大学大学院 システム情報学研究科

高島　遼一

音響伝達特性の識別に基づく
シングルチャネル音源位置推定の研究

要 旨

目的音声がどの位置から発せられているのかを推定する音源位置推定の技術は，雑音の抑圧や目的音の強調といったロバストネス向上のための前処理や，話者の推定，コミュニケーションロボットといった，会話状況の情報の豊富化など，様々な場面で応用されている重要な技術の一つである．従来の音源位置推定法の多くは，マイクロホンアレーと呼ばれる複数のマイクロホンを用いて，各観測信号間の時間差や音圧差などの情報を利用するものがほとんどである．一方，単一マイクロホンのみで音源位置を推定する方法論は未だ確立されておらず，単一マイクロホンによる音源位置推定は非常に困難な試みとされていた．単一マイクロホンで音源位置推定を行うことができれば，コストの削減やシステムの縮小化が可能となる他，従来のマイクロホンアレーシステムと組み合わせることで，システム全体をより強化することも可能であると期待できる．

本研究では，音声の持つ音響伝達特性が音源位置に依存する点に着目し，音響伝達特性を用いることで音源の位置や方向を単一マイクロホンのみで推定する手法について提案する．ここで，マイクによって観測される信号は，原音声（クリーン音声）と音響伝達特性（インパルス応答）が畳み込まれた残響信号である．そのため提案するシステムでは，残響信号からクリーン音声成分を取り除き，音響伝達特性成分のみをブラインドで推定する必要がある．本手法では，あらかじめ話者のクリーン音声を Gaussian Mixture Model (GMM) あるいは Hidden Markov Model (HMM) でモデル化しておき，これらをクリーン音声成分の事前知識として活用することで，観測信号から音響伝達特性を確率的に推定する．そして，推定された音響伝達特性を用いて音源の位置や方向を推定する．

第 1 章では，序論として，音源の位置が音声研究に活用されている例を挙げ，研究の背景及び本研究の位置づけについて述べる．

第 2 章では，本研究で用いる手法を理解する上で必要となる音声信号処理の技術について解説し，従来の複数マイクロホンを用いた音源位置・方向推定手法について解説する．

第 3 章では，音響伝達特性を用いた音源方向推定のアプローチの一つとして，マイクを中心に回転するパラボラ型反射板を用いて，反射板の回転による音響伝達特性の変動を検出することで，音源の方向を推定する手法を提案する．この手法は，人間の聴覚システムに倣った音源方向推定手法である．従来の音源方向推定の枠組みにおいて，二つのマイクでは正中面上の仰角方向の推定はできないが，人間は二つの耳でも耳介（外耳）によって仰角方向の推定が行える点から，耳介のような反射板と，それによって変動する音響伝達

特性が，単一マイクロホンによる音源方向推定に有効と考えられる．また，人間は音源の方向に耳を向けるというように，聴覚システムが能動的である点にも着目して，この章では，反射板が回転することで音源の方向を推定する，アクティブマイクロホンを提案する．焦点位置に単一マイクを備え付けた，パラボラ型の反射板が回転することで，反射板が音源方向を向いたときのみ音響伝達特性が大きく変動する．この手法では，観測信号の音響伝達特性をクリーン音声 GMM により推定し，音響伝達特性が大きく変動したときの反射板の角度を検出することで，音源方向を推定する．実環境下における音源方向推定実験を行い，提案手法の有効性を示す．

第 4 章では，第 3 章で提案するシステムとは別のアプローチとして，部屋の音響伝達特性を識別することで，音源の位置を推定する手法を提案する．口から発せられた直後のクリーン音声は，壁からの反射音（残響）やスペクトルの減衰などの，部屋音響伝達特性（インパルス応答）が畳み込まれた残響信号としてマイクに観測される．この音響伝達特性は，音源の位置に依存した特性を持つことが知られている．この手法では，音響伝達特性が音源位置に依存する点に着目し，音響伝達特性をあらかじめ音源位置毎に学習しておき，評価音声についても，その音響伝達特性を識別することで，音源位置を単一マイクで推定する手法を提案する．さらに，第 3 章では音響伝達特性をクリーン音声 GMM を用いて推定していたのに対して，本章では，より正確に音響伝達特性を推定するために，クリーン音声を HMM でモデル化し，GMM による推定よりも音源位置推定精度が向上することを実験により示す．また，音響伝達特性の特徴量ベクトルである Mel-Frequency Cepstral Coefficients (MFCC) の各次元の中には，音源位置の識別に有効な次元とそうでない次元があり，かつその次元は音源位置によって異なるという考えに基づいて，Multiple Kernel Learning (MKL) を用いた音源位置毎の次元重みを学習させる手法を提案する．音源位置の識別実験において，MKL を組み合わせた Support Vector Machine (SVM) と通常の SVM を比較し，提案手法の有効性を示す．

第 5 章では，第 4 章で提案したシングルチャネル音源位置推定の枠組みを元に，音源が複数ある場合の位置推定問題への拡張手法を提案する．この手法では，あらかじめ音源位置毎に音響伝達特性のモデルを Single Gaussian Model (SGM) で学習しておく．そしてこの音響伝達特性 SGM を各話者のクリーン音声 GMM と合成することで，各話者のそれぞれの位置における，残響音声の GMM が作成される．そして合成された各話者の残響音声 GMM を，さらに話者同士で合成することにより，複数話者がそれぞれの位置から発話したときの混合音声信号の GMM が作成される．これを，あらゆる位置の組み合わせについて行い，各位置の組み合わせにおける混合音声 GMM を作成する．位置推定の際には，複数話者が発話して収録される混合音声信号について，各混合音声 GMM の尤度を算出し，最も高い尤度を示す混合音声 GMM の位置の組み合わせを出力する．本来ならば，混合音声 GMM を学習するためには，想定されるすべての位置の組み合わせ

について，実際に各話者が同時に発話し，その混合音声を収録する必要がある．しかし，モデル合成を用いることで，同時に発話せずとも混合音声 GMM を作成することができ，学習時におけるユーザの負担を減らすことができる．実験ではモデル合成を行う提案手法と，行わない手法を比較し，特に位置の学習発話数が少ない場合において提案手法が高い性能を示した．

第 6 章では，第 4 章で提案した枠組みを用いて，話者の頭部の方向を推定する手法を提案する．音源の位置は「どこから」発話しているのかを表しているのに対して，話者の頭部方向は「どこへ向かって」発話しているのかを表すことになる．この話者の頭部方向の推定は，従来のマイクロホンアレーのシステムでも困難とされているが，これが行えれば，例えば従来では「誰が」話しているのかしか分析できなかったシステムが，「誰に向かって」話しているのかまで分析が可能となる．本章で提案するシステムでは，音響伝達特性が音源の位置だけではなく，話者の頭部方向にも依存する点を利用して，第 4 章で提案した枠組みと同様に，話者の頭部方向毎に音響伝達特性を学習し，識別することで話者の頭部方向をシングルチャネルで推定する．

最後に第 7 章にて，全体を通してのまとめと，今後の課題について述べる．

目次

図目次

表目次

第 1 章

序論

音声は，人間にとって最も身近な情報のひとつであり，人間同士のコミュニケーションをはじめ，メディアによる情報の伝達など，多くの場面で利用されている情報である．ここで，我々は音声から「何と言っているのか」を表す言語情報だけではなく，「誰が」，「どのような気持ちで」，「どのような環境下で」話をしているのか，すなわち話者・感情・環境といった多くの情報を読み取ることができる．そしてこれらの情報は，特に人間同士のコミュニケーションを円滑に進める上で重要な役割を担っている．

音声から読み取れる多くの情報の中には，「どこから話されているのか」を表す音源位置・方向の情報も含まれる．人間はこの位置や方向の情報をもとに，話者の区別や，話者の位置の探索を行うことができる．また，目的の方向からの音声だけを意識して聞こうとすることで，雑音の多い環境下でも目的の音声を認識することができる．現在行われている音声の研究の中には，この音源位置・方向の情報を用いて音声の情報の豊富化や，音声インターフェースのロバストネスの向上を行う研究が多く存在する．

音声情報の豊富化に関しては，音源位置の情報を用いることで発話者を特定し，議事録の作成など，会話状況の詳細な分析を行う手法が提案されている [1, 2]．また，音源方向の情報から話者が交替している時間フレームを検出し，その情報をもとに，システムに入力された音声がシステムへの要求なのか，単なる雑談なのかを判別する手法も提案されている [3]．ロバストネスの向上に関しては，遅延和アレーや適応型アレーのように，目的音声の方向に指向性を形成する，あるいは雑音方向に死角を形成することで目的音声を強調や雑音の抑圧を行う手法が多く提案されている [4, 5, 6, 7, 8, 9]．

これまでに多くの音源位置推定の研究がされてきたが，これらの研究では，マイクロホンアレーと呼ばれる複数のマイクロホンを用いて，音源位置を推定する手法がほとんどである [10, 11, 12, 13, 14]．マイクロホンアレーに基づく手法では，音声が各マイクロホン

に到達する時間が音源の位置によって異なる点に着目し，各マイクロホンにおける観測信号間の時間差（位相差）を用いて音源位置を推定している．一方，マイク一つのみで音源位置を推定する方法論は未だ確立されておらず，単一マイクロホンによる音源位置推定は非常に困難な試みとされていた．単一マイクロホンで音源位置推定を行うことができれば，コストの削減やシステムの縮小化が可能となる他，従来のマイクロホンアレーのシステムと組み合わせることで，システム全体をより強化することも期待できる．

システムの縮小化，コスト削減のために，単一マイクで処理を行おうとする試みは雑音抑圧や音源分離などの分野でも研究されており，多くの手法が提案されている [15, 16, 17, 18, 19]．しかしながら，単一マイクロホンで音源の位置や方向を推定しようとした場合，従来手法のような信号間の位相差といったマイク間の情報が使えないため，別の情報を用いた音源位置・方向推定手法の提案が必要である．

本研究では，音声の持つ音響伝達特性が音源位置や方向に依存する点に着目し，音響伝達特性を用いることで音源の位置や方向を単一マイクロホンのみで推定する手法について提案し，そのアプローチとして大きく二つの枠組みを提案する．一方は，マイクを中心に回転するパラボラ反射板を用いて，反射板の回転により変動した音響伝達特性を検出することで，音源の方向を推定する，アクティブマイクロホンによる音源方向推定法である．もう一方は，部屋の音響伝達特性が音源位置に依存して変化する点に着目し，音響伝達特性を識別することで，音源の位置を推定する手法である．また，後者の枠組みを応用した手法として，シングルチャネルによる複数話者の音源位置推定手法，話者の頭部方向の推定手法を提案する．

マイクによって観測される信号は，原音声（クリーン音声）と音響伝達特性（インパルス応答）が畳み込まれた残響信号である．そのため提案するシステムでは，残響信号からクリーン音声成分を取り除き，音響伝達特性成分のみをブラインドで推定する必要がある．本手法では，あらかじめ話者のクリーン音声を Gaussian Mixture Model (GMM) あるいは Hidden Markov Model (HMM) でモデル化しておき，これをクリーン音声成分の事前知識として活用することで，観測信号から音響伝達特性を確率的に推定している．そして，推定された音響伝達特性を用いて音源の位置や方向を推定する．

本論文は 7 章から構成されている．第 2 章では提案手法を説明する上で重要となる音響特徴量，音響モデルとそのパラメータの推定法について説明し，さらに従来用いられてきたマイクロホンアレーによる音源位置推定法として，代表的な手法である Cross-power Spectrum Phase (CSP) 法と Multiple Signal Classification (MUSIC) 法について説明する．第 3 章では提案する枠組みの一つである，アクティブマイクロホンによる音源方向推定法について述べる．第 4 章ではもう一方の提案手法である，部屋の音響伝達特性の識別による音源位置推定法について述べる．第 5 章では，第 4 章で提案した手法を，複数話者の音源位置推定問題へ拡張する手法について述べる．第 6 章では，第 4 章で提案した手

法を応用した，話者の頭部方向を推定する手法について述べる．最後に第 7 章にて，全体を通してのまとめと，今後の課題について述べる．

第 2 章

音声信号処理の基本技術及び従来の音源位置推定手法

2.1 まえがき

本章では，提案手法を説明するために必要となる知識として，音響特徴量である Mel-Frequency Cepstral Coefficient (MFCC)[20, 21]，音響モデルである混合正規分布 (GMM: Gaussian Mixture Model)，隠れマルコフモデル (HMM: Hidden Markov Model) とそのパラメータ推定法 [22] について説明する．これらの技術は，以下で説明する従来の音源位置・方向推定の手法では利用されていないが，本研究で提案する手法では，音響伝達特性を推定するために必要な技術である．

さらに，従来手法である，複数のマイクロホンを用いた音源位置・方向推定について述べる．ここでは，特に各マイクロホンで観測された信号間の位相差（時間差）を用いた手法として代表的な Cross-power Spectrum Phase (CSP) 法 [13, 23, 24] と Multiple Signal Classification (MUSIC) 法 [25, 26, 27] について紹介する．

2.2 音声信号処理の基本技術

2.2.1 音響特徴量抽出

2.2.1.1 **ケプストラム** (Cepstrum)

音声は元々は単なる声帯の振動であり，これ自体には音韻を区別する音色は存在しない．この声帯振動による波が，Fig. 2.1 のように口腔や舌などの声道によって共振したものが音声であり，発声によって声道の形を変えて共振周波数を変化させることで，「あ」や

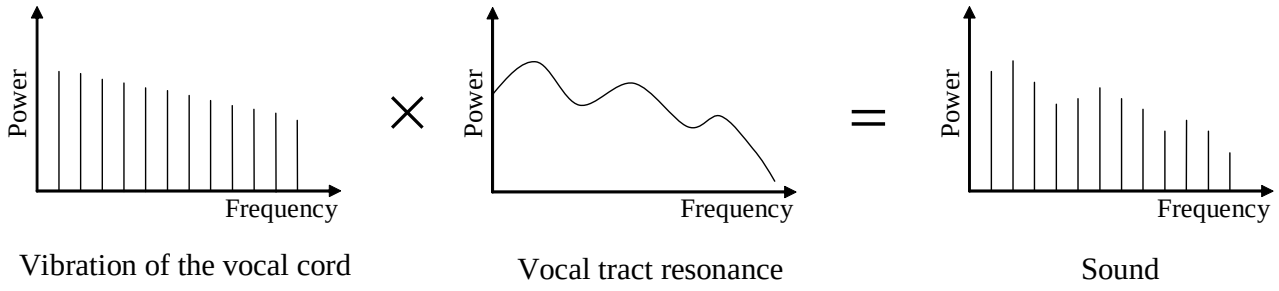

図 2.1　Vibration of the vocal cord and filtering effect of vocal tract

「い」といった様々な音韻の音声を発することができる．この声道による調音フィルタをフォルマントと呼び，声帯振動の基本周波数をピッチと呼ぶ．音声認識においては，音韻を決定付けるフォルマントが特に重要な情報であり，また声帯振動は安定して基本周波数を得ることが難しいため，音声スペクトルからフォルマント成分のみを抽出して用いることで通常のスペクトルを用いるよりも高い認識精度を得ることができる．

音声スペクトルからフォルマントを分離する手法として，代表的なものにケプストラム分析がある [20]．ケプストラムは音のスペクトルを信号とみなし，それをさらにフーリエ変換を行ったもので，音声信号の対数スペクトルを逆フーリエ変換したものとして定義される．前述の通り，音声は声帯振動が声道による調音フィルタによって共振されたものであるので，声帯振動のスペクトルを $G(\omega)$，声道共振の共振特性（フォルマント）を $H(\omega)$ とおくと，音声信号のスペクトル $S(\omega)$ は

$$S(\omega) = G(\omega) \cdot H(\omega) \tag{2.1}$$

と表される．この振幅スペクトルに対数変換，逆フーリエ変換を行ってケプストラムを求めると

$$\log|S(\omega)| = \log|G(\omega)| + \log|H(\omega)| \tag{2.2}$$

$$\begin{aligned} S_{cep}(d) &= DFT^{-1}[\log|S(\omega)|] \\ &= DFT^{-1}[\log|G(\omega)|] + DFT^{-1}[\log|H(\omega)|] \end{aligned} \tag{2.3}$$

と表される．$S_{cep}(d)$ は音声信号のケプストラムを表し，d はケプストラムの次元で，ケフレンシー (quefrency) と呼ばれる．Fig. 2.1 の各スペクトルをそれぞれ一つの信号とみなすと，声帯振動スペクトル $G(\omega)$ は変化の激しい信号であり，共振特性 $H(\omega)$ は逆に変化の緩やかな信号である．そのため，それらに逆フーリエ変換を行うと，$DFT^{-1}[\log|G(\omega)|]$ は高ケフレンシー部に，$DFT^{-1}[\log|H(\omega)|]$ は低ケフレンシー部に現れる．また，式 (2.3) より，ケプストラム空間では音声信号は声帯振動と共振特性の和で表されるため，単純な

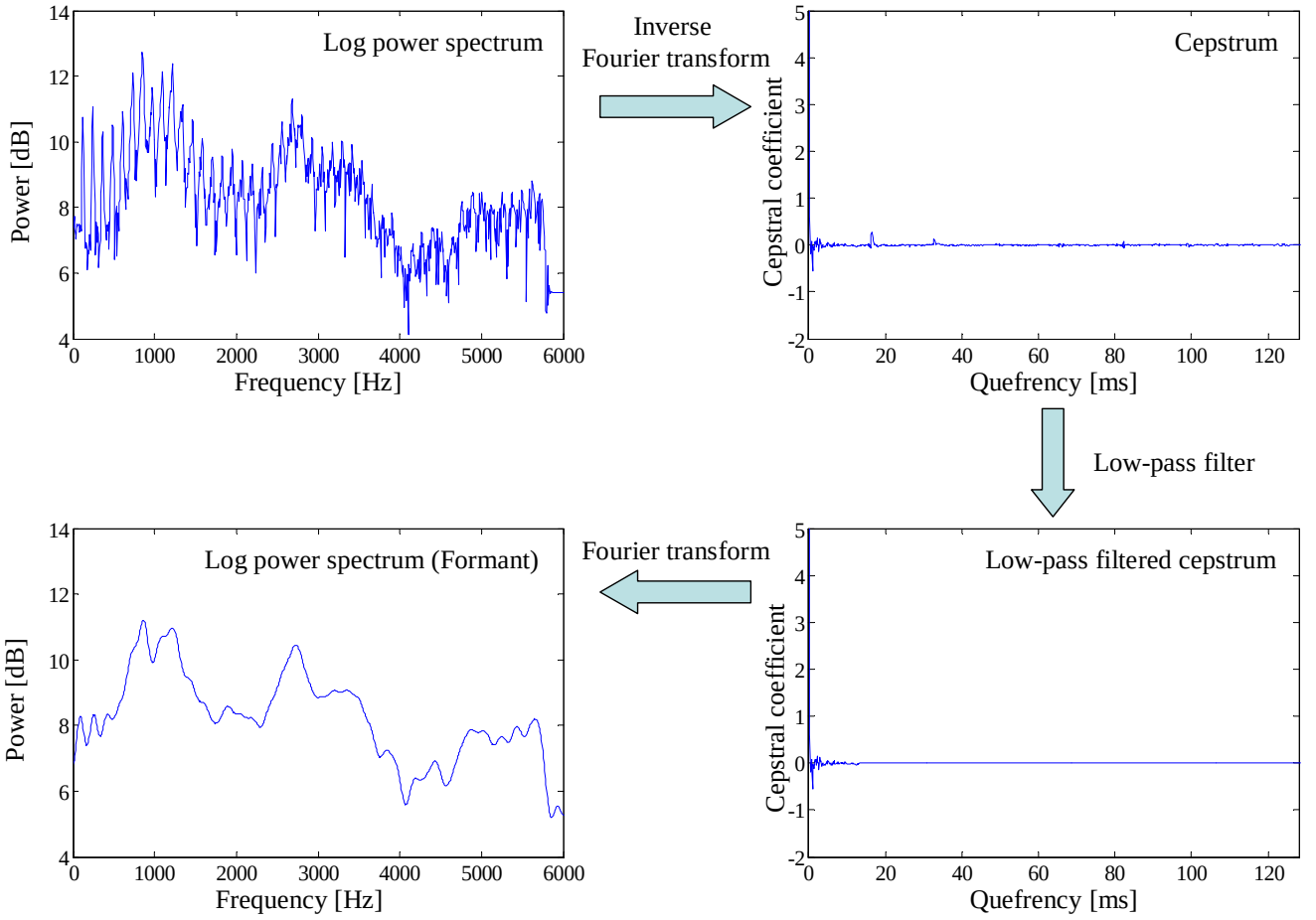

図 2.2　Formant extraction by cepstrum analysis

減算により共振特性 $DFT^{-1}[\log|H(\omega)|]$ のみを抜き出すことが可能となる．そこで，音声信号のケプストラム $S_{cep}(d)$ の高ケフレンシー部を声帯振動及びノイズ成分とみなし，低ケフレンシー成分のみを取り出すリフタリング (liftering) 処理により，フォルマント成分のみを取り出すことができる．

実際に「あ」と発話された音声の対数パワースペクトルから，ケプストラム分析により対数スペクトル包絡をとりだした図を Fig. 2.2 に示す．左上の図は音声の対数パワースペクトルを表し，それをケプストラムに変換したものが右上の図である．右上の図を見ると，20 ms 付近にピークが存在しているのが分かる．このピークが声帯振動の主成分であるため，15 ms 以下のケフレンシー部のみを取り出すフィルタリング処理を行い（右下図），フーリエ変換を行って対数スペクトルに戻すと（左下図），対数スペクトル包絡のみが取り出せていることが確認できる．

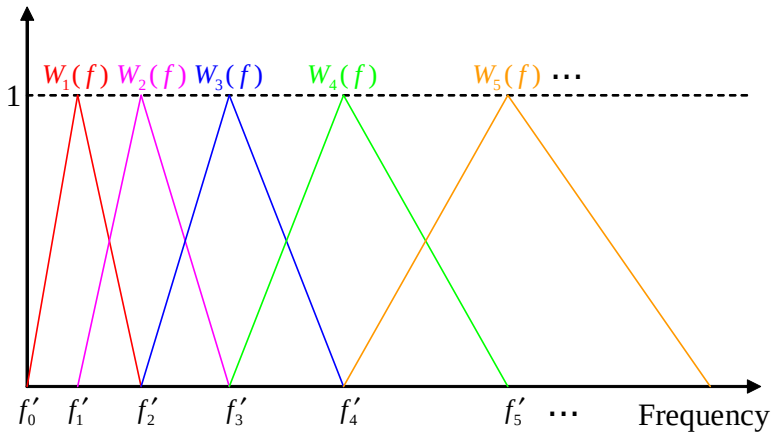

図 2.3　Mel scale filter bank

2.2.1.2　MFCC (Mel Frequency Cepstrum Coefficient)

Fig. 2.2 では低ケフレンシー成分を取り出した後，再度フーリエ変換を行い対数スペクトルに戻しているが，音声認識においてはスペクトルに戻す必要はなく，ケプストラムのまま特徴量として用いている．また，低ケフレンシー成分のみを用いるため，スペクトルを用いるより低次元のパラメータを得ることができ，計算コストを低減することができる．2.2.1.1 節で説明したケプストラムは，線形周波数軸上での変換であるので，LFCC (Linear Frequency Cepstral Coefficient) と呼ばれている．一方，メル尺度と呼ばれる，人間の音の高さに対する感覚尺度での周波数軸上で変換を行う MFCC (Mel Frequency Cepstrum Coefficient) がある [20].

人間が音の高さに対する感覚は，音高が高くなるにつれて鈍くなる．これは人間の聴覚の周波数分解能が，低い周波数に対しては細かく，高い周波数では粗いことを意味する．この人間の周波数分解能は線形ではなく，以下のような非線形の式によって近似されている．

$$f' = 1127.01048 \log(1 + \frac{f}{700}). \tag{2.4}$$

この f' をメル周波数と呼ぶ．音声認識のためには，特徴量の抽出も人間の聴覚特性に合わせて行ったほうがよいと考えられる．音声スペクトルをメル周波数軸で表現する手法として，メルフィルタバンクがある．メルフィルタバンクは，Fig. 2.3 のようなメル周波数軸上で等間隔に配置されたフィルタ群のことである．フィルタは三角窓で構成されており，各フィルタの出力をフィルタごとに総和を取ったものをメルフィルタバンクの出力と

する．

$$M(i) = \sum_{f=f'_{i-1}}^{f'_{i+1}} W_i(f) \cdot |X(f)|^2. \tag{2.5}$$

この $M(i)$ をメル周波数軸にスケーリングされたパワースペクトルとして扱い，これに対数変換，逆フーリエ変換を行うことで，メル周波数軸でのケプストラムである MFCC が計算される．

$$M_{cep}(d) = DFT^{-1}[\log M(i)] \tag{2.6}$$

MFCC は音声の生成確率をモデル化する上で有用な情報を多く含んでおり，現在の音声認識技術で最もよく用いられる特徴量である．そのため，音声モデルを用いる本研究でも MFCC を特徴量として用いている．

2.2.2 音響モデルとパラメータ推定

現在の音声認識は，音声のデータベースを用いて，単語や音素といったクラス毎にその生成確率を統計モデルによりモデル化し，評価データに対して，各クラスのモデルの尤度を計算して比較することで判別を行うという手法が主流となっている．本節では，主に用いられており，また本研究でも使用する統計モデルである GMM，HMM と，その基本となる多次元正規分布について説明する．

2.2.2.1 多次元正規分布

一般に，パターン識別に用いられる特徴量は，多次元ベクトルとなる．そこで，多次元ベクトルの正規分布である多次元正規分布について述べる．

入力ベクトルを $\mathbf{x} \in R^D$，平均ベクトルを $\mu \in R^D$，分散共分散行列を $\mathbf{\Sigma} \in R^{D\times D}$ とすると，多次元正規分布の確率密度関数は次式で表される．

$$N(\mathbf{x}; \mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^t \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\} \tag{2.7}$$

ただし，$\mathbf{x}$, μ , $\mathbf{\Sigma}$ の各成分は，

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_D \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1D}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_{D1}^2 & \cdots & \cdots & \sigma_{DD}^2 \end{bmatrix} \tag{2.8}$$

であり，$(\mathbf{x}-\mu)^t$ は $(\mathbf{x}-\mu)$ の転置行列，$|\mathbf{\Sigma}|$ は $\mathbf{\Sigma}$ の行列式を表す．

実際に評価データを用いて確率密度関数を計算する場合，アンダーフローを避けるために確率密度関数の対数を取ったものを尤度として用いている．また後述する EM アルゴリズムの計算にも対数尤度を用いている．その際，分散共分散行列の対角成分以外を 0 としておくと，対数尤度の計算式は

$$\log N(\mathbf{x};\mu,\mathbf{\Sigma}) = -\frac{1}{2}\left(D\log 2\pi + \sum_{d=1}^{D}\log\sigma_{dd}^2 + \sum_{d=1}^{D}\frac{(x_d-\mu_d)^2}{\sigma_d^2}\right) \tag{2.9}$$

となり，計算量を大幅に削減することができる．本研究におけるモデルの分散共分散行列も対角行列として計算を行っている．

2.2.2.2 **混合正規分布** (GMM: Gaussian Mixture Model)

正規分布は単一のピークを持つ単純な分布関数であり，複雑な分布形状を表現することはできない．そこで，複数の正規分布の重みつき和を用いて，複数のピークを持つ分布関数を考える．これを GMM という．混合数 M の GMM に含まれる正規分布（混合要素）m $(m = 1,\cdots,M)$ が出力される事前確率（混合重み）を $\Pr(m) = w_m$ $(\sum_m w_m = 1)$ とすると，入力に対する GMM の出力確率（尤度）は，次式で表される．

$$\begin{aligned}\Pr(\mathbf{x}|\lambda) &= \sum_{m=1}^{M}\Pr(m)\Pr(\mathbf{x}|m,\lambda)\\ &= \sum_{m=1}^{M} w_m N(\mathbf{x};\mu_m,\mathbf{\Sigma}_m)\end{aligned} \tag{2.10}$$

ここで，λ は GMM パラメータの集合を表す．

$$\lambda = \{w_m,\mu_m,\mathbf{\Sigma}_m|m=1,\cdots,M\} \tag{2.11}$$

Fig. 2.4 は 1 次元 2 混合 GMM の例であり，青線と赤線で描かれた各正規分布の重みつき和により GMM が紫線により表されている．

次に GMM の各パラメータである混合重み，平均ベクトル，分散共分散行列の推定方法について説明する．GMM のパラメータ推定は最尤推定法がよく用いられる．最尤推定法は学習データを用いて，データ毎に対するモデルの尤度の総和が最大になるようにパラメータを決定する方法である．前節で説明した多次元正規分布のパラメータを最尤推定法で推定する場合，正規分布の式を各パラメータで微分し，0 と置くことで各パラメータの計算式が得られる．一方 GMM には混合要素 m が隠れ変数として存在している．すなわち GMM から生成されたデータ $\mathbf{x}$ のみではデータとして不完全であり，それがどの混合要素 m から生成されたのかが観測されて初めて完全なデータ $(\mathbf{x},m)$ となる．しかしながら，学習データ $\mathbf{x}$ を用いて GMM を学習する場合，学習データはそれを生成する

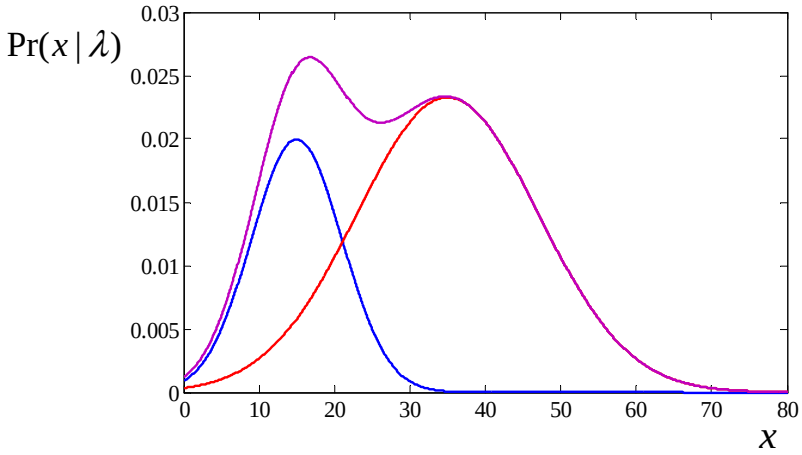

図 2.4　Gaussian mixture model

混合要素 m が未知であり不完全なデータであるため，正規分布のように単純に解くことはできない．そこで，各混合要素から出力される対数尤度を用いて完全データに対する対数尤度の期待値を算出し，それを最大化する EM (Expectation Maximization) アルゴリズムを用いて各パラメータを推定する．EM アルゴリズムは Expectation ステップと Maximization ステップの 2 段階のステップから成り，まず推定したいパラメータに適当な初期値を与え，完全データの対数尤度の期待値である Q 関数を計算する (Expectation ステップ)．次にその Q 関数が最大となるように各パラメータを更新する (Maximization ステップ)．これらのステップを繰り返し，値が収束すればパラメータの推定が完了する．

Expectation ステップにおいて，GMM の完全データに対する対数尤度の期待値（Q 関数）は以下のように表される．

$$\begin{aligned} Q(\hat{\lambda}|\lambda) &= E[\log \Pr(\mathbf{x}, m|\hat{\lambda})|\lambda] \\ &= \sum_{m=1}^{M} \Pr(m|\mathbf{x}, \lambda) \log \Pr(\mathbf{x}, m|\hat{\lambda}). \end{aligned} \tag{2.12}$$

また，N 個のデータからなる $\mathbf{x}$ の生成確率は各データの同時確率である．

$$\begin{aligned} \Pr(\mathbf{x}, m|\hat{\lambda}) &= \prod_{n=1}^{N} \Pr(\mathbf{x}_n, m|\hat{\lambda}) \\ &= \prod_{n=1}^{N} \hat{w}_m N(\mathbf{x}_n; \hat{\mu}_m, \hat{\mathbf{\Sigma}}_m). \end{aligned} \tag{2.13}$$

これを (2.12) に代入すると

$$Q(\hat{\lambda}|\lambda) = \sum_k \sum_n \Pr(m=k|\mathbf{x}_n,\lambda)\log \hat{w}_k + \sum_k \sum_n \Pr(m=k|\mathbf{x}_n,\lambda)\log N(\mathbf{x}_n;\hat{\mu}_k,\hat{\mathbf{\Sigma}}_k). \tag{2.14}$$

となる．次に Maximization ステップでは (2.14) を最大とするようにパラメータの更新式を求めると，更新後の重み $\hat{w}_k$ は,

$$\hat{w}_k = \frac{1}{N}\sum_{n=1}^{N} \Pr(k|\mathbf{x}_n,\lambda) \tag{2.15}$$

となる．更新後の平均ベクトル $\hat{\mu}_k$ は,

$$\hat{\mu}_k = \frac{\sum_n \Pr(k|\mathbf{x}_n,\lambda)\mathbf{x}_n}{\sum_n \Pr(k|\mathbf{x}_n,\lambda)} \tag{2.16}$$

となる．更新後の分散共分散行列 $\hat{\mathbf{\Sigma}}_k$ は,

$$\hat{\mathbf{\Sigma}}_k = \frac{\sum_n \Pr(k|\mathbf{x}_n,\lambda)(\mathbf{x}_n-\hat{\mu}_k)(\mathbf{x}_n-\hat{\mu}_k)^t}{\sum_n \Pr(k|\mathbf{x}_n,\lambda)} \tag{2.17}$$

となる．ここで $\Pr(k|\mathbf{x}_n,\lambda)$ はデータ $\mathbf{x}_n$ が混合要素 k から生成されている確率であり，以下のようにして計算できる．

$$\Pr(k|\mathbf{x}_n,\lambda) = \frac{w_k N(\mathbf{x}_n;\mu_k,\mathbf{\Sigma}_k)}{\sum_k w_k N(\mathbf{x}_n;\mu_k,\mathbf{\Sigma}_k)}. \tag{2.18}$$

2.2.2.3 **隠れマルコフモデル** (HMM: Hidden Markov Model)

HMM はいくつかの状態 s_i と，遷移枝からなる状態遷移モデルであり，状態遷移は一意に決まらないような構造を持っている．例として 5 状態 3 自己ループの構造を持つ left-to-right HMM を Fig. 2.5 に示す．音声認識に用いられる HMM は left-to-right 型で初期状態が 1 つ，最終状態が 1 つのものが多く使われており，各状態における出力確率は GMM で定義されている．これにより，前節で説明した GMM 単体では表現できなかった音韻の時間的変化を HMM では表現することが可能となる．

GMM では各混合要素の出力確率と，そこから出力される確率密度関数から期待値的に尤度を算出していたが，HMM では音声の各フレーム t に対して，それがどの状態 $s(t)$ から生成されているかという状態系列ごとに尤度が算出され，その中から最大尤度を示す状態系列を探索，決定する．状態系列の探索には代表的なものに Viterbi アルゴリズムが，また学習方法には Baum-Welch アルゴリズムなどがある [28]．状態系列 s から生成され

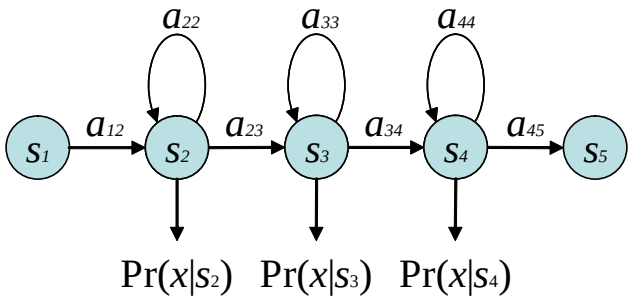

図 2.5　HMM structure of 5 states and 3 loop

るデータ $\mathbf{x}$ に対する混合分布型 HMM の尤度は

$$\begin{aligned}\Pr(\mathbf{x}, s|\lambda) &= \prod_{n=1}^{N} a_{s_{n-1},s_n} \Pr(\mathbf{x}_n|s_n, \lambda) \\ &= \prod_{n=1}^{N} a_{s_{n-1},s_n} \sum_{m=1}^{M} w_{s_n,m} N(\mathbf{x}_n; \mu_{s_n,m}, \mathbf{\Sigma}_{s_n,m}) \end{aligned} \tag{2.19}$$

となる．ここで a_{s_{n-1},s_n} は前フレームの状態 s_{n-1} から現在のフレームの状態 s_n へ遷移する確率 (遷移確率) を表す．モデルのパラメータは GMM 同様に最尤推定法及び EM アルゴリズムを用いて推定すると，Q 関数は以下のようになる．

$$\begin{aligned}Q(\hat{\lambda}|\lambda) &= E[\log \Pr(\mathbf{x}, s, m|\hat{\lambda})|\lambda] \\ &= \sum_{m=1}^{M} \Pr(s, m|\mathbf{x}, \lambda) \log \Pr(\mathbf{x}, s, m|\hat{\lambda}). \end{aligned} \tag{2.20}$$

完全データの出力確率は

$$\begin{aligned}\Pr(\mathbf{x}, s, m|\lambda) &= \prod_{n=1}^{N} a_{s_{n-1},s_n} w_{s_n,m_n} \Pr(\mathbf{x}_n|s_n, m_n, \lambda) \\ &= \prod_{n=1}^{N} a_{s_{n-1},s_n} w_{s_n,m_n} N(\mathbf{x}_n; \mu_{s_n,m_n}, \mathbf{\Sigma}_{s_n,m_n}) \end{aligned} \tag{2.21}$$

となり，これを (2.20) へ代入すると

$$\begin{aligned}Q(\hat{\lambda}|\lambda) = &\sum_i \sum_j \sum_n \Pr(s_{n-1} = i, s_n = j|\mathbf{x}_n, \lambda) \log \hat{a}_{i,j} \\ &+ \sum_j \sum_k \sum_n \Pr(s_n = j, m_n = k|\mathbf{x}_n, \lambda) \log \hat{w}_{j,k} \\ &+ \sum_j \sum_k \sum_n \Pr(s_n = j, m_n = k|\mathbf{x}_n, \lambda) \log N(\mathbf{x}_n; \hat{\mu}_{j,k}, \hat{\mathbf{\Sigma}}_{j,k}) \end{aligned} \tag{2.22}$$

となる．Maximization ステップで (2.22) の各項を最大とするようにパラメータの更新式を求めると，更新後の遷移確率 $\hat{a}_{i,j}$ は

$$\hat{a}_{i,j} = \frac{\sum n \Pr(i,j|\mathbf{x}_n,\lambda)}{\sum_i \sum n \Pr(i,j|\mathbf{x}_n,\lambda)}, \tag{2.23}$$

更新後の重み $\hat{w}_{j,k}$ は

$$\hat{w}_{j,k} = \frac{\sum n \Pr(j,k|\mathbf{x}_n,\lambda)}{\sum_k \sum n \Pr(j,k|\mathbf{x}_n,\lambda)}, \tag{2.24}$$

更新後の平均ベクトル $\hat{\mu}_{j,k}$ は

$$\hat{\mu}_{j,k} = \frac{\sum n \Pr(j,k|\mathbf{x}_n,\lambda)\mathbf{x}_n}{\sum n \Pr(j,k|\mathbf{x}_n,\lambda)}, \tag{2.25}$$

更新後の分散共分散行列 $\hat{\mathbf{\Sigma}}_{j,k}$ は

$$\hat{\mathbf{\Sigma}}_{j,k} = \frac{\sum n \Pr(j,k|\mathbf{x}_n,\lambda)(\mathbf{x}_n - \hat{\mu}_{j,k})(\mathbf{x}_n - \hat{\mu}_{j,k})^t}{\sum n \Pr(j,k|\mathbf{x}_n,\lambda)} \tag{2.26}$$

となる．このアルゴリズムは最尤推定法に基づくパラメータ推定法であるが，近年の音声認識の研究では音声認識に特化したモデルパラメータ推定法として，相互情報量最大化 (MMI: Maximum Mutual Information) 基準 [29] や，音素誤り最小 (MPE: Minimum Phone Error) 基準 [30] が提案されている．

2.3 従来の音源位置推定手法

2.3.1 CSP (Cross-power Spectrum Phase)

Fig. 2.6 のように，ある場所で発せられた音声信号をマイクロホンアレー (p_0，p_1，p_2，$\cdots$，p_{M-1}) で観測したとすると，音源と最も距離が近いマイクロホン p_0 に信号が到来してから各マイクロホンに信号が到来するまでに時間差 δ_{01}，δ_{02}，$\cdots$，δ_{0M-1} が生じる．

入力信号 $r(t)$ が音源からマイク p_i に到達するまでの時間を τ_i とすると，マイク p_i における観測信号 $s_i(t)$ は

$$s_i(t) = \alpha_i r(t - \tau_i) \tag{2.27}$$

と表せる．ただし α_i は音源からマイク p_i までの減衰度である．また，信号がマイク p_i に到達してからマイク p_k に到達するまでの時間差 δ_{ik} は

$$\delta_{ik} = \tau_k - \tau_i \tag{2.28}$$

と表せる．

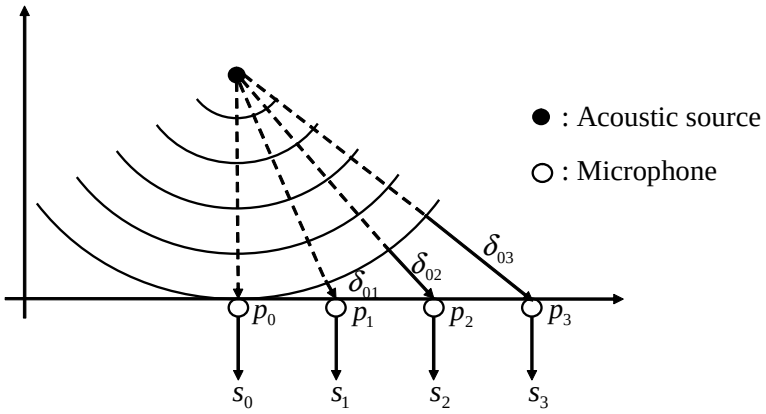

図 2.6 Wavefront propagation of an acoustic stimulus

次に観測信号間の時間差 δ_{ik} を求めるために，各観測信号の相互相関関数を求める．p_i における観測信号 $s_i(t)$ と p_k における観測信号 $s_k(t)$ の相互相関関数 $R_{ik}(\tau)$ は次のように定義される．

$$R_{ik}(\tau) = E[s_i(t)s_k(t+\tau)]. \tag{2.29}$$

ここで E は期待値を表す．式 (2.27) を用いると式 (2.29) は次のように表せる．

$$R_{ik}(\tau) = \alpha_i\alpha_k R_{rr}(\tau - \delta_{ik}). \tag{2.30}$$

フーリエ変換を行うことにより，$s_i(t)$ と $s_k(t)$ のクロススペクトル $G_{ik}(f)$ が得られる．

$$G_{ik}(f) = \alpha_i\alpha_k G_{rr}(f)e^{-j2\pi f\delta_{ik}}. \tag{2.31}$$

参考文献 [23] によると相関関数を求める前にフィルターを掛けることにより，一般化相互相関関数 $R_{ik}^{(g)}(\tau)$ が得られる．

$$R_{ik}^{(g)}(\tau) = \int_{-\infty}^{+\infty} \psi_g(f)G_{ik}(f)e^{j2\pi f\tau}df. \tag{2.32}$$

ここで $\psi_g(f)$ は周波数毎の重みを表すフィルターである．信号の白色化を行うと相互相関関数の形状が鋭くなるため，$\psi_g(f)$ を

$$\psi_g(f) = \frac{1}{|G_{ik}(f)|} \tag{2.33}$$

と設定することで位相相関 $R_{ik}^{(p)}(\tau)$ が得られる．

$$R_{ik}^{(p)}(\tau) = \int_{-\infty}^{+\infty} \frac{G_{ik}(f)}{|G_{ik}(f)|}e^{j2\pi f\tau}df. \tag{2.34}$$

また，クロススペクトルをパワーで割ると位相成分が得られる．

$$\frac{G_{ik}(f)}{|G_{ik}(f)|} = e^{-j2\pi f\delta_{ik}}. \tag{2.35}$$

式 (2.34)，(2.35) より，$R_{ik}^{(p)}(\tau)$ は入力信号に依存せず，信号間の時間差 δ_{ik} のみに依存することが分かる．

実際には $R_{ik}^{(p)}(\tau)$ の代わりに観測信号のクロススペクトルをパワーで割り，逆フーリエ変換を行うことにより位相情報を求める．s_i と s_k の時刻 t におけるスペクトルをそれぞれ $\hat{S}_i(t,f)$，$\hat{S}_k(t,f)$ とすると，

$$\phi(t,f) = \frac{\hat{S}_i(t,f)\hat{S}_k^*(t,f)}{|\hat{S}_i(t,f)||\hat{S}_k(t,f)|} \tag{2.36}$$

$$\tilde{R}_{ik}(t,\tau) = DFT^{-1}[\phi(t,f)] \tag{2.37}$$

が求められる．$\tilde{R}_{ik}(t,\tau)$ を CSP 係数と呼ぶ．ここで式 (2.34)，(2.35) より，$R_{ik}^{(p)}(\tau)$ は $\tau = \delta_{ik}$ のとき最大となるので，CSP 係数 $\tilde{R}_{ik}(n,l)$ を用いて時間差の推定値 $\hat{l}_{ik}$ を求めることができる．

$$\hat{l}_{ik} = \underset{l}{\mathrm{argmax}}[\sum_{n=1}^{N} \tilde{R}_{ik}(n,l)]. \tag{2.38}$$

2.3.2　MUSIC (MUltiple SIgnal Classification)

Fig. 2.7 のように，M 個のマイクロホンが間隔 d で並んでいるマイクロホンアレーに K 個の音源から信号が到来しているとする．音速を c，音源 q の方向を θ_q，マイクロホンで受信する複素音圧値を $S_q(t)$ とすると，隣接するマイクロホン間での到来時間差は $\tau_q = d\sin\theta_q/c$，i 番目のマイクロホンで受信される複素音圧は $S_q(t)e^{-j\omega(i-1)\tau_q}$ と表せる．K 個の音源から波が到来しているときの受信信号の複素音圧ベクトル $\mathbf{x}(t)$ は，それらの和として

$$\begin{aligned}\mathbf{x}(t) &= \sum_{q=1}^{K} S_q(t)\mathbf{a}_q + \mathbf{n}(t) \\ \mathbf{a}_q &= [1, e^{-j\omega\tau_q}, e^{-j\omega 2\tau_q}, \cdots, e^{-j\omega(M-1)\tau_q}]^T \\ \mathbf{n}(t) &= [n_1, n_2, \cdots, n_M]^T\end{aligned} \tag{2.39}$$

で表される．ここで，$\mathbf{a}$ は音源の方向を表す位相ベクトル，$\mathbf{n}$(t) はマイクロホンにおける雑音成分のベクトルとする．

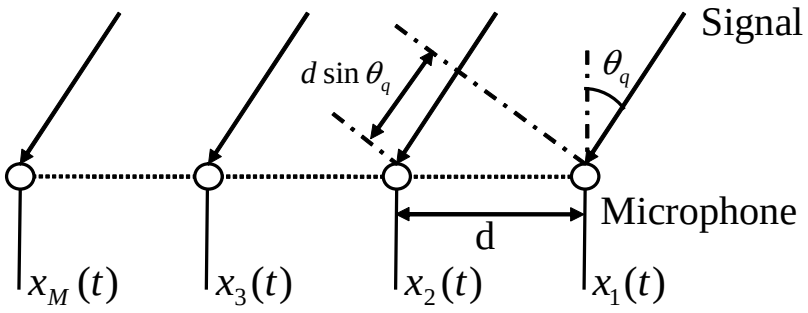

図 2.7 Microphone arrays model

到来する各音波および雑音がそれぞれ無相関であると仮定すると，$\mathbf{x}(t)$ の相関行列 $\mathbf{R}_{xx}$ は，

$$\mathbf{R}_{xx} = E[\mathbf{x}(t)\mathbf{x}(t)^H] \tag{2.40}$$

となり，$\mathbf{R}_{xx}$ について固有値分解を行うことで信号空間と雑音空間に分けることができる．

$$\mathbf{R}_{xx} = \mathbf{V}_s\mathbf{\Lambda}_s\mathbf{V}_s^H + \mathbf{V}_n\mathbf{\Lambda}_n\mathbf{V}_n^H \tag{2.41}$$

$$\mathbf{\Lambda}_s = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_K \end{bmatrix}, \mathbf{\Lambda}_n = \begin{bmatrix} \lambda_{K+1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_M \end{bmatrix}$$

$$\mathbf{V}_s \equiv [\mathbf{v}_1, \cdots, \mathbf{v}_K], \mathbf{V}_n \equiv [\mathbf{v}_{K+1}, \cdots, \mathbf{v}_M].$$

ただし，H は転置複素共役，$\mathbf{R}_{xx}$ の固有値を $\lambda_m (1 \leq m \leq M)$，$\lambda_m$ に対する固有ベクトルを $\mathbf{v}_m (1 \leq m \leq M)$ とする．λ_m は大きい順に並んでいる．

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K > \lambda_{K+1} = \cdots = \lambda_M. \tag{2.42}$$

ここで位相ベクトル $\mathbf{a}_q(\theta)$ と雑音部分空間の固有ベクトル $\mathbf{V}_n$ を用いて

$$P_{MUSIC}(\theta) = \frac{\mathbf{a}_q(\theta)^H\mathbf{a}_q(\theta)}{\mathbf{a}_q(\theta)^H\mathbf{V}_n\mathbf{V}_n^H\mathbf{a}_q(\theta)} \tag{2.43}$$

と定義する．ここで，$\mathbf{v}_1, \cdots, \mathbf{v}_M$ は互いに直交しているので，信号空間 $\mathbf{v}_1, \cdots, \mathbf{v}_K$ と $\mathbf{v}_{K+1}, \cdots, \mathbf{v}_M$ は直交する．従って $\mathbf{a}_1, \cdots, \mathbf{a}_K$ と $\mathbf{v}_{K+1}, \cdots, \mathbf{v}_M$ も直交する．ここで，式 (2.43) の θ と θ_q が一致したとき，$\mathbf{a}_1, \cdots, \mathbf{a}_K$ と $\mathbf{v}_{K+1}, \cdots, \mathbf{v}_M$ の直交性より右辺の分母が零となる．そこで θ を変化させ，$P_{MUSIC}(\theta)$ のピーク値を探し，その値を示した θ を音源 q の方向として推定する．

2.3.3　その他の手法

2.3.1 節，2.3.2 節ではマイクロホンアレーの各マイクで観測される信号間の位相差を用いた手法として，CSP 法と MUSIC 法について説明した．また位相差の他には音圧差を利用した方法や，バイノーラル信号による両耳間の差を利用した手法などが提案されている [31, 32] が，これらはいずれも複数のマイクが必要である．単一マイクでの音源方向推定の試みとしては，頭部伝達特性 (HRTF) を用いて音源位置を推定する手法や [33, 34]，生物の耳介を模倣して，耳介からの反射波を利用する手法が提案されている [35, 36, 37, 38].

第 3 章

パラボラ反射板による音響伝達特性の変動検出に基づくシングルチャネル音源方向推定

3.1 まえがき

本章では，音響伝達特性を用いた音源方向推定のアプローチの一つとして，マイクを中心に回転するパラボラ反射板を用いて，反射板の回転により変動した音響伝達特性を検出することで，音源の方向を推定する手法を提案する．この手法は，人間の聴覚システムに倣った音源方向推定手法である．従来の音源方向推定の枠組みにおいて，二つのマイクでは正中面上の仰角方向の推定はできないが，人間は二つの耳でも耳介（外耳）によって仰角方向の推定が行える．このことから，耳介のような反射板と，それによって変動する音響伝達特性が，単一マイクロホンによる音源方向推定に有効であると考えられる．また，人間は音源の方向に耳を向けるというように，聴覚システムが能動的である点にも着目して，この章では，反射板が回転することで音源の方向を推定する，アクティブマイクロホンを提案する．焦点位置に単一マイクを備え付けた，パラボラ型の反射板が回転することで，反射板が音源方向を向いたときのみ音響伝達特性が大きく変動する．この手法では，観測信号の音響伝達特性をクリーン音声 GMM により推定し，音響伝達特性が大きく変動する反射板の角度を検出することで，音源方向を推定する．

パラボラ反射板を用いて信号の発信源の方向を推定する手法はレーダーの分野においては既に提案されている [39]．これらの手法では，反射板が音源の方向を向いたときに反射板から到来する反射波が焦点に集まることにより，信号のパワーが増幅される性質を利用

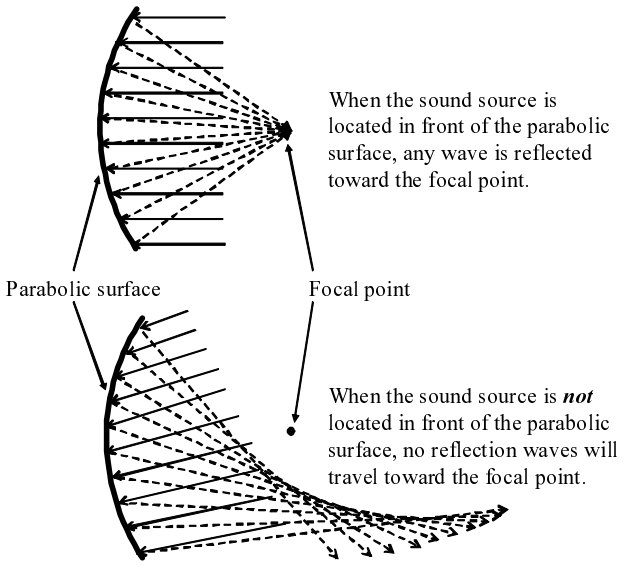

図 3.1　Concept of parabolic reflection

している．しかしながら，従来のパワーを利用した手法は，発信される原信号のパワーが一定であれば有効に働くが，音声のようなパワーが変動する信号においては，単純にパワーが最大となる方向を検出しても，それが音源の方向であるとは限らない．そこで，本手法ではパワーの代わりに，原音声に依存しない音響伝達特性を用いて音源方向を推定している．実験ではパワーに基づく手法と比較し，音響伝達特性を用いることの有用性を示す．

3.2　アクティブマイクロホン

3.2.1　パラボラ反射板

本手法では放物面状の反射板をマイクロフォンに装着することで方向推定を試みる．到来する音波が平面波であると仮定した場合，Fig. 3.1 で示されるように放物面の軸と平行に入射した波はすべて放物面の焦点に向かって反射される．一方別の角度から入射した波が反射によって焦点に到達することはない．

放物面の焦点の位置にマイクロホンを装着すると，音波が放物面の正面方向から到来したとき，反射波がすべて焦点に集まるのでマイクロホンで観測された信号の音響伝達特性が変動する．一方別の方向から到来した場合，反射波は焦点に到達しないので観測信号の音響伝達特性は変化しない．

このように，音波が正面方向から到来した場合と別の方向から到来した場合では観測信

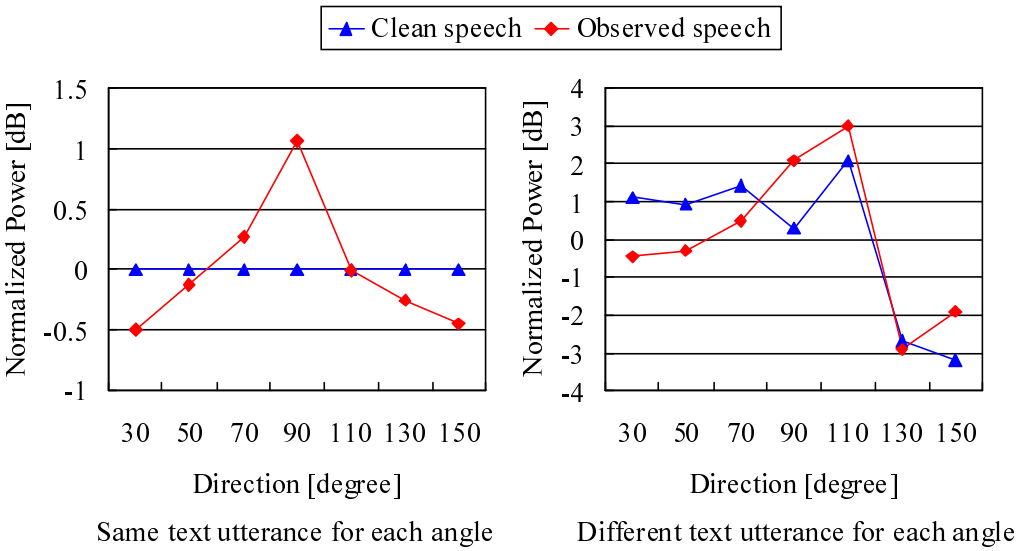

図 3.2 Power of a clean speech segment and the speech segment observed by the microphone with a parabolic reflection board for each angle. The power was normalized so that the mean values of all directions was 0 dB.

号の音響伝達特性に差が生じると考えられる．そこでマイクロホンを反射板と一緒に回転させながら観測信号の音響伝達特性を求め，音響伝達特性が大きく変動したときの角度を音源方向として推定することができる．

3.2.2 観測信号のパワーに基づく音源方向推定

レーダーの分野で用いられている信号原の方向推定では，観測信号のパワーを用いて方向を推定している [39]．そこで，比較手法として，パワーを用いて音源方向を推定する場合を考える．パラボラ反射板が音源方向を向くと，反射波が焦点に集まるため，音声のパワーが増幅される．そのため，マイクロホンを反射板と一緒に回転させながら観測信号のパワーを求め，パワーが最大となったときの角度を音源方向として推定することができると考えられる．

$$\hat{i} = \underset{i}{\operatorname{argmax}} \sum_{n} \sum_{\omega} \log |O_i(\omega; n)|^2. \tag{3.1}$$

$O(\omega; n)$ は n フレーム目の音声スペクトルの ω 番目の周波数ビンを表す．i はパラボラ反射板の向いている方向である．

この手法は，原信号がホワイトノイズのような定常信号であれば，効果的と考えられるが，音声信号のような非定常信号の場合，原信号のパワーが一定でないため，例え反射板

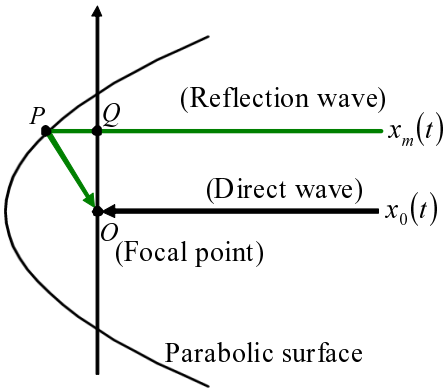

図 3.3　Observed signal at the focal point, where the input signal is coming from directly in front of the parabolic surface

によってパワーが増幅されたとしても，そのときの原信号のパワーが小さければ，音源方向として検出されるとは限らない．Fig. 3.2 は，反射板が回転している間に発話された音声信号の原信号のパワーとアクティブマイクロホンによって収録された信号の反射板の方向毎のパワーを示している．左図は反射板が回転している間，話者が同じ発話内容を同じ音圧で発声していた場合であり，右図は回転している間に発話内容や音圧が変わっている場合である．音源方向は 90°，反射板の直径は 24cm，焦点距離は 9cm である．サンプリング周波数は 12kHz，分析窓はハミング窓を使用し，窓幅は 32msec，フレームシフトは 8msec である．パワーは平均が 0dB になるように正規化されている．

図より，音源方向である 90° においてパワーが大きく増幅されていることが分かる．しかしながら，発話内容や音圧が反射板の回転中に変化する場合では，パワーが 90° において最大になっておらず，この場合では音源方向の推定に失敗することになる．そこで本手法では，パワーの代わりに，原信号に依存しない音響伝達特性を用いて音源方向を推定する．

3.2.3　パラボラ反射板によって変動する音響伝達特性

アクティブマイクロホンによって収録される観測信号について考える．背景雑音がなく，かつ反射板が音源方向を向いている場合，Fig. 3.3 で示される通り，時刻 t における観測信号は，直接焦点へ向かう信号（直接波）と，パラボラ反射板を反射して焦点へ到来する信号（反射波）の線形和で表される．

$$o(t) = x_p(t) + \sum_{m=1}^{M} x_m(t) \tag{3.2}$$

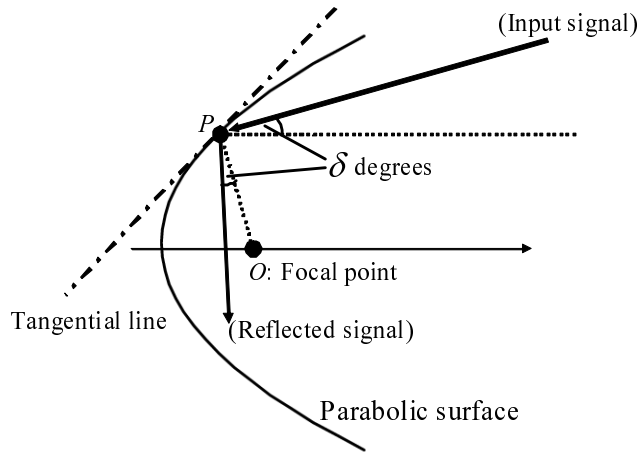

図 3.4 Observed signal at the focal point, where the input signal is coming from δ degrees

$o(t)$, x_p, x_m $(m = 1, \cdots, M)$ はそれぞれ観測信号，直接波，反射波を表す．放物線の性質より，直接波と反射波の焦点までの時間差は m の値に関わらず一定であるので，式 (3.2) は以下のように表現できる．

$$o(t) = s(t) * h_p(t) + \sum_{m=1}^{M} s(t-\tau) * h_m(t) \tag{3.3}$$

$s(t)$，τ はクリーン音声および到来時間差を表す．h_p，h_m はそれぞれ直接波と反射波のインパルス応答である．短時間フーリエ変換を行うと，n フレーム目の観測信号のスペクトルは以下のようになる．

$$\begin{aligned} &O(\omega; n) \\ &\approx S(\omega; n) \cdot (H_p(\omega; n) + e^{-j2\pi\omega\tau} \cdot \sum_{m=1}^{M} H_m(\omega; n)) \\ &= S(\omega; n) \cdot (H_p(\omega; n) + H_r(\omega; n)). \end{aligned} \tag{3.4}$$

H_p は直接波の音響伝達特性であり，パラボラ反射板の影響は受けない．H_r はパラボラ反射板によって追加される音響伝達特性である．

一方，Fig. 3.4 のように，反射板が音源方向から角度 δ ずれた方向を向いている場合，反射波も焦点の方向から δ ずれた方向へ反射される．そのため，反射板が音源方向を向いていない場合は，音響伝達特性がパラボラ反射板によって変動することはない．

$$O(\omega; n) \approx S(\omega; n) \cdot H_p(\omega; n). \tag{3.5}$$

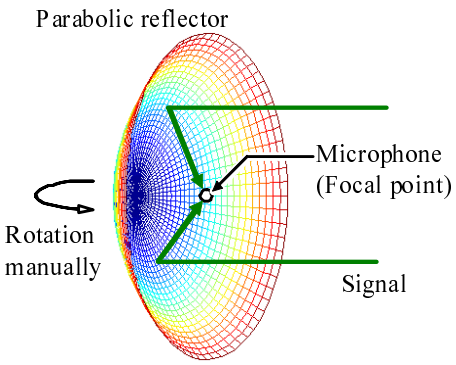

図 3.5　Active microphone with parabolic reflection

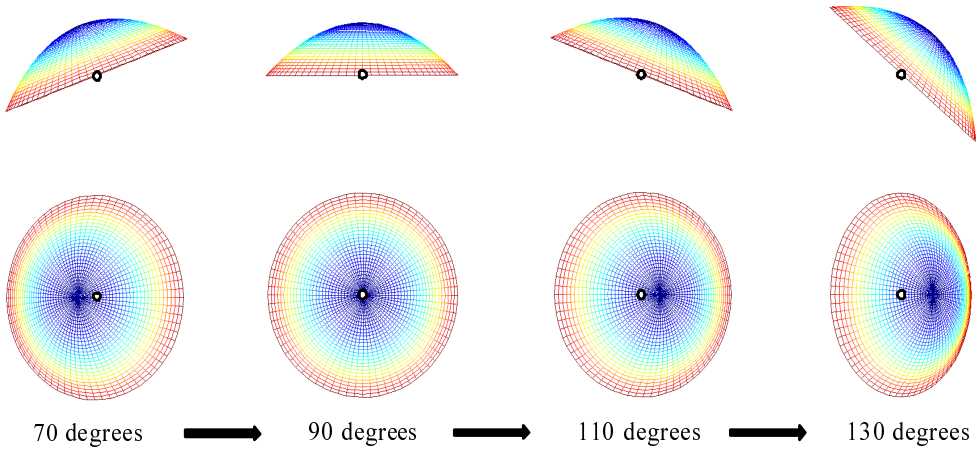

図 3.6　Rotated active microphone

3.2.4　音源方向の推定

Fig. 3.5 のようにパラボラ反射板の焦点位置にマイクロホンを装着し，マイクロホンと反射板が回転することにより音源の正面方向を探索する．本研究では反射板及びマイクロホンは Fig. 3.6 のように手動で回転させ，反射板の方向を離散的に変化させている．

式 (3.4)，(3.5) より，アクティブマイクロフォンの向きを θ，音源方向を $\hat{\theta}$ とすると伝達特性 $H_\theta(\omega;n)$ は

$$
\begin{aligned}
&O_\theta(\omega;n) \approx S_\theta(\omega;n)\cdot H_\theta(\omega;n)\\
&H_\theta(\omega;n) = \begin{cases} H_p(\omega;n)+H_r(\omega;n) & (\theta=\hat{\theta})\\ H_p(\omega;n) & (\theta\neq\hat{\theta}) \end{cases}
\end{aligned}
\tag{3.6}
$$

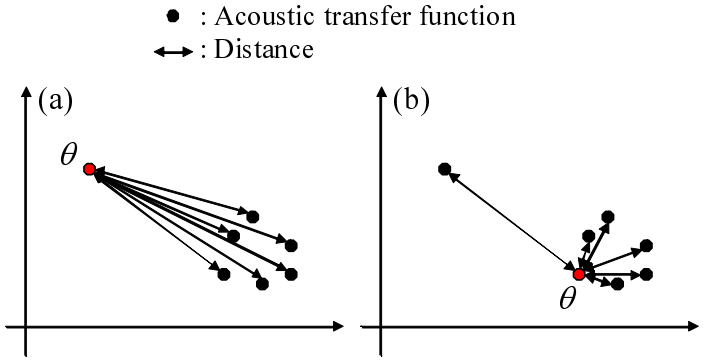

図 3.7 Acoustic transfer function in a feature space for each angle of the active microphone. (a) The case that the direction θ has the acoustic transfer function which is the farthest from those of other directions. (b) The case that the acoustic transfer function of θ is similar to most of the acoustic transfer functions of other directions.

となる．ここでアクティブマイクロフォンの角度が変化しても H_p の値がほぼ一定であれば，H_θ はマイクが音源方向を向いていない状態ではほぼ同じ空間上に分布し，音源方向を向いたときのみそれらとは離れた位置に分布することになる．そこで，各角度で推定される音響伝達特性 H_θ の中で，最も離れた位置に分布する音響伝達特性 $H_{\hat{\theta}}$ を見つけ，それに対応する角度 $\hat{\theta}$ を音源方向として出力する．

本研究では最も離れた位置に分布する音響伝達特性を求める方法として，アクティブマイクの各方向において推定された音響伝達特性の相互距離の総和を用いる．まず，アクティブマクロホンの向きを離散的に変化させ，その向き毎に音響伝達特性を推定する．そして，ある方向 θ における音響伝達特性 H_θ を評価する際，その平均値ベクトル $\bar{H}_\theta$ と，その他の方向における音響伝達特性の平均値ベクトル $\bar{H}_{\theta'}$ とのユークリッド距離を方向毎に計算する．このとき，Fig. 3.7 に示されるように，もし $\bar{H}_\theta$ が他の方向の音響伝達特性から離れた位置に分布している場合，このユークリッド距離の総和は大きくなる．一方，$\bar{H}_\theta$ が他の方向の音響伝達特性と近い位置に分布している場合，ユークリッド距離の総和は小さくなる．そこで，各アクティブマイクの方向について，他の方向の音響伝達特性との相互距離の総和を計算し，この値が最も高くなったときの方向を，音源方向として出力する．

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{\theta'} (\bar{H}_\theta - \bar{H}_{\theta'})^2 \tag{3.7}$$

ここで $\bar{H}$ は H のフレーム平均を表す．

3.3　クリーン音声 GMM による音響伝達特性の推定

3.3.1　残響音声信号のモデル化

本節では，クリーン音声の GMM を用いて音響伝達特性を推定する手法について述べる．ある位置から発話された残響信号 $x(t)$ は，クリーン音声 $s(t)$ と音響伝達特性 (インパルス応答)$h(t)$ の畳み込みで表される．

$$x(t) = \sum_{l=0}^{L-1} s(t-l)h(l) \tag{3.8}$$

L はインパルス応答の長さを表す．次に短時間フーリエ変換を行うことで短時間スペクトルが得られる．音声認識や残響除去の分野では，短時間スペクトル上での残響信号を周波数ビン毎にクリーン音声スペクトルと音響伝達特性の畳み込みで表現している [40, 41]．しかしそれらの表現は複雑で音響伝達特性のフレーム系列が求め難く，本研究の音源位置推定に組み込みにくいため，本研究では残響信号の短時間スペクトルを，近似的にクリーン音声と音響伝達特性の短時間スペクトルの乗算で表現することにする．

$$O(\omega; n) \approx S(\omega; n) \cdot H(\omega; n) \tag{3.9}$$

$O(\omega; n)$，$S(\omega; n)$，及び $H(\omega; n)$ はそれぞれフレーム n における残響信号，クリーン音声，音響伝達特性の短時間スペクトルを，ω は周波数のビンを表す．

このモデル化において，パワースペクトルに対数変換と離散コサイン変換を適用することで得られる，残響信号のケプストラムは以下のような線形加算で表現される．

$$O_{cep}(d; n) \approx S_{cep}(d; n) + H_{cep}(d; n) \tag{3.10}$$

O_{cep}，S_{cep}，及び H_{cep} はそれぞれ残響信号，クリーン音声，音響伝達特性のケプストラムを表し，d はその次元を表す．ケプストラムは，音声情報を効率よく表現できるパラメータの一つであり，音声認識などでよく用いられていることから，本手法においてもケプストラム空間上で音響伝達特性の推定を行う．ただし本研究において，実際には音声情報をさらに効率よく表現するために，周波数軸をメル尺度化したケプストラムである MFCC を特徴量として用いている．MFCC の場合では，パワースペクトルに対数変換を行う前に，2.2.1.2 節で述べたメルフィルタバンク処理を行っているため，正確には式 (3.9) から式 (3.10) への変換は成り立たない．しかし文献 [42] では，メルフィルタバンク処理とクリーン音声・音響伝達特性スペクトルの積を逆にしても，残響音声のモデル化にはさほど影響が出ないことを示した上で，MFCC 上でも式 (3.9) から式 (3.10) への変換が近似的に成立すると仮定している．本研究においてもこの仮定を用いることにする．

ここで，仮にクリーン音声のフレーム系列が既知であれば，音響伝達特性のフレーム系列は

$$H_{cep}(d;n) \approx O_{cep}(d;n) - S_{cep}(d;n). \tag{3.11}$$

として求めることができる．しかしながらクリーン音声のフレーム系列は実際の環境では未知であるため，クリーン音声の GMM を用いて最尤推定法により音響伝達特性を推定する．以降，簡単のため O_{cep}，S_{cep}，及び H_{cep} をそれぞれ O，S，H と省略して記述することにする．

3.3.2 最尤推定法による音響伝達特性の推定

本節では，クリーン音声 GMM を用いて最尤推定法の枠組みにより残響信号から音響伝達特性を推定する手法について説明する．文献 [43] では，電話音声の認識において，最尤推定法により電話回線の特性による音響ミスマッチを減らす手法を提案している．また文献 [44] では EM アルゴリズムにより音声のスペクトル歪みと加算性ノイズを同時に除去する手法を提案している．

最尤推定法の枠組みでは，残響信号に対して最も尤度が高くなるように音響伝達特性を推定する．

$$\hat{H} = \underset{H}{\operatorname{argmax}} \Pr(O|H, \lambda_S). \tag{3.12}$$

ここで λ_S はクリーン音声 GMM のモデルパラメータを表す．

$$\lambda_S = \{w_k, \mu_k^{(S)}, \mathbf{\Sigma}_k^{(S)}\}, \quad \sum_k w_k = 1 \tag{3.13}$$

w_k，μ_k，$\Sigma_k^{(S)}$ はそれぞれ混合要素 k における正規分布の重み，平均ベクトル，共分散行列 (対角共分散行列で定義) を表す．

解の推定は EM アルゴリズムによって行われる．EM アルゴリズムは 2 つのステップからなる反復法であり，Expectation ステップと呼ばれる最初のステップでは対数尤度の期待値で定義される Q 関数を計算する．

$$\begin{aligned} &Q(\hat{H}|H) \\ &= E[\log \Pr(O, c|\hat{H}, \lambda_S)|H, \lambda_S] \\ &= \sum_c \frac{\Pr(O, c|H, \lambda_S)}{\Pr(O|H, \lambda_S)} \cdot \log \Pr(O, c|\hat{H}, \lambda_S) \end{aligned} \tag{3.14}$$

c は混合要素を表す潜在変数である．残響信号 O のフレーム系列と c の同時確率は以下のように計算される．

$$\Pr(O, c|\hat{H}, \lambda_S) = \prod_n w_{c(n)} \Pr(O(n)|\hat{H}, \lambda_S) \tag{3.15}$$

$w_{c(n)}$ と $O(n)$ はフレーム n における混合要素の重みと残響信号のケプストラムを表す．ここで，式 (3.10) で示される通り，残響信号のケプストラムはクリーン音声と音響伝達特性のケプストラムの線形加算として仮定しているため，残響信号の分布の平均はクリーン音声の分布に音響伝達特性のケプストラムを加算したものとして表現することができる．

$$\begin{aligned}&\Pr(O, c|\hat{H}, \lambda_S)\\ =&\prod_n w_{c(n)} \cdot N(O(n); \mu^{(S)}_{c(n)} + \hat{H}(n),\ \Sigma^{(S)}_{c(n)})\end{aligned} \tag{3.16}$$

$N(O; \mu, \Sigma)$ は多次元正規分布を表す．これにより式 (3.14) は以下のように展開される [45].

$$\begin{aligned}&Q(\hat{H}|H)\\ &= \sum_k \sum_n \Pr(O(n), c(n) = k|\lambda_S) \log w_k\\ &\quad \sum_k \sum_n \Pr(O(n), c(n) = k|\lambda_S)\\ &\quad \cdot \log N(O(n); \mu^{(S)}_k + \hat{H}(n), \Sigma^{(S)}_k).\end{aligned} \tag{3.17}$$

共分散行列を対角行列で定義し，H に関する項のみを展開すると，以下の式が得られる．

$$\begin{aligned}Q(\hat{H}|H) = -\sum_k \sum_n \gamma_k(n) \sum_{d=1}^{D} \Bigg\{ &\frac{1}{2} \log(2\pi)^D {\sigma^{(S)}_{k,d}}^2\\ &+ \frac{(O(d;n) - \mu^{(S)}_{k,d} - \hat{H}(d;n))^2}{2{\sigma^{(S)}_{k,d}}^2} \Bigg\}\end{aligned} \tag{3.18}$$

$$\gamma_k(n) = \Pr(k|O(n), H, \lambda_S) \tag{3.19}$$

D は $O(n)$ の次元数を表し，$\mu^{(S)}_{k,d}$ 及び ${\sigma^{(S)}_{k,d}}^2$ はそれぞれ混合要素 k の平均ベクトル，対角共分散行列における d 次元目の要素と対角要素を表す．

maximization ステップと呼ばれる次のステップでは $Q(\hat{H}|H)$ を最大とする $\hat{H}$ を求め，H を更新する．この更新式は $\partial Q(\hat{H}|H)/\partial \hat{H} = 0$ を解くことで得られる．

$$\hat{H}(d;n) = \frac{\sum_k \gamma_k(n) \frac{O(d;n) - \mu^{(S)}_{k,d}}{{\sigma^{(S)}_{k,d}}^2}}{\sum_k \frac{\gamma_k(n)}{{\sigma^{(S)}_{k,d}}^2}}. \tag{3.20}$$

本研究では H の初期値は 0 ベクトルとし，解が収束したときの $\hat{H}$ を推定した音響伝達特性とし，そのフレーム方向への平均を取る．

$$\bar{H}_\theta(d) = \sum_n \hat{H}_\theta(d;n) \tag{3.21}$$

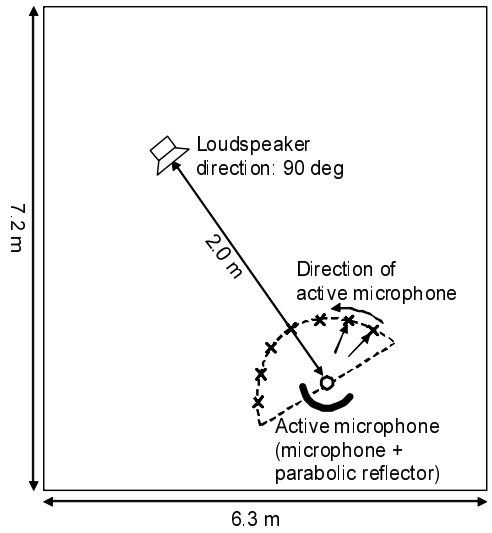

図 3.8　Experimental conditions

同様にして，全ての離散的な方向毎に音響伝達特性を推定し，式 (3.7) により音源方向を推定する．

3.4　評価実験

3.4.1　実験環境

実験は実環境下で行った．反射板の直径は 24cm，焦点距離は 9cm である．マイクロホンは無指向性マイクロホン (SONY ECM-77B) を使用した．音源として，音源方向 90°，音源距離 2m の位置にスピーカーを配置した．アクティブマイクロホンは手動で 30° から 150° へ 20° 間隔で変化させた．角度ごとに収録する時間は 0.5, 1.0, 1.5, 2.0, 2.5, 3.0sec の場合で比較した．収録部屋の大きさは 6.3 m × 7.2 m(W × D) である．Fig. 3.8 に実験環境の図を示す．

サンプリング周波数は 12kHz，分析窓はハミング窓を使用し，窓幅は 32msec，フレームシフトは 8msec である．パワーは平均が 0dB になるように正規化されている．クリーン音声 GMM の作成には，ASJ 日本語音声データベース [46] より，50 文の音声を使用した．GMM の混合数は 64 である．音響伝達特性の推定までは 16 次元の MFCC を特徴量として使用し，音源方向の推定には 2 次元の MFCC を用いた．クリーン音声 GMM の作成に用いた音声と，テストに用いた音声はそれぞれ異なる発話内容である．

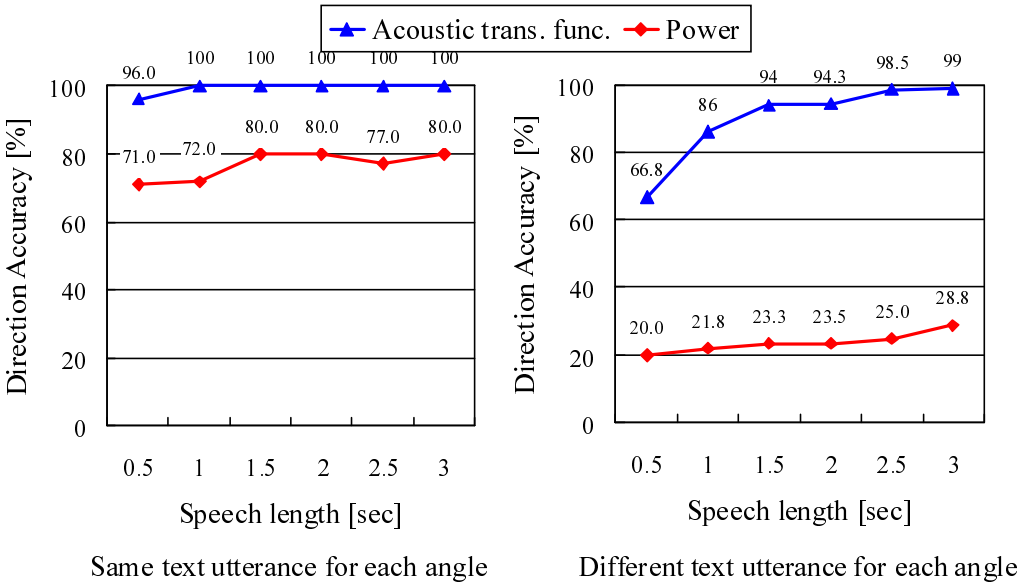

図 3.9　Performance of an active microphone with a parabolic reflection board

3.4.2　実験結果

収録する音声の長さ毎の提案手法及びパワーに基づく手法の音源方向推定精度を Fig. 3.9 に示す．収録音声の長さは 0.5, 1.0, 1.5, 2.0, 2.5, 3.0sec であり，テストデータの数はそれぞれ 600, 300, 200, 150, 120, 100 セグメントである．左図より，アクティブマイクロホンの方向毎に発話している内容が同じ場合では，提案手法もパワーに基づく手法も高い推定精度を示している．しかし，実際の環境では，アクティブマイクが回転している間，話者が同じ発話をする場面はほとんどない．

右図より，アクティブマイクロホンの方向毎に発話している内容が異なる場合，パワーに基づく手法では，精度が急激に低下していることが分かる．これは，パワーが元の発話音声に依存するため，パラボラ反射板によってパワーが増幅したにも関わらず，パワーが反射板の全方向に対して最大とならなかったためである．一方，音響伝達特性を用いる提案手法では，発話内容が異なる場合であっても高い推定精度を保っていることがわかる．これは，音響伝達特性が元のクリーン音声成分に依存せず，アクティブマイクロホンの特性にのみ依存するからである．また，収録する音声が長くなるにつれて，推定精度が向上していることが分かる．これは，収録音声が長くなるほど，その平均値ベクトルが安定するためである．

次に，パラボラ反射板を用いる代わりに，ショットガンマイク (SONY ECM-764) を用

図 3.10　Shotgun microphone (SONY ECM-674)

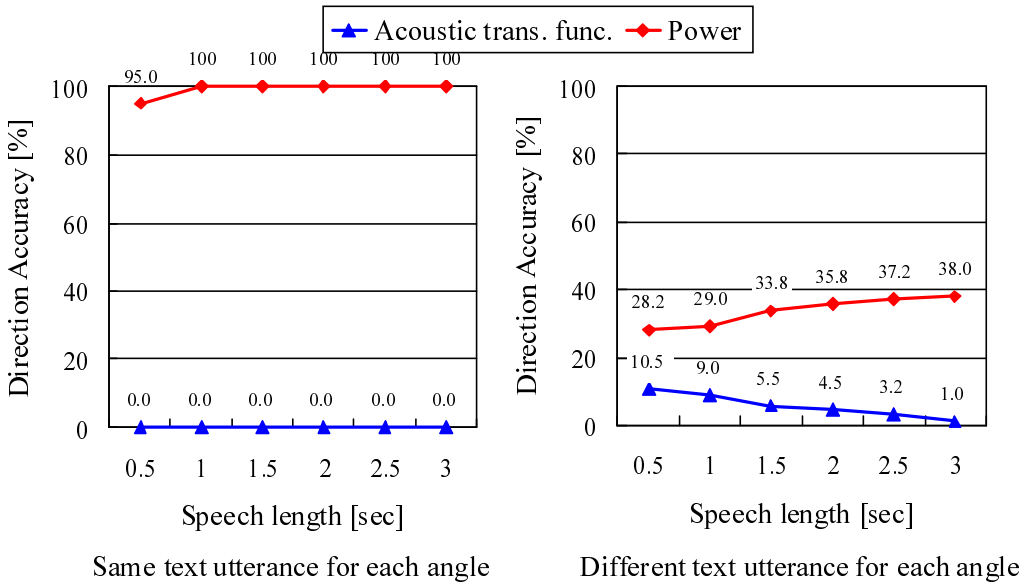

図 3.11　Performance of a shotgun microphone without a parabolic reflection board

いて同様の実験を行った．ショットガンマイクは Fig. 3.10 のような超指向型のマイクロホンであり，マイク前方からの音をよく捉え，それ以外からの音はパワーが低減される特性を持ったマイクロホンである．結果を Fig. 3.11 に示す．パワーに基づく手法においては，ショットガンマイクを用いた場合はパラボラ反射板を用いた場合と似たような結果になり，方向毎の発話内容が同じ場合ではショットガンマイクの指向性により高い性能を示すが，発話内容が異なる場合は性能が下がる．

一方，音響伝達特性を用いた手法の場合，ショットガンマイクではほとんど正しく音源

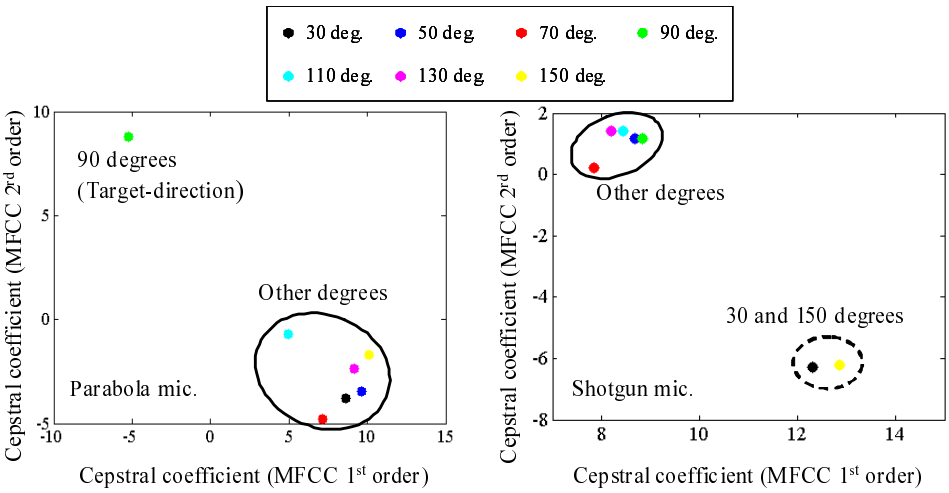

図 3.12　Mean values of the acoustic transfer functions for the microphone with a parabolic reflection board (left) and the shotgun microphone (right)

方向推定が行えていない．パラボラ反射板付きマイク，及びショットガンマイクにおける，マイクの向き毎の音響伝達特性の平均値ベクトルをプロットした図を Fig. 3.12 に示す．左図より，パラボラ反射板を用いた場合では，音源方向である 90° の音響伝達特性がそれ以外の方向の音響伝達特性から離れた位置に分布している．一方右図のショットガンマイクの場合では，マイクの横方向である 30, 150° の音響伝達特性が他の方向の音響伝達特性から離れた位置に分布している．そのため，ショットガンマイクにおいて音響伝達特性を用いた場合では 30, 150° を音源方向として出力してしまい，正しく音源方向が推定できない．

次に，提案手法の枠組みにおいて，クリーン音声 GMM を用いて音響伝達特性を推定した場合と，実際は未知であるクリーン音声情報を与えて音響伝達特性を計算した場合を比較した．Fig. 3.13 は 300 セグメントのアクティブマイクの角度毎の音響伝達特性を示した図である．左図は，式 (3.11) に本来のクリーン音声の MFCC を代入することで計算した音響伝達特性 H_{sub} である．右図はクリーン音声 GMM によって推定された音響伝達特性 H_{est} である．左図より，アクティブマイクロホンが音源方向を向いていないときは，H_{sub} はほとんど同じ位置に分布しており，音源方向を向いたときのみ，異なる位置に分布していることが分かる．しかし右図では，推定された音響伝達特性 H_{est} にばらつきが存在しており，H_{sub} のような理想的な分布にはなっていない．

Fig. 3.14 は音響伝達特性に H_{sub} を用いた場合と H_{est} の比較である．発話内容が同じ場合では，どちらの音響伝達特性も 100% の推定精度を示しているが，発話内容が異なる場合においては，H_{est} では推定精度が低下している．これは，音響伝達特性が完全には

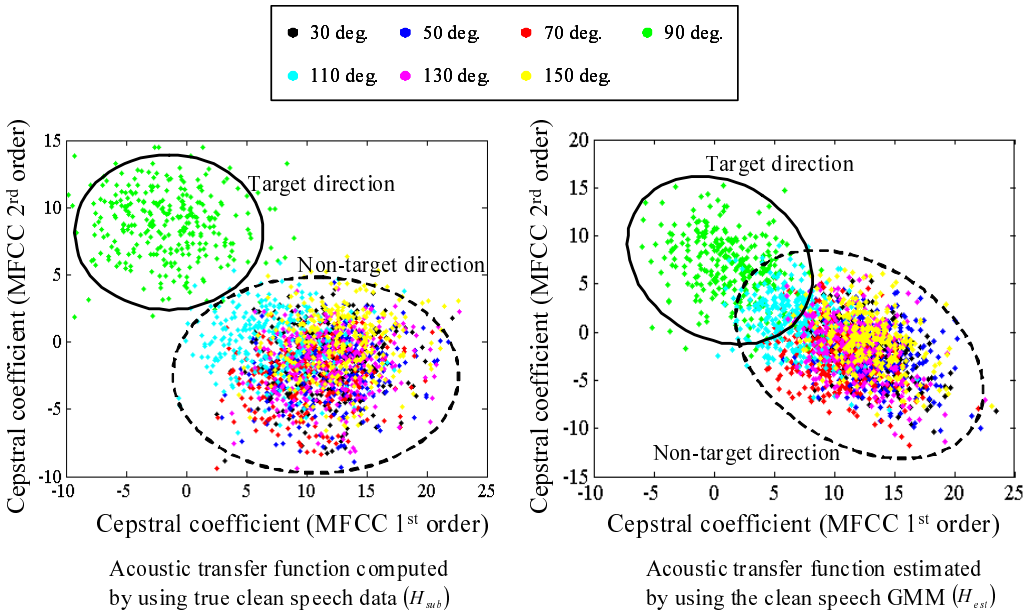

図 3.13 Acoustic transfer function computed by using true clean speech data (left) and that estimated by the proposed method using only the statistics of clean speech GMM (right) at each angle in the cepstral domain

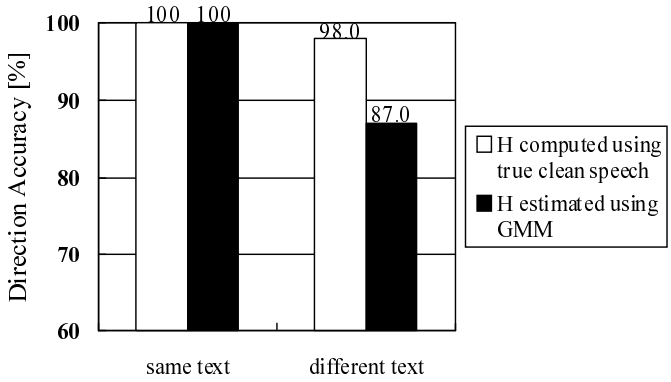

図 3.14 Comparison of true clean speech data and clean speech model

推定しきれておらず，クリーン音声の成分が若干残っているためと考えられる．

3.5　まとめ

本章では，音響伝達特性を用いた音源方向推定のアプローチの一つとして，マイクを中心に回転するパラボラ反射板を用いて，反射板の回転により変動した音響伝達特性を検出することで，音源の方向を推定する手法を提案した．パラボラ反射板を用いた信号原の方向推定として，パワーに基づく手法が他の分野では既に提案されているが，音声信号を対象とした場合では，パワーに基づく手法では十分に信号原の方向を推定することができない．一方，提案手法である音響伝達特性に基づく手法では，音響伝達特性がクリーン音声に依存しないため，パワーに基づく手法よりも高い精度で音源の方向を推定することができている．しかし，本手法では音響伝達特性が完全に推定しきれていないため，その推定誤差が音源方向推定の誤差の原因となっている．また，本研究で用いた反射板は直径 24cm と大きく，実用をする際には反射板の縮小化が必要である．反射板が小さくなった場合，焦点に集まる反射波の数も減るため，音響伝達特性の変動も小さくなると考えられる．そのため，音響伝達特性の微小な変動であっても検出できるように，音響伝達特性のより正確な推定が必要であると考えられる．

第 4 章

HMM による部屋音響伝達特性の推定及び MKL-SVM による次元重みを考慮した音源位置の識別

4.1 まえがき

本章では，第 3 章で提案するシステムとは別のアプローチとして，部屋の音響伝達特性を識別することで，音源の位置を推定する手法を提案する．口から発せられた直後のクリーン音声は，壁からの反射音（残響）やスペクトルの減衰などの，部屋音響伝達特性が畳み込まれた残響信号としてマイクに観測される．この音響伝達特性は，音源の位置に依存して変化することが知られている．この手法では，この音響伝達特性が音源位置に依存する点に着目し，音響伝達特性をあらかじめ音源位置毎に学習しておき，評価音声についても，その音響伝達特性を識別することで，音源位置を単一マイクで推定する手法を提案する．さらに，第 3 章では音響伝達特性をクリーン音声 GMM を用いて推定していたのに対して，本章では，より正確に音響伝達特性を推定するために，クリーン音声を HMM でモデル化する．GMM は静的なモデルであるため，音声の時間的変化を表現することができない．一方 HMM は複数の状態からなる状態遷移モデルであり，音声の時間的変化を表現することができる．そのため，HMM を用いた方がクリーン音声をより詳細にモデル化でき，それにより音響伝達特性の推定も正確になると考えられる．

また，音響伝達特性の特徴量ベクトルである MFCC の各次元の中には，音源位置の識別に有効な次元とそうでない次元があり，かつその次元は音源位置によって異なるという考えに基づいて，Multiple Kernel Learning (MKL)[47] を用いた音源位置毎の次元重み

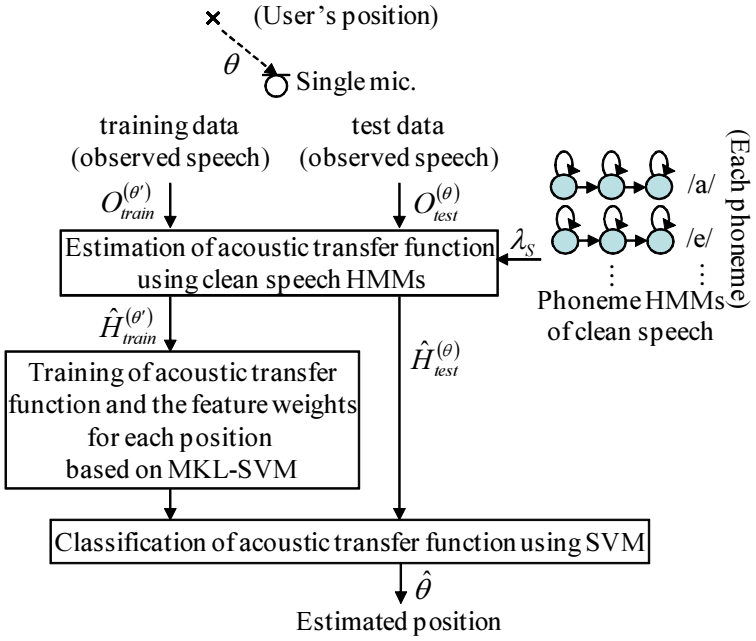

図 4.1　System overview

を学習させる手法を提案する．MKL は本来，様々なカーネル関数を統合することによって認識器の性能を強化するために用いられてきた手法であるが [48, 49, 50, 51]，本手法では MKL を MFCC の各次元の重みを求める目的で使用する．音源位置の識別実験において，MKL を組み合わせた Support Vector Machine (SVM) と通常の SVM を比較し，提案手法の有効性を示す．

4.2　音響伝達特性の推定

4.2.1　提案手法の概要

本研究では音響伝達特性を用いて音源の位置を推定している．音響伝達特性は音源の位置によって異なる値を持つため，あらかじめこれを位置毎に学習しておけば，評価音声に対してもその音響伝達特性を識別することで音源位置を推定することができる．

提案手法の概要を Fig. 4.1 に示す．まず，位置毎の音響伝達特性を学習するために，それぞれの位置 θ で発話された音声 $O^{(\theta)}_{train}$ を収録し，その音響伝達特性をクリーン音声の音素 HMM を用いて推定する．次に，位置毎に推定された音響伝達特性のケプストラム $\hat{H}^{(\theta)}_{train}$ を MKL-SVM により学習する．この際，音響伝達特性ケプストラムの次元重みも同時に学習される．そして評価したい音声 $O^{(\theta)}_{test}$ についても学習データと同様にして音響伝達特性 $\hat{H}^{(\theta)}_{test}$ を推定し，それを SVM で識別することで，音源位置 $\hat{\theta}$ を推定する．

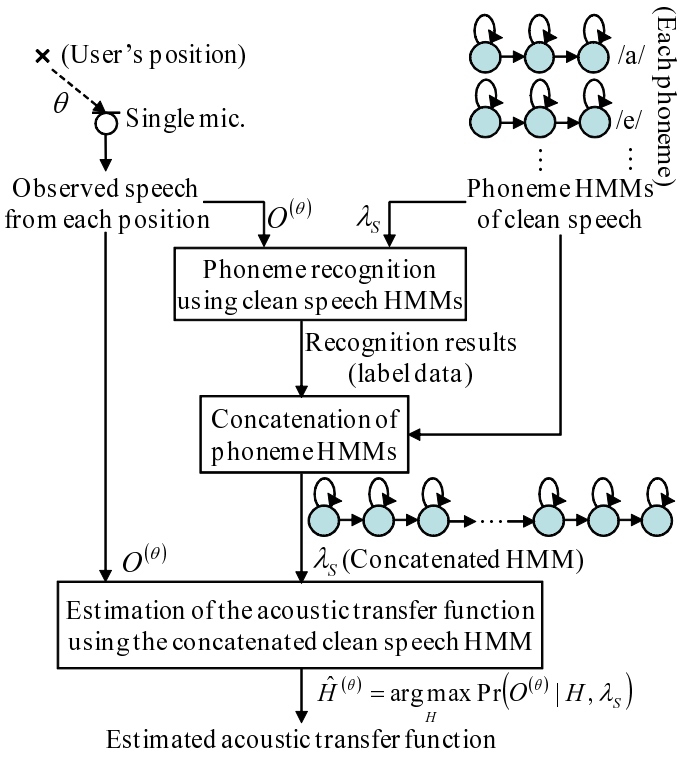

図 4.2　Estimation of the acoustic transfer function using phoneme HMMs of clean speech

音響伝達特性の推定の流れを Fig. 4.2 に示す．あらかじめ特定話者のクリーン音声の MFCC を音素 HMM でモデル化しておく．HMM を用いて音響伝達特性を推定するためには，その音声信号の音素ラベルが必要であるため，まず学習した音素 HMM を用いて観測信号を音素認識する．そして出力された音素認識結果をラベルとして音素 HMM を連結し，連結された HMM を用いて観測信号から最尤推定法により音響伝達特性の MFCC を推定する．

4.2.2　音素 HMM による音響伝達特性の推定

本節では音素 HMM を用いて観測信号 O から音響伝達特性 H を推定する手法について述べる．ある場所で発声されたクリーン音声 S は，音響伝達特性 H の影響を受けて観測される．このとき，フレーム n における観測信号 O のケプストラムは，第 3 章の 3.3.1 節と同様に

$$O_{cep}(d;n) \approx S_{cep}(d;n) + H_{cep}(d;n) \tag{4.1}$$

と近似することにする．第 3 章では S の統計モデルとして GMM を用いていたが，本章では音素毎にモデル化された音素 HMM を用いて，最尤推定法により O から H を推定

する．

静的モデルである GMM は音声の時間的変化を表現することができないが，状態遷移モデルである HMM は音声の時間的変化を表現することができる．そのため，HMM を用いた方がクリーン音声をより詳細にモデル化でき，それにより音響伝達特性の推定も正確になると考えられる．また，第 3 章では全ての音素を一つの GMM で表現していたため，ある音素の残響音声から音響伝達特性を推定しようとした際に，別の音素の情報が伝達特性の推定に影響を与えてしまう．一方本章においては HMM を音素毎にモデル化しているため，別の音素の情報が推定に影響を与える問題を防ぐことが可能となる．

しかし，HMM を用いて音響伝達特性を推定するためには，その音声信号の音素ラベルが必要であるため，まず学習した音素 HMM を用いて観測信号を音素認識する．そして出力された音素認識結果をラベルとして音素 HMM を連結し，連結された HMM を用いて観測信号から最尤推定法により音響伝達特性の MFCC を推定する．

$$\hat{H} = \underset{H}{\operatorname{argmax}} \Pr(O|\lambda_S, H) \tag{4.2}$$

λ_S はクリーン音声のモデルパラメータを表す．(4.2) 式の解は第 3 章と同様に EM アルゴリズムによって推定される．その際，Q 関数は以下のように定義される．

$$\begin{aligned} &Q(\hat{H}|H) \\ &= E[\log \Pr(O, p, b_p, c_p|\hat{H}, \lambda_S)|H, \lambda_S] \\ &= \textstyle\sum_p \sum_{b_p} \sum_{c_p} \frac{\Pr(O,p,b_p,c_p|H,\lambda_S)}{\Pr(O|H,\lambda_S)} \\ &\quad \cdot \log \Pr(O, p, b_p, c_p|\hat{H}, \lambda_S) \end{aligned} \tag{4.3}$$

b_p と c_p はそれぞれ音素 p における HMM の状態，混合要素を表す．O，p，b，c の同時確率 $\Pr(O, p, b_p, c_p|\hat{H}, \lambda_S)$ は以下のように展開される．

$$\begin{aligned} &\Pr(O, p, b_p, c_p|\hat{H}, \lambda_S) \\ &= \textstyle\prod_n a_{b_p(n-1),b_p(n)} w_{b_p(n),c_p(n)} \\ &\quad \cdot \Pr(O(n)|p, b_p(n), c_p(n); \hat{H}, \lambda_S) \end{aligned} \tag{4.4}$$

n，a，w はそれぞれフレーム番号，状態遷移確率，混合重みを表す．ここで，(4.1) 式より O は S と H の加算とみなされるため，O の事後確率をクリーン音声 HMM を用いて以下のように表すことができる．

$$\begin{aligned} &\Pr(O, p, b_p, c_p|\hat{H}, \lambda_S) \\ &= \textstyle\prod_n a_{b(n-1),b(n)} w_{b(n),c(n)} \\ &\quad \cdot N(O(n); \mu^{(S)}_{p,j,k} + \hat{H}_{(n)}, \Sigma^{(S)}_{p,j,k}) \end{aligned} \tag{4.5}$$

$N(O;\mu,\Sigma)$ は多次元正規分布を表し，$\mu^{(S)}_{p,j,k}$，$\Sigma^{(S)}_{p,j,k}$ はそれぞれ S の状態 $b(n)=j$，混合要素 $c(n)=k$ における平均ベクトルと共分散行列 (対角行列) を表す．これらを用いて (4.3) 式を展開すると，

$$\begin{aligned}
&Q(\hat{H}|H)\\
&=\textstyle\sum_p\sum_i\sum_j\sum_n\\
&\quad \Pr(O(n),p,b_p(n)=j,b_p(n-1)=i|H,\lambda_S)\log a_{p,i,j}\\
&+\textstyle\sum_p\sum_j\sum_k\sum_n\\
&\quad \Pr(O(n),p,b_p(n)=j,c_p(n)=k|H,\lambda_S)\log w_{p,j,k}\\
&+\textstyle\sum_p\sum_j\sum_k\sum_n\\
&\quad \Pr(O(n),p,b_p(n)=j,c_p(n)=k|H,\lambda_S)\\
&\quad \cdot\log N(O(n);\mu^{(S)}_{p,j,k}+\hat{H}(n),\Sigma^{(S)}_{p,j,k})
\end{aligned}\tag{4.6}$$

となり，H に関わる項のみを取り出すと以下のようになる．

$$\begin{aligned}
&Q(\hat{H}|H)\\
&=-\textstyle\sum_p\sum_j\sum_k\sum_n\gamma_{p,j,k}(n)\\
&\quad -\textstyle\sum_{d=1}^{D}\left\{\frac{1}{2}\log(2\pi)^D{\sigma^{(S)}_{p,j,k,d}}^2\right.\\
&\quad \left.+\frac{(O(d;n)-\mu^{(S)}_{p,j,k,d}-\hat{H}(d;n))^2}{2{\sigma^{(S)}_{p,j,k,d}}^2}\right\}
\end{aligned}\tag{4.7}$$

$$\gamma_{p,j,k}(n)=\Pr(O(n),p,j,k|\lambda_S)\tag{4.8}$$

D は次元数，$\mu^{(S)}_{p,j,k,d}$，${\sigma^{(S)}_{p,j,k,d}}^2$ はそれぞれ平均ベクトルの d 次元目の値と，共分散行列の d 番目の対角要素の値を表す．(4.7) 式を最大にする H は，$\partial Q(\hat{H}|H)/\partial\hat{H}=0$ を解くことで求められる．

$$\hat{H}(d;n)=\frac{\sum_p\sum_j\sum_k\gamma_{p,j,k}(n)\frac{O(d;n)-\mu^{(S)}_{p,j,k,d}}{{\sigma^{(S)}_{p,j,k,d}}^2}}{\sum_p\sum_j\sum_k\frac{\gamma_{p,j,k}(n)}{{\sigma^{(S)}_{p,j,k,d}}^2}}.\tag{4.9}$$

4.3 MKL-SVM による音響伝達特性の次元重み学習及び識別

本節では，MKL-SVM による音響伝達特性の次元重みの学習方法と，識別方法について述べる．本章における音源位置推定手法では，まず音源位置 θ 毎に推定された音響伝達特性の MFCC を用いて，SVM で位置の学習を行う．そして，音源位置が不明な評価音

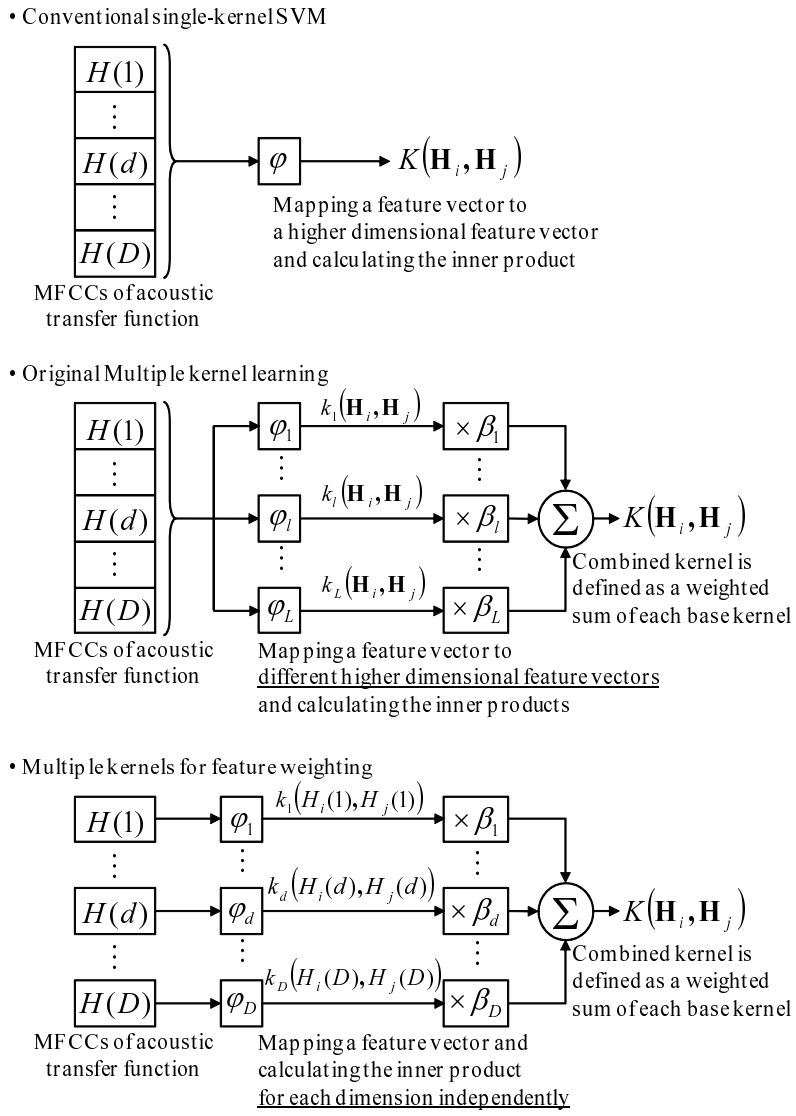

図 4.3　A conventional SVM with single-kernel, original MKL for SVMs and a new weighting method based on MKL

声についても，その推定された音響伝達特性の MFCC を識別することで，位置の推定を行う．

その際，音響伝達特性 MFCC の中にはその位置のインパルス応答の影響を強く受ける次元と，影響を受けにくい次元が存在すると考えられる．また影響を受ける次元は，音源の位置によって多少のばらつきがあると考えられる．そこで本研究では，MKL (Multiple Kernel Learning) により，音響伝達特性 MFCC の次元重みを位置毎に学習する手法を提案する．

MKL[47] は，複数のサブカーネルを線形結合して新たなカーネルを作成することで，

より複雑な非線形空間を作成する手法である．これを用いて，サンプル i, j の音響伝達特性 MFCC $\mathbf{H}_i, \mathbf{H}_j$ より計算されるカーネル関数は，以下のように表現される．

$$k(\mathbf{H}_i, \mathbf{H}_j) = \sum_l \beta_l k_l(\mathbf{H}_i, \mathbf{H}_j) \tag{4.10}$$

β_l は l 番目のサブカーネル k_l の重みである．

MKL-SVM は本来，それぞれのサブカーネルを識別器とみなし，それらを統合することで，通常の SVM の識別能力を向上させることを目的として，音声の話者認識 [48, 49] や画像認識 [50, 51] の分野などで用いられてきた．しかし近年，画像の一般物体認識の分野などでは，MKL-SVM を用いて特徴選択を行う手法も提案されている [52, 53]．この手法では，複数の特徴量を用いた画像識別において，サブカーネルを特徴量ごとに定義することで，識別に適した特徴重みを MKL により学習させている．本研究では，MFCC の次元毎にサブカーネルを定義し，MKL により重みを学習させる．

$$k(\mathbf{H}_i, \mathbf{H}_j) = \sum_d \beta_d k_d(H_i(d), H_j(d)) \tag{4.11}$$

従来の SVM における単一カーネルと，通常の MKL-SVM におけるマルチカーネル，そして本手法における次元重み学習のためのマルチカーネルの図を Fig. 4.3 に示す．従来の単一カーネル SVM（上図）ではある特徴量ベクトルに対して一つのカーネル写像関数が適用されているのに対し，MKL-SVM（中図）においては特徴量ベクトルに対して複数のカーネル関数を適用し，その重み付き和によって新たなカーネル関数を定義している．そして提案法におけるマルチカーネル（下図）では，特徴量ベクトルの 1 次元に対して 1 つのカーネル関数を適用させ，それらの重み付き和によってカーネル関数を定義している．特徴ベクトルの次元毎に独立してサブカーネルを計算させた場合，次元間の相関関係を表す情報は失われてしまう．しかし MFCC は次元の相関性が弱いため，次元毎にサブカーネルを定義しても識別能力に大きく影響はしないと考えられる．

MKL の重み β_d の学習は，SVM の枠組み，すなわちマージン最大化の枠組みで解かれるのが一般的である [47].

$$\begin{aligned} &\min_{\beta, w_d, b, \xi} \tfrac{1}{2} \textstyle\sum_d \frac{1}{\beta_d} ||w_d||_2^2 + C \sum_i \xi_i \\ &s.t. \begin{cases} y_i \Big(\sum_d w_d^T \phi_d\big(H_i(d)\big) + b \Big) \geq 1 - \xi_i \ \forall i \\ \xi_i \geq 0 \ \forall i, \quad \beta_d \geq 0 \ \forall k, \quad \sum_d \beta_d = 1 \end{cases} \end{aligned} \tag{4.12}$$

ここで ϕ は高次元空間への写像関数を表し，y_i はクラスを表す変数 $(-1, 1)$，ξ はスラック変数，C はマージンと学習データの誤り率とのトレードオフを決定する変数である．

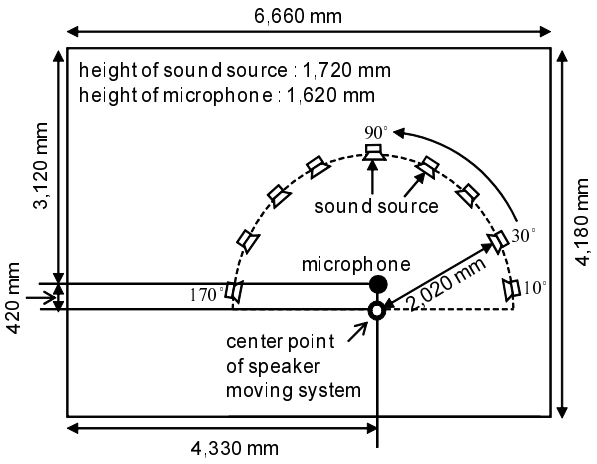

図 4.4　Experimental room environment

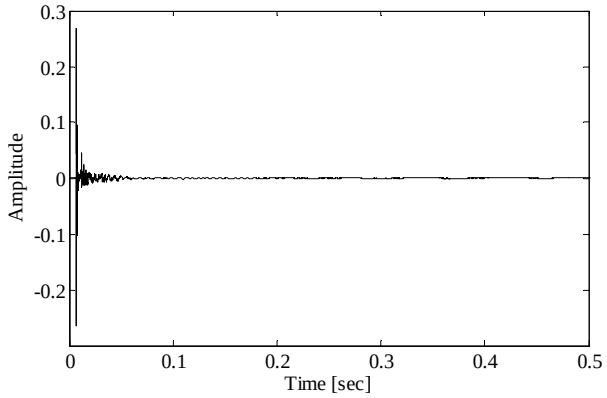

図 4.5　Impulse response (90 degrees, reverberation time: 300 msec).

(4.12) 式に対する双対問題は以下のように導出される．

$$\max_{\alpha,\beta} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \sum_d \beta_d k_d(H_i(d), H_j(d))$$
$$s.t. \begin{cases} \sum_i y_i \alpha_i = 0, & 0 \leq \alpha_i \leq C \\ \sum_d \beta_d = 1, & \beta_d \geq 0 \end{cases} \tag{4.13}$$

α_i はラグランジュ係数である．(4.13) 式を満たす α_i，β_d は 2 ステップの反復による解法を用いて求められる．まず第一ステップでは β_d を固定して α_i を通常の SVM の解法により更新する．そして第 2 ステップでは α_i を固定して β_d を更新する．本手法では，α_i の更新には SVM^{light}[54] を用い，β_d の更新には projected-gradient[55] を用いた．これらのステップを繰り返すことにより，特徴次元の重みと，識別境界が同時に学習される．

表 4.1 Mean square error of the acoustic transfer function separated using a clean speech GMM, clean speech HMMs with the 1-best hypothesis and HMMs with the correct transcription. The MSE of the observed speech was calculated by substituting $O(d;n)$ for $\hat{H}(d;n)$ in Eq. (4.14).

	Observed speech	GMM	HMM (1-best hypothesis)	HMM (correct transcription)
MSE	9485.97	2264.33	2096.14	1968.36

4.4 評価実験

4.4.1 実験環境

提案手法を評価するために特定話者によるシミュレーション実験を行った．音声データは ATR 研究用日本語音声データベースセット A[56] より男性話者 1 名の単語音声を用い，RWCP 実環境音声・音響データベース [57] において収録されたインパルス応答を畳み込むことで，それぞれの位置における残響音声を作成した．Fig. 4.4 にインパルス応答の収録環境を，Fig. 4.5 に音源位置 90° のインパルス応答を示す．残響時間は 300 msec である．約 6.7 m × 4.2 m の部屋内のある点を中心に，半径 2,020 mm の半円上に音源の位置が存在しており，マイクは中心点から 420 mm 離れた点に位置する．

音声信号はサンプリング周波数 12 kHz，窓幅 32 msec，フレームシフト 8 msec の分析条件で MFCC 16 次元を特徴量として使用した．音響伝達特性の推定におけるクリーン音声の音素 HMM は，2,620 単語を用いて学習した．音素数は 54，各音素 HMM の状態数は 3，混合数は 32 である．音響伝達特性の学習には 50 単語を，評価には 1,000 単語を用いた．なお，クリーン音声の学習，位置の学習，評価に用いたデータはそれぞれ異なる発話内容の単語を使用している．

4.4.2 音響伝達特性の推定精度の評価

位置推定性能を評価する前に，予備実験として音響伝達特性の推定精度をクリーン音声のモデル化に GMM を用いた場合と音素 HMM を用いた場合で比較した．クリーン音声 GMM は HMM と同様の 2,620 単語を用いて学習しており，混合数は 64 である．Table 4.1 に各クリーン音声モデルによって推定された音響伝達特性 $\hat{H}(d;n)$ と本来の音響伝達特性 $H_{sub}(d;n)$ との平均二乗誤差 (MSE: Mean Square Error) を示す．MSE は各音源位置におけるテストデータ 1000 単語に対して，フレーム毎に二乗誤差を計算し，

全フレーム，位置で平均して求めている．

$$\mathrm{MSE} = \frac{1}{N}\sum_{n}\sum_{d}(H_{sub}(d;n) - \hat{H}(d;n))^2, \tag{4.14}$$

ここで $H_{sub}(d;n)$ は第 3 章の実験（3.4.2 節）で説明した通り，式 (3.11) に本来未知であるクリーン音声の MFCC を与えて計算した本来の音響伝達特性である．Table. 4.1 において，Observed speech は式 (4.14) の $\hat{H}(d;n)$ の代わりに観測信号の MFCC $O(d;n)$ を代入して求めた，観測信号と本来の音響伝達特性との誤差である．GMM はクリーン音声 GMM を用いて推定した音響伝達特性の MSE を表す．HMM(1-best hypothesis) は音素 HMM による音響伝達特性の推定において，音素認識結果の 1-best を音素ラベルとして音素 HMM を連結し，音響伝達特性を推定した場合の MSE である．一方 HMM(correct transcription) は正解の音素ラベルを用いて音素 HMM を連結し，音響伝達特性を推定した場合の MSE である．Table 4.1 より，クリーン音声のモデルに GMM を用いるよりも音素 HMM を用いた方が，音素ラベルに音素認識結果を用いた場合でも推定誤差が小さくなっていることが分かる．

ある入力フレームにおける本来の音響伝達特性と，各手法で推定した音響伝達特性，そして観測信号のメルスペクトルをプロットした図を Fig. 4.6 に示す．メルスペクトルは MFCC に逆離散コサイン変換を適用することで得られる．このとき，16 次元 MFCC の 17 次元目以降を 0 で埋めて 32 次元にした上で逆離散コサイン変換を適用し，32 次元のメルスペクトルに変換している．また，本研究において使用していない，MFCC の 0 次元目も 0 を代入しているため，得られるメルスペクトルのパワーは総和が 1 になるように正規化された形となっている．Fig. 4.6 の上図は観測信号のメルスペクトルを含めた図を，下図は音響伝達特性のメルスペクトルのみを拡大して表示した図である．図より，GMM よりも音素 HMM を用いた方がより正解の音響伝達特性に近くなっていることが分かる．

4.4.3　音源位置推定性能の評価

音響伝達特性の識別手法として，従来の単一カーネル SVM と提案手法である MKL-SVM を用いて比較を行った．カーネル関数として Gaussian kernel

$$k(\mathbf{H}_i, \mathbf{H}_j) = exp(-\gamma||\mathbf{H}_i - \mathbf{H}_j||^2) \tag{4.15}$$

を用い，(4.13) 式における C の値は 1 とした．MKL-SVM を用いる提案手法では，次元毎に定義する Gaussian kernel のパラメータ γ を同一のものとした場合と，次元毎に異なるパラメータを用いた場合の 2 種類で実験を行った．これは，特徴ベクトルの MFCC が次元無相関であるため，識別に最適なカーネルのパラメータは次元によって異なるかもし

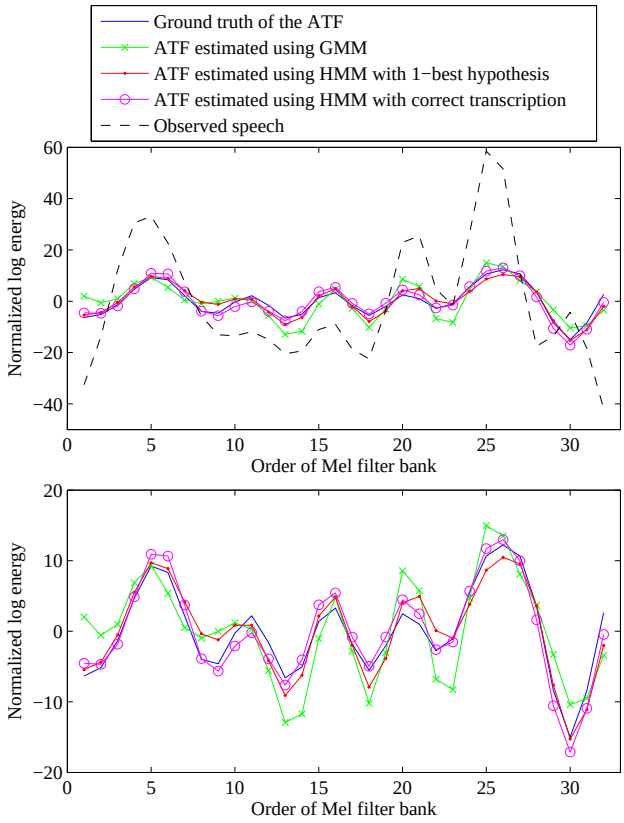

図 4.6　Mel spectra of the ground truth of the acoustic transfer function (ATF), estimated acoustic transfer functions and observed speech of a sample frame (top figure). The bottom figure is a close-up of the estimated acoustic transfer functions.

れないと考えたためである．これら 2 種類の提案手法，及び従来の SVM のカーネルのパラメータは実験的に定めた．

まず位置の数を 3 箇所 (30°，90°，130°) として，3 箇所の位置推定の性能を比較した．各比較手法における，位置毎の True positive rate とその平均 (すなわち認識率を表す) を Table 4.2 に，False positive rate とその平均を Table 4.3 に示す．MKL-SVM は単一カーネル SVM よりも高い認識率を示しており，その差はカイ二乗検定において有意 ($P \ll 0.01$) であった．また，MKL-SVM において次元毎に異なるカーネルのパラメータを設定した場合，次元毎に同一のカーネルを定義した場合に対して若干の認識率の向上が見られたが，この差は有意ではなかった ($P = 0.11$)．そのため以降の実験では，単一カーネル SVM と次元毎に同一のカーネルを定義した MKL-SVM の 2 種類の手法のみで比較を行うことにする．

提案手法では，SVM を用いてマルチクラスの識別を行うために，one-vs-rest 法を用いてクラス (位置) 毎に識別境界を学習している．一方，識別境界を学習する度に特徴次元

表 4.2　True positive rate [%] of comparison methods for each position and average of true positive rates (accuracy)

	30°	90°	130°	Average (accuracy)
SVM with single kernel	95.2	79.4	82.4	85.7
MKL-SVM with an identical kernel dimensionally	92.1	90.8	87.0	90.0
MKL-SVM with different kernels dimensionally	92.8	96.8	84.1	91.2

表 4.3　False positive rate [%] of comparison methods for each position and average of false positive rates

	30°	90°	130°	Average (accuracy)
SVM with single kernel	20.3	0.2	22.5	14.3
MKL-SVM with an identical kernel dimensionally	13.0	1.5	15.6	10.0
MKL-SVM with different kernels dimensionally	14.1	3.0	9.0	8.8

の重みも学習されるため，結果として識別境界と同じ数の種類だけ特徴次元の重みが得られる．これは，音源位置毎に最適な次元重みを学習していることを意味している．

Table 4.4 は次元毎に同一のカーネルパラメータを設定した提案手法を用いて得られた，音源位置毎の次元重みを表している．また Fig. 4.7 は，(4.1) 式へ実際のクリーン音声 $S_{cep}(d;n)$ を代入して得られた $H_{cep}(d;n)$ の単語毎のフレーム平均値を，位置毎にプロットした図である．これらの表と図を見ると，ある位置において高い重みが得られている次元では，音響伝達特性がその位置を識別しやすい分布をしていることが分かる．例えば Table 4.4 では，90° において 7 次元目が最も高い重みを得ており，一方 Fig. 4.7 では，7 次元目が 90° の音響伝達特性を判別されやすい分布になっている．同様に 30° では 10 次元目が，130° では 8 次元目が高い重みを得ており，それぞれの位置を判別しやすい分布になっている．一方，1 次元目はいずれの位置においても重みがほぼ 0 となっており，Fig. 4.7 を見ると 1 次元目の値はほとんど位置の違いの影響が現れていないことが分かる．これらのことから，それぞれの位置において，その位置の音響伝達特性を判別しやす

表 4.4 Feature weights for some cepstral dimensions trained using MKL for each position. **Bold type** shows the highest weight for the position.

Degree\order	1^{st}	4^{th}	7^{th}	8^{th}	10^{th}
30 deg	0.00	0.07	0.07	0.07	**0.08**
90 deg	0.00	0.06	**0.10**	0.07	0.07
130 deg	0.01	0.07	0.06	**0.11**	0.07

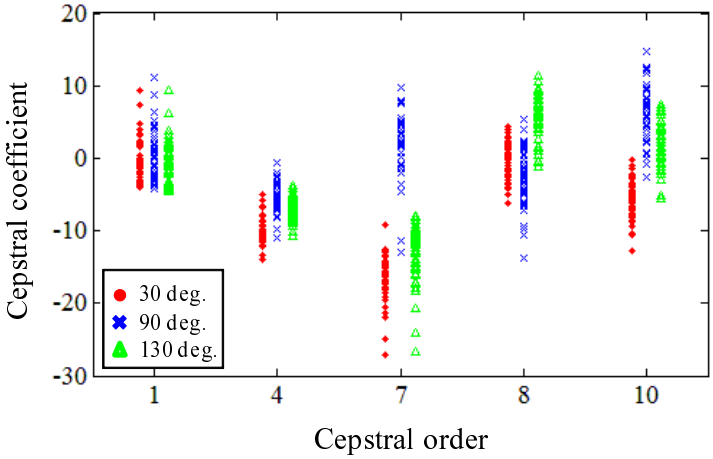

図 4.7 Mean acoustic transfer function values for some cepstral dimensions

い次元に対する重みが，MKL によって学習できていることが分かる．

次に，位置毎の音響伝達特性の学習データ数を 50 単語から 20 単語，5 単語と減らし，音源位置推定性能の変化を評価した．それぞれの音源位置の認識率を Fig. 4.8 に示す．20 単語の場合，5 単語の場合で，それぞれ 50 単語の場合と比較して 4.7%，8.9% の認識率の低下が見られた．提案手法では音響伝達特性をクリーン音声 HMM を用いて推定しており，そこには推定誤差が含まれている．そのため，学習データが少ない場合ではこの推定誤差が音源定位の性能の劣化を引き起こしたと考えられる．より少量の学習データで位置推定を行うためには，より正確な音響伝達特性の推定が必要となる．以降の実験では位置毎の学習データ数は 50 単語に固定して行う．

次に，音源の位置を 3 箇所から 5 箇所 (10°，50°，90°，130°，170°)，7 箇所 (30°，50°，70°，...，130°，150°) と増やし，音源位置推定性能を評価した．それぞれの音源位置の認識率を Fig. 4.9 に示す．位置が増えるにつれて位置の認識率が低下していることが分かるが，この理由として二つの要因が考えられる．一つは単純にクラス数が増えたことによっ

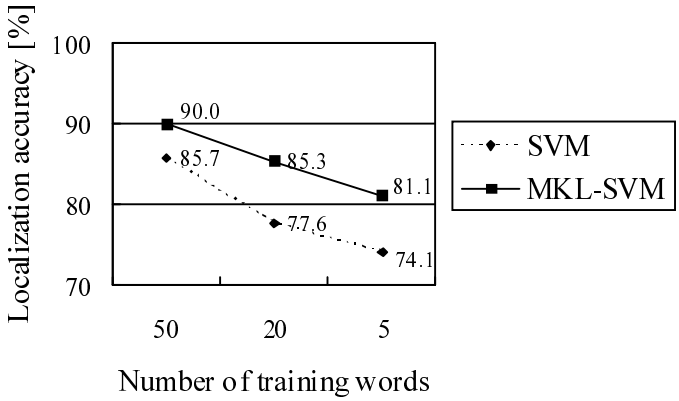

図 4.8　Localization accuracies [%] as a function of the number of training data (words)

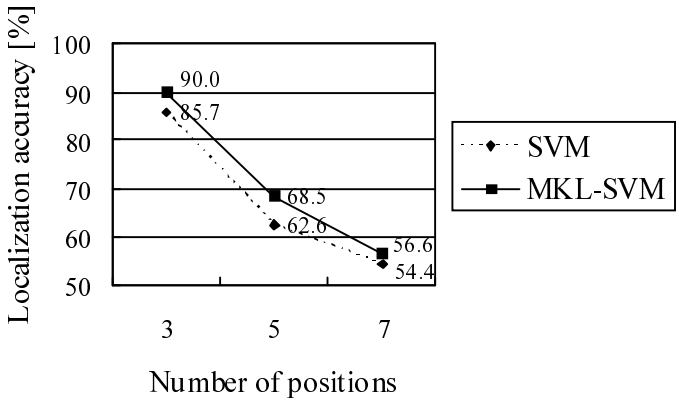

図 4.9　Localization accuracies [%] as a function of the number of positions

て認識が困難になるという，パターン認識の分野に共通する要因である．二つ目は位置が増えたことによって位置の間隔が狭まったことで位置の区別が困難になったという要因である．

そこで，二つ目の考えられる要因である，位置同士の間隔の影響を調べるため，3 種類の位置の組み合わせで音源位置推定性能を評価した．一つ目の組み合わせは 30°，90°，130°，二つ目は 10°，30°，50°，三つ目は 70°，90°，110° のそれぞれ 3 箇所である．二つ目，三つ目の組み合わせは一つ目の組み合わせと比べて位置の間隔が狭く，また二つ目の組み合わせでは壁から近い 3 箇所を，三つ目の組み合わせでは壁から遠い 3 箇所を選択している．

表 4.5 Localization accuracies [%] for each set of positions

	SVM	MKL-SVM
30, 90, 130 deg	85.7	90.0
10, 30, 50 deg	58.0	68.7
70, 90, 110 deg	42.8	54.9

表 4.6 Confusion matrix [%], where the testing position (Actual) that was not pre-trained is estimated as the most likely position from the set of pre-trained positions (Predicted).

	Predicted					
Actual	**degree**	**10**	**50**	**90**	**130**	**170**
	30	32.7	35.2	0.2	17.9	14.0
	70	5.2	42.3	26.7	23.9	1.9
	110	1.2	23.9	51.8	18.7	4.4
	150	4.4	19.0	1.7	40.0	34.9

それぞれの音源位置の認識率を Table 4.5 に示す．位置の間隔が狭い場合，音源位置の認識率が低下することが分かる．また，壁から遠い位置の組み合わせの場合，さらに認識率が低下している．これは，位置が壁から遠い場合，位置の違いによるインパルス応答の変化が生じにくいためと考えられる．

本手法では，位置をあらかじめ学習しておき，評価したい音声に対して，学習した位置の中から最も似ている位置を出力している．このとき，もし評価したい音声が事前に学習していない位置から到来された場合，どのような推定結果になるのかを実験した．この実験では，10°，50°，90°，130°，170° の 5 箇所の位置を学習しておき，学習位置に含まれない 30°，70°，110°，150° の 4 箇所の位置の音声を入力し，その認識結果を評価する．それぞれの未学習位置の音声について，どの学習位置に分類されたかの割合 (%) を表す confusion matrix を Table 4.6 に示す．

Table 4.6 より，未学習の位置の音声を入力した場合，比較的その位置に近い位置が出力される傾向にあることが分かる．このことから，未学習の位置は，そこから近い既学習位置とある程度の相関性を持っていると考えられる．そのため，今後，似ている位置の情報を使って未学習位置の情報を補間することも可能かもしれないと考えられる．

次に，2 種類の雑音環境下における 3 箇所 (30°，90°，130°) の位置推定の性能を評価

表 4.7　Localization accuracies [%] for a noisy environment and multi-talker situation.

	SVM	MKL-SVM
clean	85.7	90.0
whitenoise	45.6	49.4
multi-talker	40.2	46.0

表 4.8　Confusion matrix [%] for multi-talker situation, where a non-target speaker spoke at a position of 30 degrees.

	Predicted			
Actual	**degree**	**30**	**90**	**130**
	30	98.9	1.1	0.0
	90	63.6	36.7	0.3
	130	94.2	3.4	2.4

した．一方の雑音環境では，テストデータの残響音声に音圧レベル 40dB のホワイトノイズを足し合わせて作成した雑音重畳音声を用いて評価した．もう一方の雑音環境では，異なる話者が 30° の位置で発話した音声を，テストデータに足し合わせて作成した 2 話者の混合音声を用いて評価した．このとき目的音声の音圧レベルは平均 54.0dB，雑音となる話者の音声の音圧レベルは平均 59.3dB であった．このときの音源位置の認識率を Table 4.7 に示す．表より，雑音のない環境と比較して，ホワイトノイズを重畳した場合で 40.6%，別の話者の音声を重畳した場合で 44.0% の認識率の低下が見られた．

別の話者の音声を重畳した場合における，各入力位置がそれぞれどの位置に認識されたのかを表す Confusion matrix を Table. 4.8 に示す．表より，全ての位置において，もう一方の話者の位置である 30° が出力される傾向が高いことが分かる．これらの問題を解決するためには，雑音除去や雑音モデル適応 [58] などの処理が必要になると考えられる．

本手法では，位置毎の音響伝達特性を学習するために，実際にその位置でユーザが発話した音声を用いている．また，音響伝達特性を推定するために，ユーザのクリーン音声から学習された特定話者の音声 HMM を用いている．しかし実際の環境ではクリーン音声 HMM の学習，位置の学習のために十分な量のユーザの発話を収録するのは，ユーザにとって負担がかかるため困難である．

音声認識の分野においては，ユーザの音声を学習に用いない手法として，不特定話者 HMM と，適応 HMM[59] が存在する．不特定話者 HMM とは，ユーザを含まない，不特

定多数の音声を用いて学習した音声 HMM であり，一般的に実用されている音声認識システムでは不特定話者 HMM が多く用いられている．一方適応 HMM とは，少量のユーザの発話データを用いて，不特定話者 HMM をユーザの発話に適応させた HMM のことで，特定話者 HMM に比べて必要なユーザの発話が少量で，かつ不特定話者 HMM よりも高い認識率が得られることが知られている．

そこで，本手法における音響伝達特性の推定について，不特定話者 HMM を用いた場合と，ユーザの適応 HMM を用いた場合で比較を行った．この実験では，テストデータの話者を含まない，男女それぞれ 4 人の音声を用いて不特定話者 HMM を学習している．学習データの数は 2,620 単語 (×8 人) である．適応 HMM の作成には，テストデータの話者と同一話者の音声 50 単語を用いて不特定話者 HMM を適応させて作成した．モデル適応の手法は Maximum Likelihood Linear Regression (MLLR)[59] を用いた．

不特定話者 HMM，あるいは適応 HMM を用いて推定された音響伝達特性を用いたときの，3 箇所の音源位置認識率を Table. 4.9 に示す．この実験において，位置の学習に用いるデータと，テストデータは同一話者の音声である．特定話者 HMM を用いた場合に比べて，不特定話者 HMM や適応話者 HMM を用いた場合では性能が低下するが，8 割以上の認識率で 3 箇所の位置推定が行えている．この実験においては，不特定話者 HMM と適応 HMM の性能の差はほとんど見られなかった．不特定話者 HMM を用いた場合，推定される音響伝達特性には話者のミスマッチによる誤差が多く含まれていると考えられる．しかし，この実験ではテストデータと同一話者の音声を用いて位置を学習しているため，話者のミスマッチによる悪影響が抑えられたと考えられる．

話者のミスマッチについて検証するため，さらに音源位置の学習に用いるデータと，テストデータが異なる話者の音声の場合についても同様に評価した．この場合における認識率を Table. 4.10 に示す．表より，不特定話者 HMM の場合では適応 HMM に比べて性能が低下していることが分かる．これは，位置の学習とテストで話者が異なるため，不特定話者 HMM を用いたときの，話者のミスマッチによる悪影響が大きく現れたためと考えられる．

4.5 まとめ

本章では，部屋の音響伝達特性が音源の位置に依存する点に着目し，音響伝達特性を識別することで，音源の位置を推定する手法を提案した．本章ではより正確に音響伝達特性を推定するために，クリーン音声を HMM でモデル化し，音響伝達特性の推定に用いた．実験により，第 3 章で用いていたクリーン音声 GMM よりも音響伝達特性の推定誤差が小さくなることが示された．また，位置の推定の際には，MFCC の次元毎にカーネル関数を定義することにより，特徴次元の重みを MKL により学習させた．提案手法では音源

表 4.9　Localization accuracies [%] using the speaker-independent HMM and speaker-adapted HMM for estimating the acoustic transfer function. (The speech data for training the position and testing were uttered by the same speaker.)

	SVM	MKL-SVM
speaker-dependent HMM	85.7	90.0
speaker-independent HMM	80.3	81.7
speaker-adapted HMM	80.7	82.1

表 4.10　Localization accuracies [%] using the speaker-independent HMM and speaker-adapted HMM for estimating the acoustic transfer function. (The speech data for training the position and testing were uttered by different speakers.)

	SVM	MKL-SVM
speaker-independent HMM	58.6	58.8
speaker-adapted HMM	73.1	71.7

位置毎に，異なる次元重みのセットを学習することができ，従来の単一カーネル SVM よりも高い識別精度を得ることができた．

この手法は第 3 章で提案した手法とは異なり，反射板を用意する必要がなく，マイク一つのみで実装することができる．一方，推定したい位置毎にあらかじめ学習する必要があるため，今後は周辺の位置情報を用いた未学習の位置の補間を検討する．また，雑音環境下や，別の話者など，学習時と評価時の環境が異なる場合について，よりロバストな音源位置推定手法についても検討する．

第 5 章

音響モデル合成を用いた単一マイクによる 2 話者位置推定

5.1 まえがき

本章では，第 4 章で提案したシングルチャネル音源位置推定の枠組みを元に，音源が複数ある場合の位置推定問題への拡張手法を提案する．第 4 章で述べた手法は一人の話者のみが発話していることを前提としており，複数の話者が同時に発話した場合の位置推定が困難であった．話者の位置が固定されている場合や，一定以上の長さの発話データが得られる場合であれば，この手法でも話者オーバラップのない単一音声区間を検出することで，その音源位置を推定することは可能である．しかし，話者が移動している場合や，単一音声区間のデータが得られない場合，複数話者の混合音声からそれぞれの話者の位置を推定しなければならない．

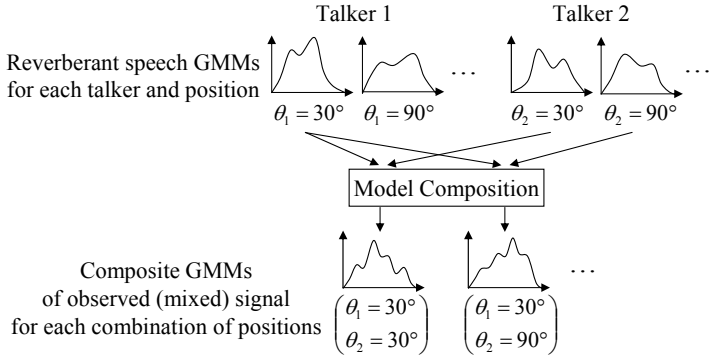

図 5.1 Training of mixed speech models of two talkers using a model composition

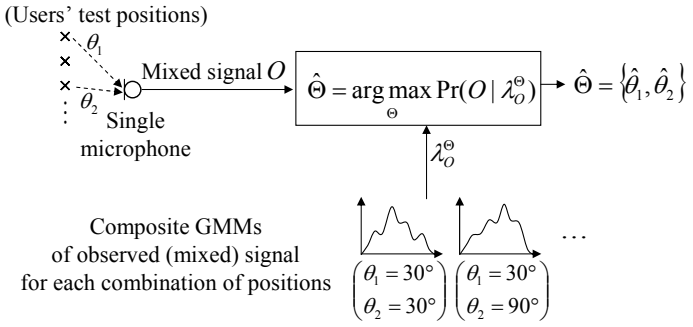

図 5.2　Two-talker localization using composite models of the mixed speech

本章では第 4 章で提案した枠組みを元に，新たに音響モデル合成を用いることで，単一マイクで 2 話者の音源位置推定を行う手法を提案する．提案手法ではそれぞれの話者に対して，話者位置毎の残響音声ケプストラムの GMM を学習しておき，それらを組み合わせて合成することで，あらゆる位置の組み合わせにおける混合音声ケプストラムの GMM を作成する．モデル合成を用いることで，実際に二人の話者があらゆる位置の組み合わせについて同時に発話をせずとも，その混合音声のモデルを作成することが可能となる．そして二人の話者が発話した評価音声について，位置の組み合わせ毎に合成された混合音声 GMM との尤度を比較することでそれぞれの話者の位置を推定する．

本章ではそれぞれの話者の位置毎の残響音声の GMM を学習する方法について，話者のクリーン音声 GMM と位置毎の音響伝達特性 SGM (Single Gaussian Model) を合成する方法と，位置毎に発話された音声信号から直接 GMM を学習する方法の二通りを検討し，1 話者及び 2 話者の位置推定実験を行ったところ，特に位置毎の学習データが少量の場合において前者の方法が優位であることが確認できた．

5.2　提案手法

5.2.1　提案手法の概要

提案する単一マイクによる 2 話者位置推定手法の概要を Fig. 5.1 と Fig. 5.2 に示す．提案手法はモデルの学習と位置の推定の二つの過程に分かれており，Fig. 5.1 はモデルの学習部を，Fig. 5.2 は位置の推定部をそれぞれ表している．モデルの学習部ではまず，二人の話者それぞれに対して，話者位置毎の残響音声ケプストラムの GMM を学習しておく．そしてこれらを組み合わせて合成することにより，あらゆる位置の組み合わせにおける 2 話者の混合音声信号ケプストラムの合成 GMM を得る．

位置の推定部では，二人の話者が発話した評価音声について，位置の組み合わせ毎に合

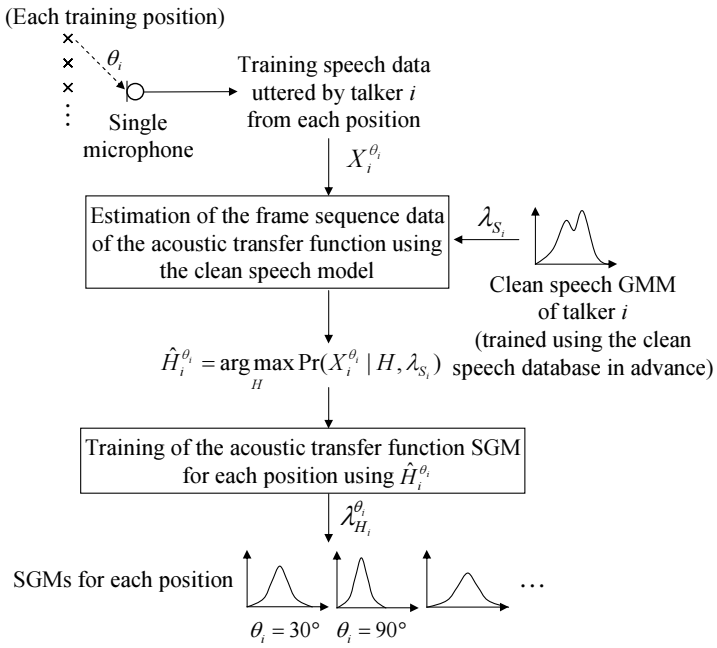

図 5.3 Training process for the acoustic transfer function SGM

成された混合音声信号モデルとの尤度を計算する．そして最も高い尤度を示した位置の組み合わせを，評価音声の話者の位置として出力する．

5.2.2 合成残響音声モデル (CRS モデル) と直接学習残響音声モデル (DTRS モデル)

前節で説明したモデル学習部における，話者位置毎の残響音声ケプストラムの GMM を得る方法として，本研究では二通りを検討する．その内，クリーン音声ケプストラムの GMM と位置毎の音響伝達特性ケプストラムの SGM を分けて学習しておき，それらを合成することで残響音声ケプストラムの GMM を得る方法を以降 “合成残響音声 (Composite Reverberant Speech; CRS) モデル” と呼び，位置毎に発話された音声信号から直接 GMM を学習する方法を “直接学習残響音声 (Directly-Trained Reverberant Speech; DTRS) モデル” と呼ぶことにする．これら二つの手法はどちらも話者毎の残響音声 GMM を合成することで 2 話者の混合音声 GMM を求めているが，CRS モデルではさらにそれぞれの話者の残響音声 GMM を，クリーン音声 GMM と音響伝達特性 SGM の合成により作成している点が DTRS モデルとの違いである．

CRS モデルで用いる音響伝達特性 SGM の学習の流れを Fig. 5.3 に示す．まず，学習用の音響伝達特性を得るために，ある話者が特定の位置から発話した音声を収録する．次に，収録した残響音声 $X_i^{\theta_i}$ $(i = 1, 2)$ からその音響伝達特性のケプストラム $\hat{H}_i^{\theta_i}$ を，同じ

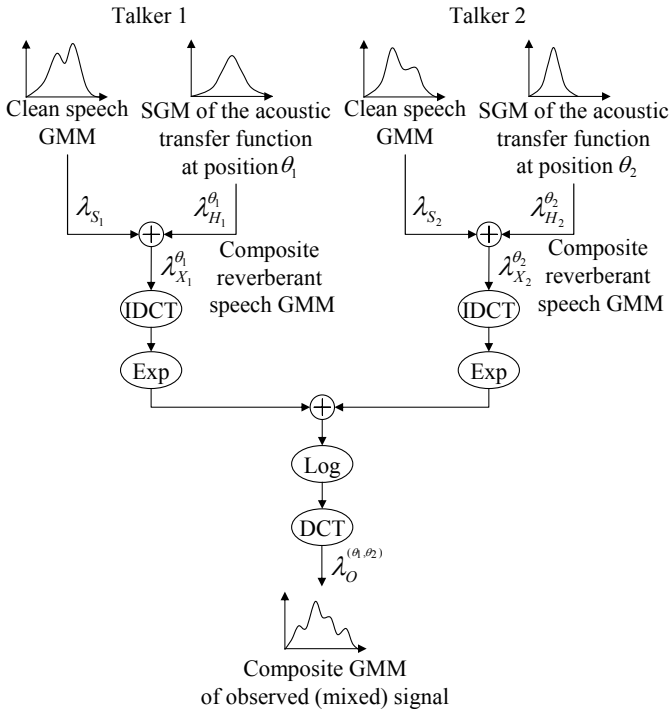

図 5.4　Composite model of the mixed speech of two talkers using CRS model

話者のクリーン音声ケプストラムの GMM（λ_{S_i} はモデルパラメータを表す）を用いて最尤推定法により推定する．そして推定された音響伝達特性のケプストラムを用いて，その位置における音響伝達特性 SGM のモデルパラメータ $\lambda_{H_i}^{\theta_i}$ を学習する．この処理を全ての音源位置に対して行う．また，もう一方の話者に対しても同様に位置毎の音響伝達特性 SGM を学習する．本来ならば音響伝達特性は話者に依存しないため，この処理は話者毎に行う必要はない．しかし本研究では，推定された音響伝達特性は，必ずしも話者の依存性を完全には除去しきれていないという点を考慮して，話者毎に音響伝達特性の SGM を学習することにした．

CRS モデルのモデル合成による 2 話者の混合音声 GMM 作成の流れを Fig. 5.4 に示す．全ての音響モデルはケプストラム領域で表現されている．CRS モデルを用いた手法では，まず先に得られた音響伝達特性 SGM と，音響伝達特性の推定に用いたクリーン音声 GMM をケプストラム領域にて合成することで，その話者とその位置における残響音声 GMM(CRS モデル) を作成する．次に，線形スペクトル領域での混合音声の加算性に注目し，得られた各話者の残響音声 GMM パラメータ $\lambda_{X_i}^{\theta_i}$ に逆コサイン変換（IDCT），指数変換（Exp）を適用し，線形スペクトル領域まで変換を行う．線形スペクトル領域にて，2 話者の特定の位置における残響音声モデルを合成し，対数変換（Log），コサイン変換（DCT）を行い，混合音声ケプストラムの GMM（パラメータ $\lambda_O^{(\theta_1,\theta_2)}$）を作成する．

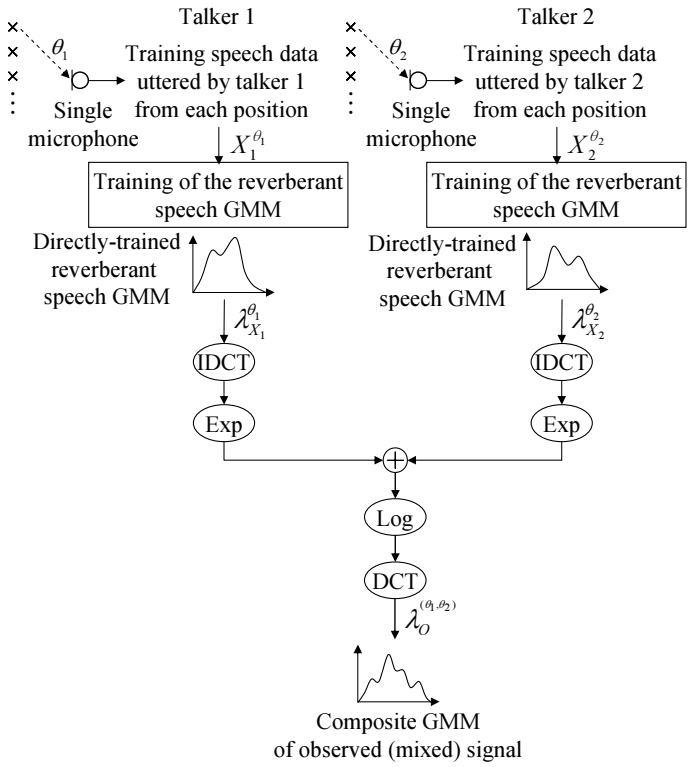

図 5.5 Composite model of the mixed speech of two talkers using DTRS model

モデル変換・合成については 5.2.4 節にて説明する．

一方 DTRS モデルのモデル合成による 2 話者の混合音声 GMM 作成の流れを Fig. 5.5 に示す．DTRS モデルを用いた手法では，それぞれの話者について，各位置から発話された音声から直接，残響音声ケプストラムの GMM(DTRS モデル) を学習する．その後は CRS モデルの手法と同様に，それぞれの話者の残響音声 GMM(DTRS モデル) を合成し，混合音声の GMM を作成する．

混合音声 GMM の合成は全ての位置の組み合わせについて行う．そして評価したい混合音声について，合成された混合音声モデルとの尤度を算出し，最も尤度の高い GMM の位置の対をそれぞれの話者の位置として出力する．次節では CRS モデルで用いる音響伝達特性の推定法について説明する．

5.2.3 音響伝達特性の推定

本節で述べる内容は基本的に第 3 章で述べた音響伝達特性の推定と基本的に同じ内容であるが，本章では各話者の残響音声を $x(t)$ としている点に注意されたい．第 3 章，第 4 章と同様，残響音声のケプストラムを以下のような線形加算で近似する．

$$X_{\mathrm{cep}}(d;n) \approx S_{\mathrm{cep}}(d;n) + H_{\mathrm{cep}}(d;n) \tag{5.1}$$

X_{cep}，S_{cep}，及び H_{cep} はそれぞれ残響音声，クリーン音声，音響伝達特性のケプストラムを表し，d はその次元を表す．第 4 章ではよりより正確な推定を行うためにクリーン音声を HMM でモデル化していたが，本章では第 3 章と同様に GMM でモデル化する．これは，後述のモデル合成にクリーン音声の GMM を利用しているためである．よって，音響伝達特性の推定は第 3 章の 3.3 節と同様である．

話者 i が位置 θ_i において発話したときの残響音声 $X_i^{\theta_i}(d;n)$ からその音響伝達特性 $\hat{H}_i^{\theta_i}(d;n)$ を推定し，これを用いて SGM を学習する．

$$\mu^{(H_i^{\theta_i})} = \frac{1}{N}\sum_n H_i^{\theta_i}(n) \tag{5.2}$$

$$\Sigma^{(H_i^{\theta_i})} = \frac{1}{N}\sum_n (H_i^{\theta_i}(n) - \mu^{(H_i^{\theta_i})})^T (H_i^{\theta_i}(n) - \mu^{(H_i^{\theta_i})}) \tag{5.3}$$

$\mu^{(H_i^{\theta_i})}$，$\Sigma^{(H_i^{\theta_i})}$ はそれぞれ音響伝達特性 SGM の平均ベクトルと共分散行列を表す．ただし本研究では共分散行列は対角行列を仮定する．

5.2.4　モデル合成による混合音声モデルの作成

CRS モデルでは前節で求めた音響伝達特性 $\hat{H}_i^{\theta_i}$ を用いて話者，位置毎に音響伝達特性の SGM を学習する．そして得られた音響伝達特性ケプストラム SGM のモデルパラメータ $\lambda_{H_i}^{\theta_i} = \left\{\mu^{(H_i^{\theta_i})}, \Sigma^{(H_i^{\theta_i})}\right\}$ と話者毎のクリーン音声ケプストラム GMM のモデルパラメータ λ_{S_i} を用いて，モデル合成により最終的に複数話者の混合音声 GMM のモデルパラメータ λ_O^{Θ} ($\Theta = \{\theta_1, \ldots, \theta_M\}$) を求める．$M$ は話者の数を表し，本研究においては $M = 2$ である．前節で説明したクリーン音声 GMM 及び音響伝達特性 SGM のパラメータ $\mu_k^{(S)}$，$\Sigma_k^{(S)}$，$\mu^{(H_i^{\theta_i})}$，$\Sigma^{(H_i^{\theta_i})}$ はケプストラム領域でのモデルであることを強調するため，本節では以降 $\mu_{\mathrm{cep},k}^{(S)}$，$\Sigma_{\mathrm{cep},k}^{(S)}$，$\mu_{\mathrm{cep}}^{(H_i^{\theta_i})}$，$\Sigma_{\mathrm{cep}}^{(H_i^{\theta_i})}$ と記述することにする．

混合音声のスペクトルは以下のように各話者の残響音声の線形和で表現される．

$$O_{\mathrm{spc}}^{\Theta}(\omega;n) = \sum_{i=1}^{M} X_{\mathrm{spc},i}^{\theta_i}(\omega;n) \tag{5.4}$$

$$X_{\mathrm{spc},i}^{\theta_i}(\omega;n) \approx S_{\mathrm{spc},i}(\omega;n) \cdot H_{\mathrm{spc},i}^{\theta_i}(\omega;n) \tag{5.5}$$

$S_{\mathrm{spc},i}(\omega;n)$ は話者 i のクリーン音声スペクトル，$X_{\mathrm{spc},i}^{\theta_i}(\omega;n)$，$H_{\mathrm{spc},i}^{\theta_i}(\omega;n)$ はそれぞれその話者の位置 θ_i における残響音声及び音響伝達特性のスペクトル，$O_{\mathrm{spc}}^{\Theta}(\omega;n)$ は各話者がそれぞれの位置から発話したときの混合音声のスペクトルを表す．

まず CRS モデルではそれぞれケプストラム領域でモデル化されているクリーン音声と音響伝達特性のモデルパラメータを用いて，残響音声のモデルパラメータを計算する．式

(5.1) より，ケプストラム領域では残響音声はクリーン音声と音響伝達特性の線形加算で表現されているため，残響音声の平均ベクトルと共分散行列はそれぞれクリーン音声と音響伝達特性の平均ベクトル及び共分散行列の線形加算で求められる [60].

$$\mu_{\mathrm{cep}}^{(X_i^{\theta_i})} = \mu_{\mathrm{cep}}^{(S_i)} + \mu_{\mathrm{cep}}^{(H_i^{\theta_i})} \tag{5.6}$$

$$\Sigma_{\mathrm{cep}}^{(X_i^{\theta_i})} = \Sigma_{\mathrm{cep}}^{(S_i)} + \Sigma_{\mathrm{cep}}^{(H_i^{\theta_i})} \tag{5.7}$$

DTRS モデルの場合では，残響音声のモデルパラメータは残響音声から直接 GMM を学習することで求める.

$$\mu_{\mathrm{cep},k}^{(X_i^{\theta_i})} = \sum_n \frac{\gamma_k(n) X_i^{\theta_i}(n)}{\sum_n \gamma_k(n)} \tag{5.8}$$

$$\Sigma_{\mathrm{cep},k}^{(X_i^{\theta_i})} = \sum_n \frac{\gamma_k(n)(X_i^{\theta_i}(n) - \mu_{\mathrm{cep},k}^{(X_i^{\theta_i})})^T (X_i^{\theta_i}(n) - \mu_{\mathrm{cep},k}^{(X_i^{\theta_i})})}{\sum_n \gamma_k(n)} \tag{5.9}$$

ただし，本研究では DTRS モデルにおいても共分散行列は対角行列を仮定している.

次に式 (5.6) と式 (5.7)，あるいは式 (5.8) と式 (5.9) により得られた話者毎の残響音声のモデルパラメータ $\lambda_{X_i}^{\theta_i}$ を用いて，混合音声信号のモデルパラメータ λ_O^{Θ} を求める．式 (5.4) より，混合音声は各話者の残響音声のスペクトル領域での線形和で表されているため，まず $\lambda_{X_i}^{\theta_i}$ をケプストラム領域から線形スペクトル領域に変換する．各モデルパラメータのケプストラム領域から対数スペクトル領域への変換は GMM の各正規分布に逆離散コサイン変換を行うことで計算することができる.

$$\mu_{\log}^{(X_i^{\theta_i})} = \Gamma^{-1} \mu_{\mathrm{cep}}^{(X_i^{\theta_i})} \tag{5.10}$$

$$\Sigma_{\log}^{(X_i^{\theta_i})} = \Gamma^{-1} \Sigma_{\mathrm{cep}}^{(X_i^{\theta_i})} (\Gamma^{-1})^T \tag{5.11}$$

Γ は離散コサイン変換の変換行列を表し，$\mu_{\log}^{(X_i^{\theta_i})}$ 及び $\Sigma_{\log}^{(X_i^{\theta_i})}$ はそれぞれ残響音声の対数スペクトル領域における平均ベクトル及び共分散行列を表す．ここで，ケプストラム領域における各モデルの共分散行列は対角行列で定義されているが，逆離散コサイン変換後の対数スペクトル領域では対角行列ではなく，非対角成分にも分散値を持った行列として，以降の対数スペクトル領域，及び線形スペクトル領域での計算を行っている.

次に対数スペクトル領域のモデルパラメータを線形スペクトル領域に変換する．本研究ではケプストラム，及びその線形変換である対数スペクトル領域ではモデルは正規分布を仮定している．対数スペクトルが正規分布に従うと仮定した場合，その指数変換である線

形スペクトルは対数正規分布に従うと仮定される．対数正規分布とは，その分布に従う変数の対数値が正規分布に従う分布である．線形スペクトルを対数正規分布で仮定した場合，その平均，及び共分散行列は対数スペクトル領域の正規分布の平均，共分散行列を用いて以下のように得られる．

$$\mu_{\mathrm{spc},p}^{(X_i^{\theta_i})} = \exp\left\{\mu_{\log,p}^{(X_i^{\theta_i})} + \sigma_{\log,pp}^{(X_i^{\theta_i})^2}/2\right\} \tag{5.12}$$

$$\sigma_{\mathrm{spc},pq}^{(X_i^{\theta_i})^2} = \mu_{\mathrm{spc},p}^{(X_i^{\theta_i})} \cdot \mu_{\mathrm{spc},q}^{(X_i^{\theta_i})} \cdot \left\{\exp\left(\sigma_{\log,pq}^{(X_i^{\theta_i})^2}\right) - 1\right\} \tag{5.13}$$

$\mu_{\mathrm{spc},p}^{(X_i^{\theta_i})}$ 及び $\sigma_{\mathrm{spc},pq}^{(X_i^{\theta_i})^2}$ はそれぞれ線形スペクトル領域における平均ベクトルの p 次元目の要素と共分散行列の (p,q) 番目の要素を表す．次に話者毎の残響音声のモデルパラメータから，線形スペクトル領域における混合音声信号のモデルパラメータを合成する．このとき，線形スペクトル領域において，混合音声信号の平均ベクトル及び共分散行列はそれぞれ各話者の残響音声の平均ベクトル及び共分散行列の線形加算として近似する [61].

$$\mu_{\mathrm{spc}}^{(O^{\Theta})} \approx \sum_{i=1}^{M} \mu_{\mathrm{spc}}^{(X_i^{\theta_i})}, \qquad \Sigma_{\mathrm{spc}}^{(O^{\Theta})} \approx \sum_{i=1}^{M} \Sigma_{\mathrm{spc}}^{(X_i^{\theta_i})} \tag{5.14}$$

その後，線形スペクトル領域で得られた混合音声信号モデル（対数正規分布で仮定）のパラメータを用いて，対数スペクトル領域でのモデルパラメータ（正規分布で仮定）を求める．これは式 (5.12)，(5.13) と逆の処理を行うことで求められる．

$$\sigma_{\log,pq}^{(O^{\Theta})^2} = \log\left\{\frac{\sigma_{\mathrm{spc},pq}^{(O^{\Theta})^2}}{\mu_{\mathrm{spc},p}^{(O^{\Theta})} \cdot \mu_{\mathrm{spc},q}^{(O^{\Theta})}} + 1\right\} \tag{5.15}$$

$$\mu_{\log,p}^{(O^{\Theta})} = \log \mu_{\mathrm{spc},p}^{(O^{\Theta})} - \sigma_{\log,pp}^{(O^{\Theta})^2}/2 \tag{5.16}$$

最後に離散コサイン変換を行い，ケプストラム領域に変換する．

$$\mu_{\mathrm{cep}}^{(O^{\Theta})} = \Gamma \mu_{\log}^{(O^{\Theta})}, \qquad \Sigma_{\mathrm{cep}}^{(O^{\Theta})} = \Gamma \Sigma_{\log}^{(O^{\Theta})} \Gamma^T \tag{5.17}$$

なお本研究ではケプストラム領域に戻した後，共分散行列は非対角成分を 0 として対角行列に定義し直している．

5.2.5　尤度最大基準による話者の位置推定

前節までの方法により話者毎の位置の全組み合わせについて混合音声信号の GMM を計算しておく．そして評価したい混合音声について，合成された混合音声モデルとの尤度を算出し，最も尤度の高い GMM の位置の対をそれぞれの話者の位置として出力する．

$$\hat{\Theta} = \underset{\Theta}{\mathrm{argmax}} \Pr(O|\lambda_O^{\Theta}) \tag{5.18}$$

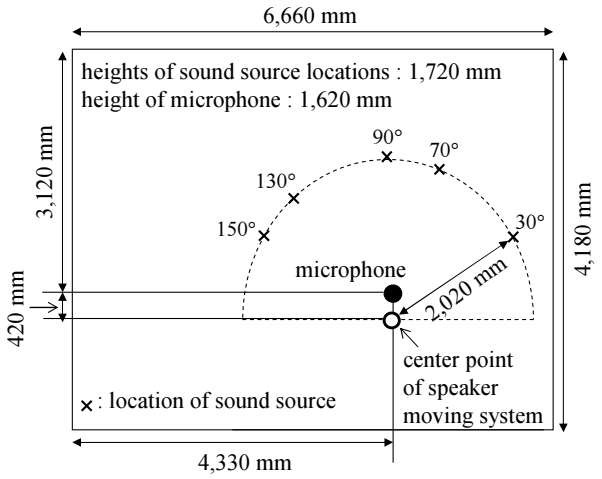

図 5.6 Experiment room environment for simulation

λ_O^{Θ} は位置の組み合わせ Θ におけるケプストラム領域での混合音声信号の合成モデルである．

5.3 評価実験

5.3.1 実験環境

提案手法を評価するために特定話者による音源位置推定実験を行った．音声データは ASJ 研究用連続音声データベース [46] より男性話者，女性話者それぞれ 1 名の発話データを用い，第 4 章と同じく RWCP 実環境音声・音響データベース [57] において収録されたインパルス応答を畳み込むことで，それぞれの位置における残響音声を作成した．音源位置は音源方向 30°，90°，150° の 3 種類存在する場合と，30°，70°，90°，130°，150° の 5 種類存在する場合で実験を行った．このとき，2 話者の位置の組み合わせはそれぞれ 9 種類と 25 種類存在する．Fig. 5.6 に本章の実験における音源位置を示す．

音声信号はサンプリング周波数 12 kHz，窓幅 32 msec，フレームシフト 8 msec の分析条件で MFCC 16 次元を特徴量として使用した．CRS モデルにおいて，音響伝達特性の推定及び残響音声モデルの合成に用いるクリーン音声の特定話者 GMM は，話者毎に 40 文を用いて学習し，混合数は 64 とした．CRS モデルにおける音響伝達特性 SGM の学習，及び DTRS モデルにおける残響音声 GMM の学習には 1 文，5 文，10 文の 3 通りで実験を行った．評価は長さ 1 sec の音声 100 セグメントを用いた．評価する混合音声の話者のパワーの比率は 1 セグメントあたり平均約 -5.90 dB，標準偏差約 ± 2.54 dB であった．1 セグメント毎に最も尤度の高い混合音声 GMM の位置の組み合わせを出力し，その正解率を評価した．なお，クリーン音声の学習，位置の学習，評価に用いたデータは

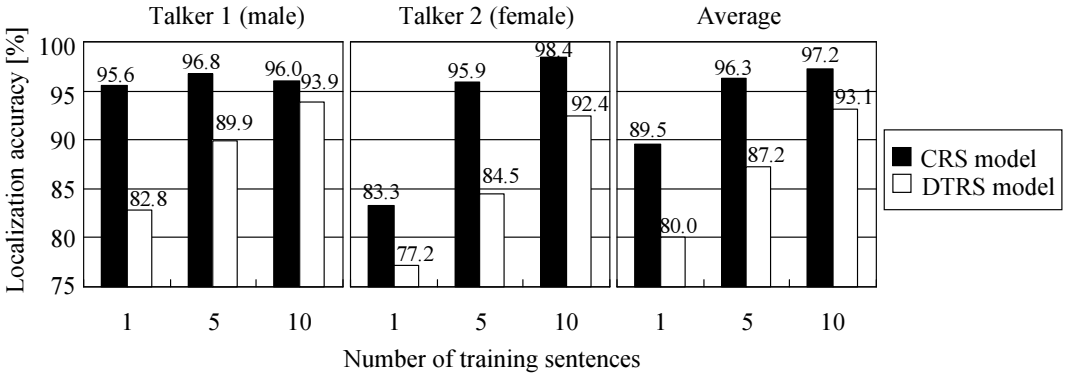

図 5.7　Single-talker localization accuracies [%], where the number of positions is three.

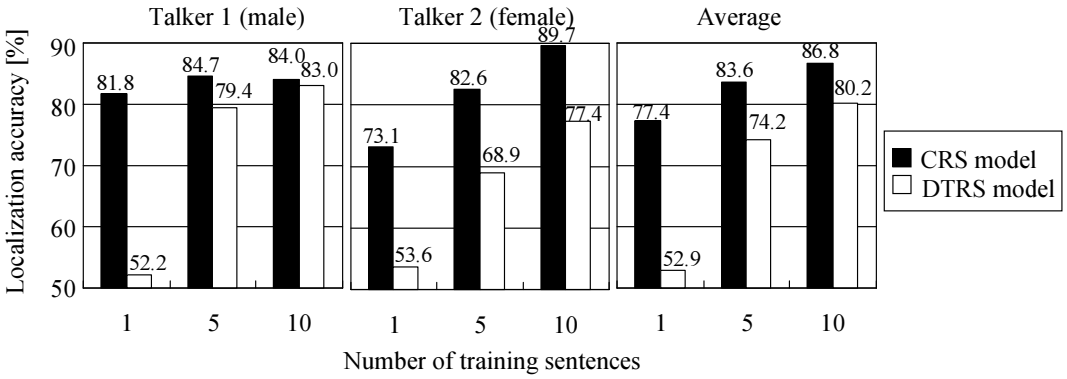

図 5.8　Single-talker localization accuracies [%], where the number of positions is five.

それぞれ異なる発話内容の音声を使用している．

5.3.2　実験結果

CRS モデルと DTRS モデルの違いは一人の話者の残響音声信号のモデルをクリーン音声モデルと音響伝達特性モデルの合成により求めるか，残響音声から直接学習により求めるかの違いである．この違いを評価するため，まずそれぞれの話者について 1 話者における音源位置推定実験を独立に行った．一人の話者が発話した評価データ (残響音声) と，それぞれの手法で求めた位置毎の残響音声モデル $\lambda_{X_i}^{\theta_i}$ との尤度を算出し，最も高い尤度を示した位置を出力し，その正解率を評価する．なお，どの話者が話しているかは既知とする．

位置数 3 及び位置数 5 における音源位置推定の正解率をそれぞれ Fig. 5.7，Fig. 5.8 に

示す．図より，全体的に CRS モデルの方が高い正解率を示していることが分かる．また，位置毎の残響音声モデルの学習に用いる文章数が減るにつれて正解率の差が広がっている．これは DTRS モデルの場合，少量の学習データでは残響音声の音韻のバリエーションを学習しきれなかったためと考えられる．一方 CRS モデルではあらかじめ 40 文を用いてクリーン音声 GMM を学習し，それを合成して残響音声モデルを作成しているため，少量の学習データであっても精度の低下をある程度抑えることができたと考えられる．

音源の位置が 5 種類の場合，識別クラス数が増え，位置同士の距離も縮まるため位置推定正解率は全体的に減少する．しかし残響音声モデルの学習データが 1 文の場合，DTRS モデルでは 10 文の場合と比較して平均 27.3% 下がったのに対して CRS モデルでは 9.4% の減少に留まった．このことから，少量の学習データ数に対する頑健性は位置の数が増えても保たれていることが分かる．本手法は残響音声モデルを作成するためにユーザの発話を用いているため，位置の数が増えるほど，位置毎の学習データは少量であることが望ましい．CRS モデルはクリーン音声 GMM を学習するために数十文の発話が必要であるが，これは位置の数に依存しないため，位置の数が多いほど有効な手法であると言える．一方位置の数が少ない場合は，DTRS モデルのように直接残響音声を学習したほうが総合的な学習データ数は CRS モデルよりも少量で済む場合もあると考えられる．

次に 2 話者の位置推定実験を行い，手法の有効性を評価した．この実験では CRS モデル, DTRS モデルの他に，モデル合成を一切行わずに混合音声 GMM を学習する手法を用いて比較を行った．以降この手法を “直接学習混合音声 (Directly-Trained Mixed Speech; DTMS) モデル” と呼ぶことにする．DTMS モデルでは二人の話者がそれぞれの位置から同時に発話した音声を収録し，その混合音声から直接 GMM を学習する．この手法はモデル合成を行わずに混合音声 GMM を学習することができるが，あらゆる位置の組み合わせ毎に複数のユーザが同時に発話をする必要があるため，位置の数が増えるにつれて学習のためのユーザの発話回数が乗算的に増えていくという欠点がある．

位置数 3 及び位置数 5 における 2 話者の位置推定正解率をそれぞれ Fig. 5.9，Fig. 5.10 に示す．Both talker positions は 2 話者の位置が両方とも正しく推定できた場合，At least one talker position は少なくとも 1 話者 (2 話者も含む) の位置が正しく推定できた場合を示す．図中の位置毎の学習発話数を W, 位置の数を Y とした場合，1 ユーザにつき実際に発話する回数は CRS モデル及び DTRS モデルではそれぞれ $W \times Y(+40)$, $W \times Y$ であり，DTMS モデルの場合は $W \times (Y^2)$ である．また，両方の位置が正解する割合の期待値は 3 位置の場合で 11.1%, 5 位置の場合で 4.0% ，少なくとも一方の位置が正解する割合の期待値はそれぞれ 55.6% ，36.0% である．Fig. 5.9，Fig. 5.10 より，ほぼ全ての条件下において CRS モデルが他の手法と比較して高い正解率を示していることが分かる．2 話者の混合音声は音韻同士の組み合わせが考慮される分，そのバリエーションが 1 話者の音声よりも膨大になる．そのため，1 話者毎の音響モデルを十分なデータを

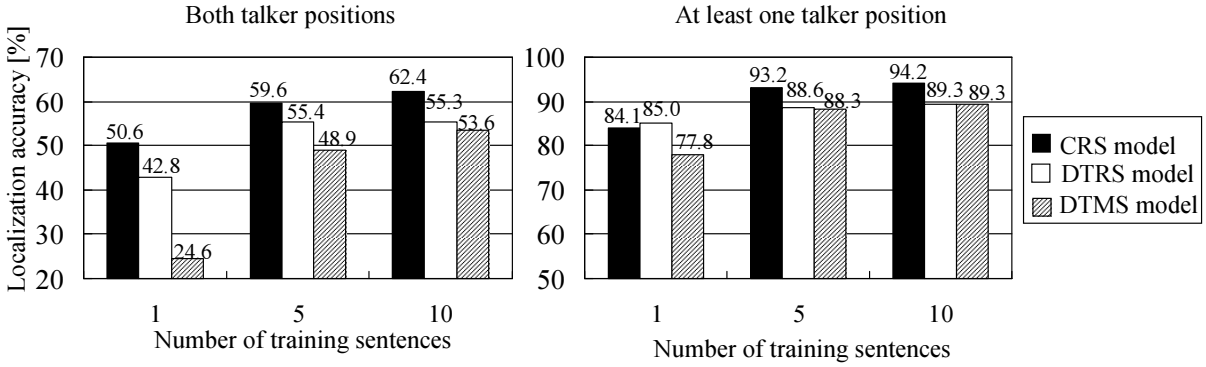

図 5.9　Two-talker localization accuracies [%], where the number of positions is three. Test data consists of 100 speech segments having a time length 1 sec.

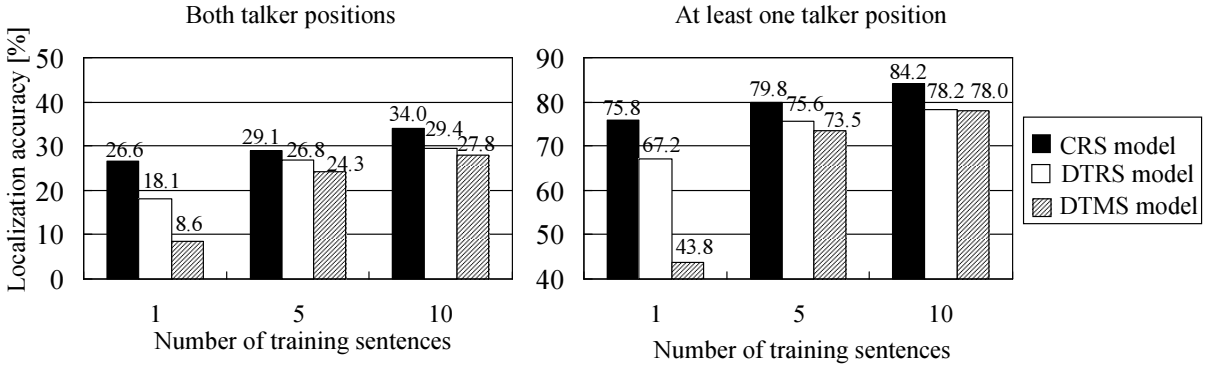

図 5.10　Two-talker localization accuracies [%], where the number of positions is five. Test data consists of 100 speech segments having a time length 1 sec.

用いて予め学習させている CRS モデルが特に高い正解率を示したと考えられる．DTRS モデルと DTMS モデルの比較では，学習データが少量の場合において正解率に差が見られた．しかし，両方の位置を推定できた割合は最大でも 3 位置の場合で 62.4% ，5 位置の場合で 34.0% であり，実用の面からは決して高い正解率とは言い難い．

次に評価音声を長さ 5 sec の音声 50 セグメントに変えて実験を行った．このときの音源位置推定正解率を Fig. 5.11，Fig. 5.12 に示す．図より，評価音声の長さが 1 sec の結果と比較して正解率が上昇していることが分かる．特に学習データ数が多い場合における DTRS モデルと DTMS モデルが CRS モデルの場合と比べて正解率が大きく上昇している．これは，評価音声が長くなることで情報量が増加したため，10 文で学習させた DTRS モデルや DTMS モデルのようなある程度混合音声を表現できたモデルであれば比

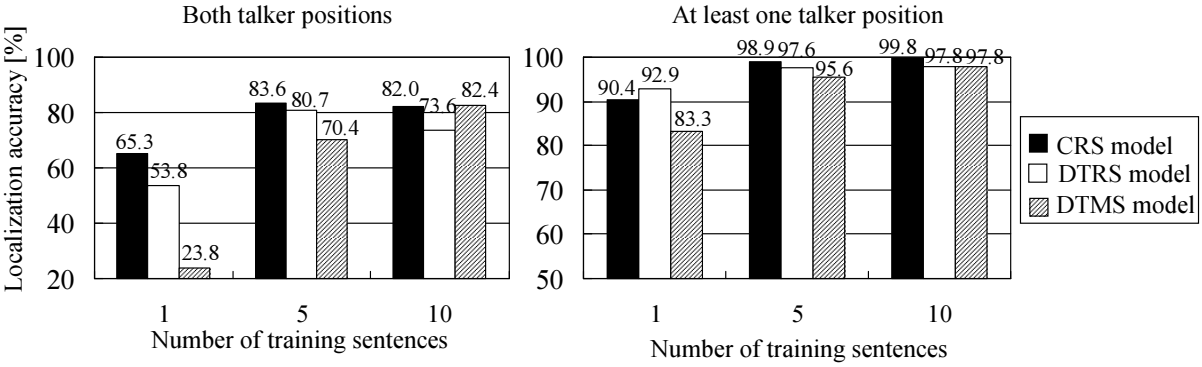

図 5.11 Two-talker localization accuracies [%], where the number of positions is three. Test data consists of 50 speech segments having a time length 5 sec.

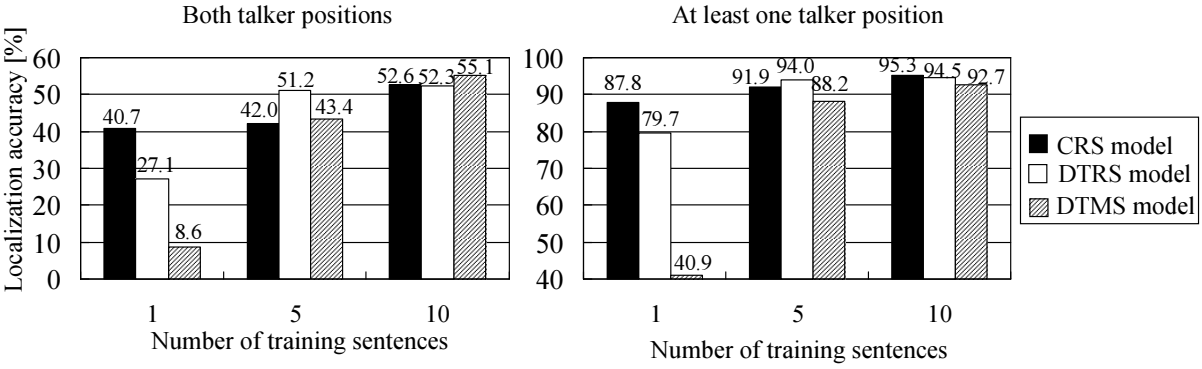

図 5.12 Two-talker localization accuracies [%], where the number of positions is five. Test data consists of 50 speech segments having a time length 5 sec.

較的精度よく位置推定ができたためと考えられる.

5.4 まとめ

本章では第 4 章で提案したシングルチャネル音源位置推定の枠組みを元に，各話者・各位置における残響音声の GMM から，その混合音声信号の GMM をモデル合成により求めて用いることで，2 話者の音源位置推定を単一マイクで行う手法について提案した．音響モデル合成を用いた手法として CRS モデルと DTRS モデルの二通りを検討し，1 話者及び 2 話者の音源位置推定実験により提案手法の有効性を評価した．実験では CRS モデルがモデル合成を用いない手法と比べて，特に音源位置毎の学習データが少量の場合にお

いて優位性を示しており，学習にユーザの発話データが必要な本手法において有効な手法であると言える．

また，評価音声を長くすることで学習データが少量でも位置推定正解率が上昇することが確認できた．今後は音源が移動する場合のような，短い評価データの場合でもより高い正解率を得るために，最適な音響伝達特性のモデル化やモデル合成手法を検討する．また，話者の音声のパワー比が大きく異なる場合の対応や，話者数や話者の推定法についても検討を行う．

第 6 章

音響伝達特性の識別に基づく話者の頭部方向の推定

6.1 まえがき

本章では，第 4 章で提案した枠組みを用いて，話者の頭部の方向を推定する手法を提案する．人と人のコミュニケーション，あるいはロボットとのコミュニケーションにおいて，話者の頭部方向は聞き手にとって重要な手がかりの一つであり，我々は話し手の頭部の向きから「誰が話しているのか」だけでなく，「誰に向かって話しているのか」という情報まで得ることができる．この「誰が誰に向かって話しているのか」という情報は，特に複数のユーザーが会話をしている状況において有効であり，会議システム，ロボット対話，雑談とシステム要求の判別など，様々なタスクにおいて利用することができると期待される．

これまでに多くの音源位置推定手法が提案されてきた一方，話者の頭部方向の推定へ関心が向けられ出したのは比較的近年のことであり，いくつかの手法が提案されている [62, 63, 64, 65]．これらの手法は複数組のマイクロホンアレーからなるネットワークを用いており，従来の音源位置推定のアルゴリズムを拡張することで話者の頭部方向を推定している．文献 [62] で提案されている手法は，従来の音源位置推定法の一つである SRP-PHAT (Steered Response Power with the PHAse Transform) をベースとした手法であり，従来の SRP-PHAT の目的関数を，話者の頭部方向に依存する重み係数によって重み付けを行うことにより，話者の位置推定問題から頭部方向推定問題へ拡張している．文献 [63, 64] では話者の頭部の方向ごとに変化する観測信号の音圧のパターンに着目しており，文献 [64] では提案手法を文献 [62] の手法と組み合わせることでさらなる精度

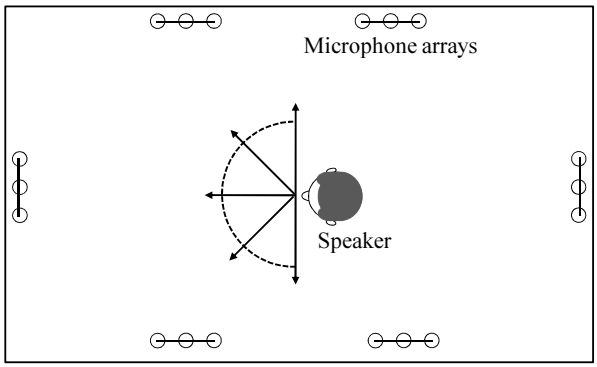

図 6.1　A head orientation estimation system based on a network of microphone arrays

の向上を示している．また，文献 [65] では各マイクロホンアレーから算出された音源方向推定結果を用いて作成されるヒストグラムから，話者の頭部方向を推定する手法を提案している．しかしながらこれらの手法は複数のマイクロホンアレーを Fig. 6.1 のように，ユーザーを囲むようにして部屋の壁などに設置する必要があり，システムが大規模になってしまうという欠点がある．

第 4 章では観測信号の音響伝達特性が音源の位置によって異なる点に着目し，音響伝達特性の識別による音源位置推定の手法を提案した．本章では観測信号の音響伝達特性が，話者の位置だけではなく，頭部の方向にも依存することに着目し，音響伝達特性の識別により話者の頭部方向を推定する手法を提案する．第 4 章で提案した音源位置推定手法では，話者の位置ごとの音響伝達特性を学習・識別していたのに対し，本章における提案手法では，各音源位置とその位置における頭部の各回転方向の音響伝達特性を学習・識別する．従来の頭部回転方向の推定法と異なり，本手法はあらかじめ音響伝達特性を学習しておく必要があるが，マイクの位置を任意の場所に設置することができるという利点がある．評価実験では音源位置のみの推定，頭部方向のみの推定，音源位置及び頭部方向の推定の 3 つのタスクにおいて実験を行い，その有効性を示す．

6.2　音源位置と頭部方向の推定

6.2.1　提案手法の概要

本研究では第 4 章と同様に，音響伝達特性を用いて音源の位置と頭部方向を推定する．音響伝達特性は音源の位置や頭部方向によって異なる値を持つため，あらかじめこれを各音源位置とその位置における頭部方向毎に学習しておけば，評価音声に対してもその音響伝達特性を識別することで音源位置及び頭部方向を推定することができる．

提案手法の概要を Fig. 6.2 に示す．まず，位置と頭部方向の組み合わせ毎の音響伝達特

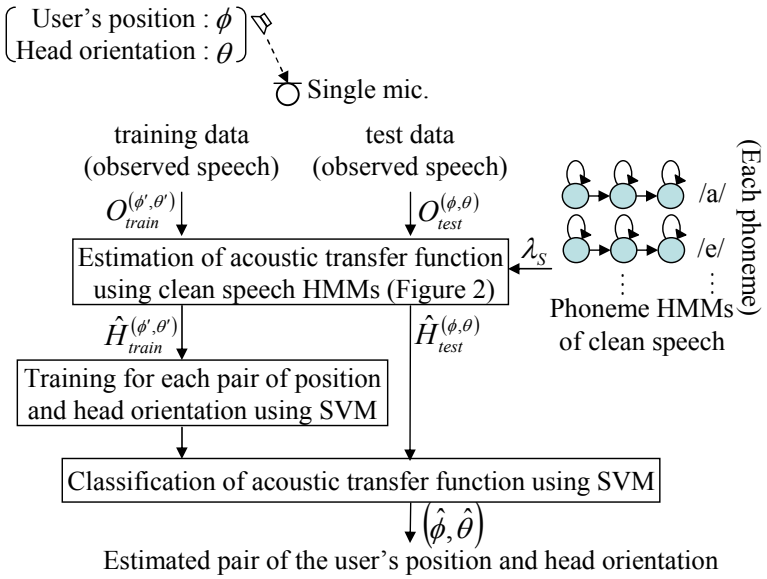

図 6.2 System overview

性を学習するために，それぞれの位置 θ において頭部を各方向 ϕ へ向けた状態で発話された音声 $O^{(\phi,\theta)}_{train}$ を収録し，その音響伝達特性をクリーン音声の音素 HMM を用いて推定する．次に，位置と頭部方向毎に推定された音響伝達特性 $\hat{H}^{(\phi,\theta)}_{train}$ を SVM により学習する．そして評価したい音声 $O^{(\phi,\theta)}_{test}$ についても学習データと同様に，音素認識結果により得られるラベル情報を用いて音響伝達特性 $\hat{H}^{(\phi,\theta)}_{test}$ を推定し，それを SVM で識別することで，音源位置と頭部方向 $(\hat{\phi},\hat{\theta})$ を推定する．

6.2.2 音響伝達特性の推定

本研究においても，これまで提案してきた手法と同様，観測信号から音響伝達特性を推定する必要がある．本研究では，第 4 章で述べたクリーン音声 HMM を用いて音響伝達特性を推定する．音源位置と頭部方向毎に音響伝達特性を推定し，それらを SVM によって学習・識別することで，音源位置と頭部方向の推定を行う．

6.3 評価実験

6.3.1 実験環境

提案手法を評価するために，音源位置のみの推定，頭部方向のみの推定，音源位置及び頭部方向の推定の 3 つのタスクにおいて実験を行った．実験環境の図を Fig. 6.3 に示す．音源位置は 3 箇所で，音源距離は 1.5m，音源方向はそれぞれ 50°，90°，130° であ

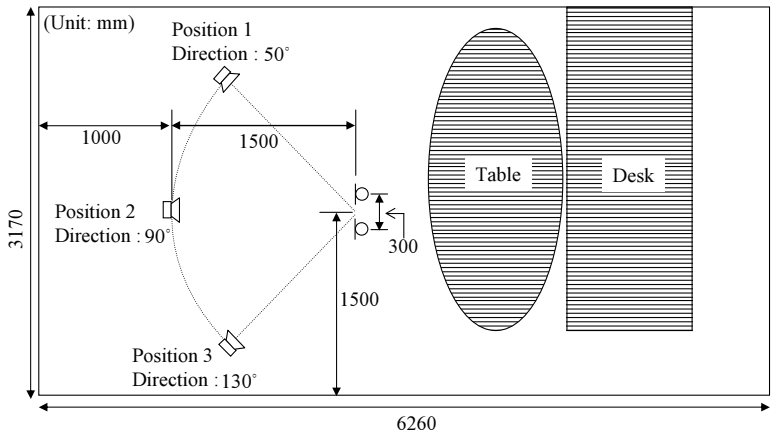

図 6.3　Experimental room environment and the loudspeaker position. The direction of each position means the direction from the microphones.

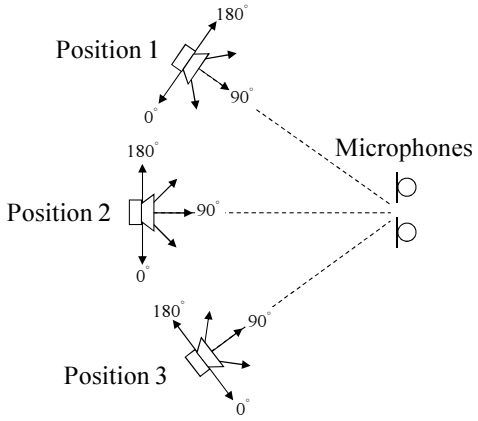

図 6.4　The head orientation of the loudspeaker for each position

る．それぞれの位置におけるスピーカーの向きを Fig. 6.4 に示す．スピーカーの向きは 0°，45°，90°，135°，180° の 5 方向であり，スピーカーがマイクロホンの方向を向いているときを 90° とした．

インパルス応答の収録環境を Fig. 6.5 に示す．約 6.3 m × 3.2 m × 2.8 m (W × D × H) の部屋において，位置・スピーカーの向き毎にインパルス応答を収録し，これをクリーン音声に畳みこむことで，残響音声を作成した．インパルス応答は 64 回の同期加算による強調処理を行った後，TSP 法 [66] により計算した．部屋の残響時間は約 380 msec であった．スピーカーは BOSE Mediamate II を，マイクロホンには指向性マイク (SONY ECM-66B) を使用した．

音声データは ATR 研究用日本語音声データベースセット A[56] より男性話者 1 名の単語音声を用い，サンプリング周波数 12 kHz，窓幅 32 msec，フレームシフト 8 msec の分

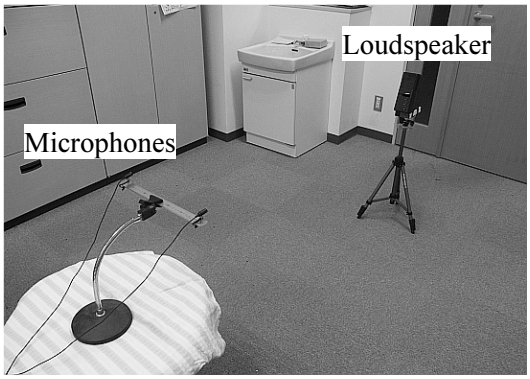

図 6.5 Photo of the recording environment

析条件で MFCC 16 次元を特徴量として使用した．音響伝達特性の推定におけるクリーン音声の音素 HMM は，2,620 単語を用いて学習した．音素数は 54，各音素 HMM の状態数は 3，混合数は 32 である．音響伝達特性の学習には 50 単語を，評価には 166 単語を，組み合わせを変えて 4-fold のクロスバリデーションにより推定精度を算出した．なお，クリーン音声の学習，位置の学習，評価に用いたデータはそれぞれ異なる発話内容の単語を使用している．SVM には SVM^{light}[54] を，カーネル関数に RBF (Gaussian) カーネルを使用し，one-vs-rest 法によりマルチクラス識別を行った．

6.3.2 実験結果

6.3.2.1 音源位置推定における実験結果

初めに，音源位置推定のみの性能を，想定されるスピーカーの向きが 1 方向 (90°)，3 方向 (0°，90°，180°)，5 方向 (0°，45°，90°，135°，180°) の場合に分けて評価した．この実験では，位置とスピーカーの向きの組み合わせ毎に学習し，評価データについて，その位置と向きを識別して出力する．スピーカーの向きが 1 方向，3 方向，5 方向の場合，組み合わせはそれぞれ 3，9，15 通り存在する．その時，スピーカーの向きの推定結果は考慮せずに，少なくとも音源位置のみが正しく認識されていれば正解として，認識率を計算した．Table. 6.1 にそれぞれの場合における位置の認識率を示す．

Table. 6.1 より，音源の位置によって認識精度のばらつきが存在することが分かる．スピーカーの向きが 90° のときの，各位置における音響伝達特性の分布を Fig. 6.6 に示す．音響伝達特性は (4.1) 式へ実際のクリーン音声 $S_{cep}(d;n)$ を代入して求めている．また，主成分分析を用いて 2 次元に圧縮してプロットしている．Fig. 6.6 より，位置 1 の音響伝達特性は他の位置の音響伝達特性から離れて分布しており，識別がし易い分布になっていることが分かる．

表 6.1　Localization accuracy [%] of the proposed method for each position (pos.), where the number of the possible orientations (ori.) is one (90 degrees, top table), three (0, 90 and 180 degrees, middle table), and five (0, 45, 90, 135 and 180 degrees, bottom table)

ori. \ pos.	pos. 1	pos. 2	pos. 3
90 deg.	95.1	78.6	83.3

ori. \ pos.	pos. 1	pos. 2	pos. 3
0 deg.	92.2	75.0	93.2
90 deg.	95.6	79.5	76.1
180 deg.	83.4	88.4	79.1
average	90.4	81.0	82.8

ori. \ pos.	pos. 1	pos. 2	pos. 3
0 deg.	90.2	73.0	92.2
45 deg.	90.5	77.3	84.9
90 deg.	92.8	76.5	74.1
135 deg.	83.4	74.2	58.6
180 deg.	80.9	88.6	77.1
average	87.6	77.9	77.4

比較手法として，2 チャンネルマイクを用いた CSP 法による音源方向推定の性能を評価した．CSP 法は提案手法とは異なり，音源の方向が連続値として出力される．そのため，提案手法と評価の条件を一致させるために，この値が実際の音源方向に対して誤差が $\pm 20^\circ$ 以内であれば正解として，認識率を算出した．結果を Table. 6.2 に示す．表より，全体的には CSP 法の方が提案手法よりも音源の方向を正しく認識出来ている．しかし，CSP 法の場合ではスピーカーの向きが 0° と 180° の場合に位置の推定性能が低下していることが分かる．これは，スピーカーが横を向いたことによって，壁からの反射波の影響が強くなったために，CSP 法の性能が低下したためと考えられる．

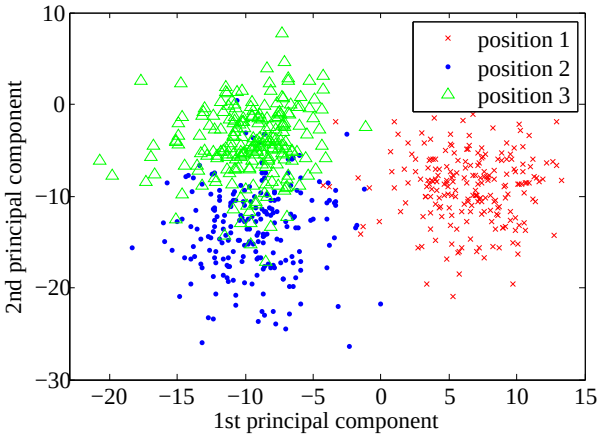

図 6.6　Mean values of the acoustic transfer function for each position fixing the head orientation at 90 degrees

表 6.2　Estimation accuracy [%] of the sound-source-direction using CSP analysis for each position and the orientation within an error of 20 degrees.

ori. \ pos.	pos. 1	pos. 2	pos. 3
0 deg.	65.7	100.0	87.5
45 deg.	100.0	100.0	99.1
90 deg.	100.0	100.0	100.0
135 deg.	100.0	100.0	99.1
180 deg.	89.4	100.0	75.9
average	91.1	100.0	92.3

6.3.2.2　頭部方向の推定及び音源位置・頭部方向の同時推定における実験結果

次に，それぞれの場所にスピーカーの位置を固定し，スピーカーの向きのみを変えることで，頭部方向の推定における精度の評価を行った．想定されるスピーカーの方向が 3 種類における推定精度と，5 種類における推定精度を Table. 6.3 に示す．表より，スピーカーの向きが 3 種類の場合において，平均 75.6% の認識率でスピーカーの向きを推定することができているが，5 種類の場合では性能が低下しており，特に 45° と 135° の認識率が低いことが分かる．Fig. 6.7 は音源位置 1 において，スピーカーの向きが 90°，135°，180° のときの音響伝達特性の分布を表す．図より，スピーカーの回転方向による音響伝達特性の変化は，音源位置による音響伝達特性の変化ほど顕著ではないことと，特に 135°

表 6.3　Orientation estimation accuracies [%] for each fixed position (pos.), where the number of possible orientations (ori.) is three (0, 90 and 180 degrees, top table) and five (0, 45, 90, 135 and 180 degrees, bottom table)

ori. \ pos.	pos. 1	pos. 2	pos. 3
0 deg.	73.5	78.5	93.7
90 deg.	81.2	68.1	68.4
180 deg.	75.9	74.8	66.4
average	76.9	73.8	76.2

ori. \ pos.	pos. 1	pos. 2	pos. 3
0 deg.	65.7	70.8	86.9
45 deg.	40.8	8.9	55.6
90 deg.	59.0	54.7	48.9
135 deg.	23.6	28.8	32.1
180 deg.	72.6	72.9	50.5
average	52.3	47.2	54.8

の音響伝達特性が判別しづらい分布になっていることが分かる．

最後に，音源の位置とスピーカーの向きの両方を変化させて，音源位置と頭部方向の同時推定の精度を評価した．スピーカーの向きの候補が 3 種類の場合における推定精度と，5 種類の場合における推定精度を Table 6.4 に示す．表より，これまでの実験結果と同様に，音源の位置によって推定精度の差が存在することと，スピーカーの向きが 45° と 135° の場合において推定精度が低いことが分かる．これらの向きも正確に推定するためには，第 4 章で述べた通り，音響伝達特性のより正確な推定が必要であると考えられる．また，スピーカーの向きの変動をより顕著に表現するような特徴量の検討も必要である．

6.4　まとめ

本章では，音響伝達特性が話者の位置だけでなく，頭部方向によっても異なる特性を持つ点に着目し，第 4 章で述べた枠組みを元に，各音源位置・頭部方向において発話された音声信号から，その音響伝達特性をクリーン音声の音素 HMM を用いて推定し，推定された音響伝達特性を SVM により学習・識別することで，音源の位置と頭部方向をシングルチャネルで推定する手法について提案した．実験では，音源位置の候補が 3 種類，頭部方

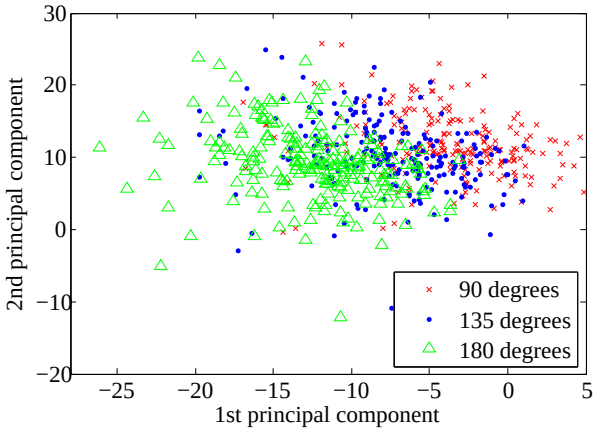

図 6.7 Mean values of the acoustic transfer function of three head orientations (90, 135 and 180 degrees) fixing the location at position 1.

向の候補が 0°，90°，180° の場合において，頭部方向のみの推定が平均 75.6% の認識率，音源位置と頭部方向の同時推定がが平均 66.0% の認識率でそれぞれ行えた．しかしながら，頭部方向の候補に 45° と 135° を追加した場合，認識率が大幅に低下していることから，現状の手法では細かい頭部の方向の識別は困難であることが分かる．

本手法では音源位置・頭部方向毎に音響伝達特性を推定する必要があるが，より多くの音源位置・頭部方向の推定を行う場合，膨大な量の学習データを要するという問題がある．そのため，モデルの適応やオンライン学習といった，より少量の学習データで実装が行える手法について，今後検討を行う．

表 6.4　Localization and head orientation estimation accuracy [%], where the number of head orientations is three (0, 90 and 180 degrees, top table) and five (0, 45, 90, 135 and 180 degrees, bottom table)

ori. \ pos.	pos. 1	pos. 2	pos. 3
0 deg.	70.6	58.1	90.2
90 deg.	80.7	54.7	49.4
180 deg.	61.9	71.7	56.8
average	71.1	61.5	65.5

ori. \ pos.	pos. 1	pos. 2	pos. 3
0 deg.	63.9	52.9	82.1
45 deg.	40.4	5.9	54.2
90 deg.	57.7	40.4	34.9
135 deg.	15.2	17.3	6.2
180 deg.	57.1	72.6	44.4
average	46.8	37.8	44.4

第 7 章

結論

本研究では，音声の持つ音響伝達特性が音源位置に依存する点に着目し，音響伝達特性を用いることで音源の位置や方向を単一マイクロホンのみで推定する手法について提案した．第 3 章では，音響伝達特性を用いた音源方向推定のアプローチの一つとして，マイクを中心に回転するパラボラ反射板を用いて，反射板の回転により変動した音響伝達特性を検出することで，音源の方向を推定する手法を提案した．焦点位置に単一マイクを備え付けた，パラボラ型の反射板が回転することで，反射板が音源方向を向いたときのみ音響伝達特性が大きく変動する．そこで，音響伝達特性が大きく変動する反射板の角度を検出することで，音源方向を推定する．類似した手法として，パワーに基づく手法と比較した結果，提案手法のほうが高い精度で音源方向を推定することができた．これは，音響伝達特性がクリーン音声に依存しないためである．しかし，本手法では音響伝達特性が完全に推定しきれていないため，その推定誤差が音源方向推定の誤差の原因となっている．また，本研究で用いた反射板は直径 24cm と大きく，実用をする際には反射板の縮小化が必要である．反射板が小さくなった場合，焦点に集まる反射波の数も減るため，音響伝達特性の変動も小さくなると考えられる．そのため，音響伝達特性の微小な変動であっても検出できるように，音響伝達特性のより正確な推定が必要であると考えられる．

第 4 章では，部屋の音響伝達特性が音源の位置に依存する点に着目し，音響伝達特性を識別することで，音源の位置を推定する手法を提案した．本章ではより正確に音響伝達特性を推定するために，クリーン音声を HMM でモデル化し，音響伝達特性の推定に用いた．実験により，第 3 章で用いていたクリーン音声 GMM よりも音響伝達特性の推定誤差が小さくなることが示された．また，位置の推定の際には，MFCC の次元毎にカーネル関数を定義することにより，特徴次元の重みを MKL により学習させた．提案手法では音源位置毎に，異なる次元重みのセットを学習することができ，従来の単一カーネル

SVM よりも高い識別精度を得ることができた．この手法は第 3 章で提案した手法とは異なり，反射板を用意する必要がなく，マイク一つのみで実装することができる．一方，推定したい位置毎にあらかじめ学習する必要があるため，今後は周辺の位置情報を用いた未学習の位置の補間を検討する．また，雑音環境下や，別の話者など，学習時と評価時の環境が異なる場合について，よりロバストな音源位置推定手法についても検討する．

第 5 章では，第 4 章で提案したシングルチャネル音源位置推定の枠組みを元に，各話者・各位置における残響音声の GMM から，その混合音声信号の GMM をモデル合成により求めて用いることで，2 話者の音源位置推定を単一マイクで行う手法について提案した．音響モデル合成を用いた手法として CRS モデルと DTRS モデルの二通りを検討し，1 話者及び 2 話者の音源位置推定実験により提案手法の有効性を評価した．実験では CRS モデルがモデル合成を用いない手法と比べて，特に音源位置毎の学習データが少量の場合において優位性を示しており，学習にユーザの発話データが必要な本手法において有効な手法であると言える．また，評価音声を長くすることで学習データが少量でも位置推定正解率が上昇することが確認できた．今後は音源が移動する場合のような，短い評価データの場合でもより高い正解率を得るために，最適な音響伝達特性のモデル化やモデル合成手法を検討する．

第 6 章では，音響伝達特性が話者の位置だけでなく頭部方向によっても異なる特性を持つ点に着目し，第 4 章で提案した手法と同様の枠組みを用いて話者の頭部方向の推定を行う手法について述べた．提案手法では，各音源位置・頭部方向において発話された音声信号から，その音響伝達特性をクリーン音声の音素 HMM を用いて推定し，推定された音響伝達特性を SVM により学習・識別することで，音源の位置と頭部方向をシングルチャネルで同時に推定する．実験では，音源位置の候補が 3 種類，頭部方向の候補が 0°，90°，180° の場合において，頭部方向のみの推定が平均 75.6% の認識率，音源位置と頭部方向の同時推定がが平均 66.0% の認識率でそれぞれ行えた．しかしながら，頭部方向の候補に 45° と 135° を追加した場合，認識率が大幅に低下していることから，現状の手法では細かい頭部の方向の識別は困難であることが分かる．本手法では音源位置・頭部方向毎に音響伝達特性を推定する必要があるが，より多くの音源位置・頭部方向の推定を行う場合，膨大な量の学習データを要するという問題がある．そのため，モデルの適応やオンライン学習といった，より少量の学習データで実装が行える手法について，今後検討を行う．

以上の研究から，音響伝達特性を用いることで，従来は困難な試みとされていたシングルチャネルによる音源位置・方向推定の可能性を示した．しかしながら，観測された信号から音響伝達特性を推定する際の誤差が，その後の音源位置・推定の性能の低下に繋がっており，従来のマイクロホンアレーに並ぶ精度を得るためには，より正確な音響伝達特性の推定が必要であると考えられる．また，第 4 章で提案した手法，及びその応用である第 5，第 6 章で提案した手法では，評価音声と同じ話者，同じ位置，同じ部屋環境で得られ

た学習データを用いているため，未学習の話者や位置，部屋環境の変化といった今後解決しなければならない問題が多く存在している．これらの問題を解決するために，不特定話者モデルを用いて話者の適応を行いながら音響伝達特性を推定する手法や，未学習の位置や環境におけるオンライン学習・適応といった手法についても今後検討していく．

謝辞

神戸大学大学院システム情報学研究科において，本研究を行う機会を与えてくださり，また本研究を進めるにあたり，有益な御指導，御教示をいただき本研究の副査をして頂いた 有木 康雄 教授，滝口 哲也 准教授に慎んで感謝の意を表すとともに，厚く御礼申し上げます．

お忙しい中，貴重な時間を割いて本論文の主査および副査をして頂いた神戸大学大学院システム情報学研究科 小島 史男 教授，玉置 久 教授に感謝致します．

また，日頃より研究生活を共にした神戸大学工学部情報知能工学科有木研究室の皆様に感謝致します．

参考文献

[1] X. Anguera, C. Wooters, and J. Hernando, “Acoustic beamforming for speaker diarization of meetings,” IEEE Trans. Audio, Speech and Language Processing, vol.15, pp.2011–2022, 2007.

[2] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, “A DOA based speaker diarization system for real meetings,” HSCMA2008,, pp.29–32, 2008.

[3] T. Yamagata, A. Sako, T. Takiguchi, and Y. Ariki, “System request detection in conversation based on acoustic and speaker alternation features,” Proc. Interspeech07, pp.2789–2792, 2007.

[4] 田中雅史，“マイクロホンアレー処理技術，”電子情報通信学会技術研究報告，第 95(62) 巻，pp.1–8，1995.

[5] 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，vol.J85-D-2，pp.1176–1187，2002.

[6] L.J. Griffiths and C.W. Jim, “An alternative approach to linearly constrained adaptive beamforming,” IEEE Transaction on Antennas and Propagation, vol.30, no.1, pp.27–34, 1982.

[7] 猿渡　洋，梶田将司，武田一哉，板倉文忠，“雑音適応型の相補的指向特性形成法を用いた音声強調，”電子情報通信学会技術研究報告，第 99(300) 巻，pp.1–8，1999.

[8] O. Hoshuyama, “A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters,” IEEE Transaction on Signal Processing, vol.47, no.10, pp.2677–2684, 1999.

[9] 西浦敬信，西岡良典，山田武志，中村哲，鹿野清宏，“CSP 法による音源位置同定を備えたマルチビームフォーミング，”電子情報通信学会論文誌，vol.J83-D-2，no.7，pp.1610–1619，2000.

[10] D. Johnson and D. Dudgeon, Array Signal Processing, Prentice Hall, 1996.

[11] F. Asano, H. Asoh, and T. Matsui, “Sound source localization and separation in

near field," IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences, vol.E83-A, pp.2286–2294, 2006.

[12] Y. Denda, T. Nishiura, and Y. Yamashita, "Robust talker direction estimation based on weighted CSP analysis and maximum likelihood estimation," IEICE Trans. on Information and Systems, vol.E89-D, pp.1050–1057, 2000.

[13] M. Omologo and P. Svaizer, "Acoustic event localization in noisy and reverberant environment using csp analysis," Proc. ICASSP96, pp.921–924, 1996.

[14] 西浦敬信, 山田武志, 中村哲, 鹿野清宏, "マイクロホンアレーを用いた csp 法に基づく複数音源位置推定," 電子情報通信学会論文誌, vol.J83-D-2, no.8, pp.1713–1721, 2000.

[15] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," Proc. ICASSP04, pp.817–820, 2004.

[16] B. Raj, M.V.S. Shashanka, and P. Smaragdis, "Latent direchlet decomposition for single channel speaker separation," Proc. ICASSP06, pp.821–824, 2006.

[17] G.-J. Jang, T.-W. Lee, and Y.-H. Oh, "A subspace approach to single channel signal separation using maximum likelihood weighting filters," Proc. ICASSP03, pp.45–48, 2003.

[18] T. Nakatani and B.-H. Juang, "Speech dereverberation based on probabilistic models of source and room acoustics," Proc. ICASSP06, pp.821–824, 2006.

[19] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," IEEE transaction on audio, speech, and language proceeding, vol.14, no.1, pp.191–199, 2006.

[20] 古井貞熙, ディジタル音声処理, 東海大学出版社, 神奈川, 1985.

[21] 有木康雄, "音声認識のフロントエンド," 日本音響学会誌, vol.66, no.1, pp.13–17, 2010.

[22] S.J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, The HTK Book Version 3.4, Cambridge University Press, 2006.

[23] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," IEEE Transaction on Acoustics, Speech and Signal Processing, vol.24, no.4, pp.320–327, 1976.

[24] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," Proc. ICASSP94, pp.273–276, 1994.

[25] 大賀寿朗, 山崎芳男, 金田豊, 音響システムとディジタル処理, 電子情報通信学会, 1995.

[26] 土屋義行, 王輝, 柴山秀雄, 谷本益巳, "リニアアレーを用いた移動音源の方向推定,"

電子情報通信学会技術研究報告，pp.21–28，2001.

[27] 中野振一郎，王輝，柴山秀雄，谷本益巳，“平面アレーを用いた静止音源の 2 次元方向推定，” 電子情報通信学会技術研究報告，第 100(724) 巻，pp.15–20，2001.

[28] 中川聖一，確率モデルによる音声認識，コロナ社，1988.

[29] L. Bahl, P. Brown, P. Souza, and L. Mercer, “Maximum mutual information estimation of hidden markov model parameters for speech recognition,” ICASSP, vol.11, pp.49–52, 1986.

[30] D. Povey, “Discriminative training for large vocabulary speech recognition,” PhD thesis, Cambridge University Engineering Dept, 2003.

[31] F. Keyrouz, Y. Naous, and K. Diepold, “A new method for binaural 3-D localization based on HRTFs,” Proc. ICASSP06, pp.V–341–V–344, 2006.

[32] M. Takimoto, T. Nishino, and K. Takeda, “Estimation of a talker and listener’s positions in a car using binaural signals,” The Fourth Joint Meeting ASA and ASJ, p.3216, 2006.

[33] A. Fuchs, C. Feldbauer, and M. Stark, “Monaural sound localization,” Proc. Interspeech 2011, pp.2521–2524, Florence, Italy, Aug. 2011.

[34] R. Kliper, H. Kayser, D. Weinshall, I. Nelken, and J. Anemuller, “Monaural azimuth localization using spectral dynamics of speech,” Proc. Interspeech 2011, pp.33–36, Florence, Italy, Aug. 2011.

[35] O. Ichikawa, T. Takiguchi, and M. Nishimura, “Sound source localization using a profile fitting method with sound reflectors,” IEICE transaction on information and systems, vol.E87-D(5), pp.1138–1145, 2004.

[36] N. Ono, Y. Zaitsu, T. Nomiyama, A. Kimachi, and S. Ando, “Biomimicry sound source localization with fishbone,” IEEJ Transaction on Sensors and Micromachines, vol.121-E, no.6, pp.313–319, 2001.

[37] 早川以久朗，郭少陽，及川靖広，山崎芳男，“耳の機構に倣った音源方向の推定，” 日本音響学会講演論文集，p.611，2006.

[38] 及川靖広，早川以久朗，鴇田泰弘，山崎芳男，“凹面反射板を用いた単一受音点での音源位置推定，” 日本音響学会講演論文集，p.641，2007.

[39] B. Saka and A. Kaderli, “Direction of arrival estimation and adaptive nulling in array-fed reflectors,” Electrotechnical Conference, pp.274–277, 1998.

[40] A. Sehr, R. Maas, and W. Kellermann, “Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition,” IEEE Trans. on Audio Speech and Language Processing, vol.18, no.7, pp.1676–1691, 2010.

[41] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," Proc. ICASSP08, pp.85–88, Las Vegas, Nevada, U.S.A., March 2008.

[42] A. Sehr and W. Kellermann, "Towards robust distant-talking automatic speech recognition in reverberant environments," Topics in Speech and Audio Processing in Adverse Environments, pp.679–728, Springer, Berlin, Germany, 2008.

[43] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," IEEE Trans. on Speech and Audio Processing, vol.4, no.3, pp.190–202, 1996.

[44] T. Kristiansson, B.J. Frey, L. Deng, and A. Acero, "Joint estimation of noise and channel distortion in a generalized EM framework," Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU01), pp.155–158, Trento, Italy, Dec. 2001.

[45] B.-H. Juang, "Maximum-likelihood estimation of mixture multivariate stochastic observations of Markov chains," AT&T Technical Journal, vol.64, no.6, pp.1235–1249, 1985.

[46] 小林哲則, 板橋秀一, 速水　悟, 竹澤寿幸, "日本音響学会研究用連続音声データベース," 日本音響学会誌, vol.48, no.12, pp.888–893, 1992.

[47] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "More efficiency in multiple kernel learning," Proc. ICML, pp.775–782, Corvallis, Oregon, USA, June 2007.

[48] T. Ogawa, H. Hino, N. Reyhani, N. Murata, and T. Kobayashi, "Speaker recognition using multiple kernel learning based on conditional entropy minimization," ICASSP, pp.2204–2207, 2011.

[49] C. Longworth and M.J.F. Gales, "Multiple kernel learning for speaker verification," Proc. ICASSP08, pp.1581–1584, Las Vegas, Nevada, USA, March 2008.

[50] V. Rapp, T. Senechal, K. Bailly, and L. Prevost, "Multiple kernel learning SVM and statistical validation for facial landmark detection," Automatic Face Gesture Recognition and Workshops, pp.265–271, 2011.

[51] P. Liang, G. Teodoro, L. Haibin, E. Blasch, G. Chen, and L. Bai, "Multiple kernel learning for vehicle detection in wide area motion imagery," International Conference on Information Fusion, pp.1629–1636, 2012.

[52] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," Proc. ICCV07, pp.1–8, Rio de Janeiro, Brazil, Oct. 2007.

[53] A.D. Dileep and C.C. Sekhar, "Representation and feature selection using multiple kernel learning," Proc. International Joint conference on Neural Networks, pp.717–722, 2009.

[54] T. Joachims, Making large-scale SVM learning practical, B. Scholkopf, C. Burges, and A. Smola, eds., MIT Press, Cambridge, MA, USA, 1999.

[55] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," Journal of Machine Learning Research, vol.9, pp.2491–2521, 2008.

[56] 阿部匡伸，匂坂芳典，梅田哲夫，桑原尚夫，"研究用日本語音声データベース利用解説書，" Technical report，ATR 自動翻訳電話研究所，1990.

[57] S. Nakamura, "Acoustic sound database collected for hands-free speech recognition and sound scene understanding," Proc. International Workshop on Hands-Free Speech Communication (HSC01), pp.43–46, Kyoto, Japan, April 2001.

[58] K. Yu, M.J.F. Gales, and P.C. Woodland, "Unsupervised adaptation with discriminative mapping transforms," IEEE Trans. on Audio Speech and Language Processing, vol.17, pp.714–723, 2009.

[59] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hmms," Computer Speech and Language, vol.9, pp.171–16, 1995.

[60] T. Takiguchi, M. Nishimura, and Y. Ariki, "Acoustic model adaptation using first-order linear prediction for reverberant speech," IEICE Trans. INF. and SYST., vol.E89-D, pp.908–914, 2006.

[61] M.J.F. Gales, "Predictive model-based compensation schemes for robust speech recognition," Speech Communication, vol.25, pp.55–64, 1998.

[62] A. Brutti, M. Omologo, and P. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays," Proc. Interspeech05, pp.2337–2340, 2005.

[63] J.M. Sachar and H.F. Silverman, "A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array," Proc. ICASSP04, vol.4, pp.65–68, 2004.

[64] C. Segura, A. Abad, J. Hernando, and C. Nadeu, "Speaker orientation estimation based on hybridation of GCC-PHAT and HLBR," Proc. Interspeech08, pp.1325–1328, 2008.

[65] M. Togami and Y. Kawaguchi, "Head orientation estimation of a speaker by utilizing kurtosis of a DOA histogram with restoration of distance effect," Proc. ICASSP10, pp.133–136, 2010.

[66] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, “An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses,” J. Acoust. Soc. Am., vol.97(2), pp.1119–1123, 1995.

発表論文

論文（査読あり）

1. R. Takashima, T. Takiguchi, Y. Ariki, “Dimensional feature weighting utilizing Multiple Kernel Learning for single-channel talker location discrimination using the acoustic transfer function,” Journal of the Acoustical Society of America, Volume 133, Issue 2, pp.-, 2013. (Accepted)
2. R. Takashima, T. Takiguchi, Y. Ariki, “Single-channel talker localization based on the separation of the acoustic transfer function using Hidden Markov Model and its classification,” Journal of the Acoustical Society of Japan, Volume -, No. -, pp. -, 2013. (Accepted)
3. 高島遼一，滝口哲也，有木康雄，“音響モデル合成を用いた単一マイクによる 2 話者位置推定,” 電子情報通信学会論文誌, Volume J96-D, No. 3, pp.-, 2013. (採録決定)
4. R. Aihara, R. Takashima, T. Takiguchi, Y. Ariki, “GMM-Based Emotional Voice Conversion Using Spectrum and Prosody Features,” American Journal of Signal Processing, Volume 2, No. 5, pp. 134-138, 2012.
5. R. Takashima, T. Takiguchi, Y. Ariki, “Evaluation of an Active Microphone with a Parabolic Reflection Board for Estimating Sound Source Direction,” Far East Journal of Electronics and Communications, Volume 5, Issue 2, pp. 91-104, 2010.
6. R. Takashima, T. Takiguchi, Y. Ariki, “Monaural Sound-source-direction Estimation Using the Acoustic Transfer Function of a Parabolic Reflection Board,” Journal of the Acoustical Society of America, Volume 127, Issue 2, pp. 902-908, 2010.
7. Y. Ariki, T. Takiguchi, T. Muroi, R. Takashima, “Speech Feature Extraction Using Weighted Higher-order Local Autocorrelation,” Far East Journal of Electronics and Communications, Volume 3, Issue 2, pp. 125 - 140, 2009.
8. T. Takiguchi, Y. Sumida, R. Takashima, Y. Ariki, “Single-Channel Talker Localization Based on Discrimination of Acoustic Transfer Functions, ” EURASIP Journal on Advances in Signal Processing Volume 2009 (2009), Article ID 918404, 9 pages.

著書

1. “Advances in Sound Localization” (Chapter 3: R. Takashima, T. Takiguchi, Y. Ariki, “ Single-Channel Talker Localization Based on Discrimination of Acoustic Transfer Functions ”) Intech Open Publisher, pp. 39-54, 2011.

国際会議（査読あり）

1. R. Takashima, T. Takiguchi, Y. Ariki, “Exemplar-Based Voice Conversion in Noisy Environment,” IEEE Workshop on Spoken Language Technology (SLT2012), pp. 313-317, 2012-12.

2. R. Aihara, R. Takashima, T. Takiguchi, Y. Ariki, “Consonant Enhancement for Articulation Disorders Based on Non-negative Matrix Factorization,” APSIPA2012, 4 pages, 2012-12.
3. T. Yoshioka, R. Takashima, T. Takiguchi, Y. Ariki, “Robust Feature Extraction to Utterance Fluctuations Due to Articulation Disorders Based on Sparse Expression,” APSIPA2012, 4 pages, 2012-12.
4. T. Takiguchi, T. Imada, R. Takashima, Y. Ariki, J.-F. L. Lin, P. K. Kuhl, M. Kawakatsu, M. Kotani, “An AdaBoost-Based Weighting Method for Localizing Human Brain Magnetic Activity,” APSIPA2012, 4 pages, 2012-12.
5. R. Takashima, T. Takiguchi, Y. Ariki, “Estimation of Talker ’ s Head Orientation Based on Discrimination of the Shape of Cross-power Spectrum Phase Coefficients,” Interspeech 2012, 2012-09.
6. T. Takiguchi, T. Imada, R. Takashima, Y. Ariki, J.-F. L. Lin, P. K. Kuhl, M. Kawakatsu, M. Kotani, “ A New Multiple-Kernel-Learning Weighting Method for Localizing Human Brain Magnetic Activity,” ICASSP2012, pp. 761-764, 2012-03.
7. R. Takashima, T. Takiguchi, Y. Ariki, “Single-channel Head Orientation Estimation Based on Discrimination of Acoustic Transfer Function,” INTERSPEECH 2011, pp. 2721-2724, 2011.
8. R. Takashima, T. Nagano, R. Tachibana, M. Nishimura, “Agglomerative Hierarchical Clustering of Emotions in Speech Based on Subjective Relative Similarity,” INTERSPEECH 2011, pp. 2473-2476, 2011.
9. R. Takashima, T. Takiguchi, Y. Ariki, “Feature Selection Based on Multiple Kernel Learning for Single-channel Sound Source Localization Using the Acoustic Transfer Function,” ICASSP2011, pp. 2696-2699, 2011.
10. R. Takashima, T. Takiguchi, Y. Ariki, “HMM-based Separation of Acoustic Transfer Function for Single-channel Sound Source Localization,” ICASSP 2010, pp. 2830-2833, 2010.
11. T. Muroi, R. Takashima, T. Takiguchi, Y. Ariki, “Gradient-Based Acoustic Features for Speech Recognition,” ISPACS 2009, pp. 445-448, 2009.
12. R. Takashima, T. Takiguchi, Y. Ariki, “Single-Channel Multi-Talker-Localization Based on Maximum Likelihood,” SSP 2009, pp. 461-464, 2009.
13. R. Takashima, T. Takiguchi, Y. Ariki, “Monaural Sound-Source-Direction Estimation Using the Acoustic Transfer Function of an Active Microphone,” ICIF 2009, pp. 48-53, 2009.
14. T. Takiguchi, R. Takashima, Y. Ariki, “Active Microphone with Parabolic Reflection Board for Estimation of Sound Source Direction,” HSCMA2008, pp. 65-68, 2008.
15. T. Takiguchi, R. Takashima, Y. Ariki, “Estimation of Sound Source Direction Using Parabolic Reflection Board,” NCSP 2008, pp. 9-12, 2008.

口頭発表（査読なし）

1. 高島遼一，滝口哲也，有木康雄，“音響伝達特性を用いたシングルチャネル音源位置推定に

おける局所的回帰に基づく未学習位置の補間,” 電子情報通信学会技術研究報告，vol. 112, no. 369，SP2012-92，pp. 75-80，2012-12.
2. 高島遼一，滝口哲也，有木康雄，“スパース表現を用いた雑音環境下の声質変換,” 日本音響学会 2012 年秋季研究発表会，1-2-1，pp.213-216，2012-09.
3. 吉岡利也，高島遼一，滝口哲也，有木康雄，李義昭，“構音障害者の音素認識誤りの傾向,” 日本音響学会 2012 年秋季研究発表会，3-P-9，pp.140-141，2012-09.
4. 相原龍, 高島遼一, 滝口哲也, 有木康雄，“非負値行列因子分解による構音障害者の声質変換,” 日本音響学会 2012 年秋季研究発表会，3-2-5，pp.331-334 ，2012-09.
5. 石井良, 高島遼一, 滝口哲也, 有木康雄, 中井靖, 高田哲，“音響特徴量を用いた自閉症児と定型発達児の識別,” 日本音響学会 2012 年秋季研究発表会，3-P-1, pp.117-118, 2012-09
6. R. Takashima，T. Takiguchi，Y. Ariki，T. Imada，J.F. L. Lin，P. K. Kuhl，M. Kawakatsu，M. Kotani，“An AdaBoost-Based Weighting Method for Localizing Human Brain Magnetic Activity,” 日本音響学会 2012 年春季研究発表会，3-Q-26，pp. 649-650，2012-03.
7. 高島遼一，滝口哲也，有木康雄，“音響伝達特性を用いたシングルチャネル音源位置推定における未学習位置の推定,” 日本音響学会 2012 年春季研究発表会，1-Q-8，pp. 837-840，2012-03.
8. 相原龍, 高島遼一, 滝口哲也, 有木康雄, “スペクトルと韻律を特徴量とした GMMによる感情音声変換,” 日本音響学会 2012 年春季研究発表会, 1-R-29, pp.503-504, 2012-03.
9. 吉岡利也, 高島遼一, 滝口哲也, 有木康雄, “スパース表現に基づく構音障害者の発話スタイル変動にロバストな特徴量抽出,” 日本音響学会 2012 年春季研究発表会, 1-P-4, pp.127-128, 2012-03.
10. 高島遼一，滝口哲也，有木康雄，“音響伝達特性を用いた話者の頭部方向推定の検討,” 第 14 回音響学会関西支部若手研究者交流研究発表会, 45, 2011-12.
11. 高島遼一，滝口哲也，有木康雄, “2ch マイクによる CSP 係数の識別に基づく話者の頭部方向の推定,” 日本音響学会 2011 年秋季研究発表会，1-4-8，pp. 619-622，2011-09.
12. 高島遼一，滝口哲也，有木康雄, “音響伝達特性を用いた単一マイクロホンによる話者の頭部方向の推定” 日本音響学会 2011 年秋季研究発表会，1-4-7，pp. 615-618，2011-09.
13. 高島遼一，滝口哲也，有木康雄, “CSP 係数の識別に基づく話者の頭部方向推定の検討,” 電子情報通信学会技術研究報告, vol. 111, no. 153, pp. 57-62, 2011-07.
14. 高島遼一，滝口哲也，有木康雄, “音響伝達特性の識別に基づく話者の頭部回転方向の推定,” 電子情報通信学会技術研究報告, vol. 111, no. 28, pp. 167-172, 2011-05.
15. 高島遼一，滝口哲也，有木康雄, “音響伝達特性の判別に基づく単一チャネル音源位置推定における MKL-SVM を用いた特徴量重みの自動学習,” 日本音響学会 2011 年春季研究発表会，3-P-59(c)，pp. 949-952，2011-03.
16. 高島遼一，滝口哲也，有木康雄, “音響伝達特性を用いた単一チャネル音源位置推定における特徴量選択の検討,” 電子情報通信学会技術研究報告, vol. 110, no. 401, pp. 49-54, 2011-01.
17. 高島遼一，滝口哲也，有木康雄, “音響伝達特性を用いた単一チャネル音源位置推定における特徴量選択の検討,” 日本音響学会 2010 年秋季研究発表会，2-10-8，pp. 583-584，2010-09.
18. 高島遼一，滝口哲也，有木康雄, “残響適応パラメータを用いた単一チャネル音源位置推定の検討,” 日本音響学会 2010 年春季研究発表会，2-P-3，pp. 795-796，2010-03.

19. 高島遼一，滝口哲也，有木康雄, “HMM を用いた音響伝達特性の推定と音源位置推定,” 日本音響学会 2009 年秋季研究発表会，3-Q-13，pp. 769-770，2009-09.
20. 高島遼一，滝口哲也，有木康雄, “パラボラ反射板による音響伝達特性の変化を用いたシングルチャネル音源方向推定,” 日本音響学会 2009 年春季研究発表会，3-P-5，pp. 767-768，2009-03.
21. 高島遼一，滝口哲也，有木康雄, “音響伝達特性モデルを用いたシングルチャネル音源位置推定の検討,” 日本音響学会 2009 年春季研究発表会，2-P-34，pp. 755-756，2009-03.
22. 高島遼一，滝口哲也，有木康雄, “アクティブマイクロフォンによる音響伝達特性を用いたシングルチャネル音源方向推定,” 日本音響学会 2008 年秋季研究発表会，3-P-23，pp. 811-812，2008-09.
23. 高島遼一，滝口哲也，有木康雄, “パラボラ反射板を用いたアクティブマイクロフォンによる音源方向推定,” 日本音響学会 2008 年春季研究発表会，1-P-6，pp.765-766，2008-03.