リアルタイム性を考慮したスパース尺度に基づく音源到来方向推定に関する研究

著者	岩? 宣生
発行年	2015
その他のタイトル	Studies on Real-Time Oriented Sound Source DOA Estimation Based on Sparseness
学位授与年度	平成26年度
学位授与番号	17104甲情工第295号
URL	http://hdl.handle.net/10228/5455

リアルタイム性を考慮したスパース尺度に基づく音源到来方向推定に関する研究

岩﨑 宣生

目 次

図 目 次

表 目 次

第1章

序論

雑音を抑圧してクリアな音声信号を抽出することは，音響信号処理の黎明期からの課題であり，現在まで様々なアプローチによる雑音除去法が提案されている．雑音除去法は，単一のマイクロホンを用いる方法と複数のマイクロホンを用いる方法に分かれる．前者の例としては，スペクトルサブトラクション(Spectral Subtraction)法[1][2]が良く知られており，定常雑音に対し雑音区間が既知ならば，比較的簡単な処理で高い雑音除去効果が得られる．しかし，工場内や車の行き交う雑踏では勿論のこと，ありふれた生活音や活動音しか存在しない一般家庭やオフィスでも，多くの雑音は非定常信号であり，雑音区間も未知であることが多い．そのため，スペクトルサブトラクション法により期待通りの雑音除去性能を得ることは難しい．

後者の例としては，適応ビームフォーミング(ABF: Adaptive Beam Forming)により指向特性を制御する方法がある[3][4][5][6][7][8][9]．複数のマイクロホンを用いれば，音の空間的な情報を得ることができる．すなわち，音の到来方向(DOA: Direction Of Arrival)や距離などの情報を利用することで，複数の音の空間的な性質の違いに基づいた分離が可能になり，非定常雑音にも適用できる．このようにABFは，空間的なフィルタと位置づけられ，特定の方向にフィルタの通過域を向け，その方向から到来する信号だけを通過させる技術である．したがって，ABFの分離能力は信号のDOAの推定精度に左右されるため，制約や計算コストが少なく，高精度なDOA推定法が望まれている．

これまで，多数のDOA推定法が提案されてきている．相互相関関数[10][11][12]を用いる方法や，遅延和法(DSB: Delay and Sum Beamformer)[6][13]などは，比較的簡単に

DOA推定できる方法として広く利用されている．しかし，これらの方法には，雑音や残響の影響を受けやすく，空間分解能が低いという欠点がある．高い空間分解能をもつ方法として，MUSIC（MUltiple SIgnal Classification）法[14][15]やESPRIT（Estimation of Signal Parameters via Rotational Invariance Technique）法[16][17]があり，レーダーやソナー，地震探査などの分野でも応用されている．これらは遠距離場の狭帯域信号を対象に部分空間法に基づいて開発された方法であり，音声のように広帯域の信号源に対して適用する研究も進められているが[13][18]，この場合，DOA推定結果に周波数依存性が見られる．特に，低周波数域ではDOA推定精度が著しく低下するため，使用可能な周波数を限定する必要がある．また，部分空間法では信号数が既知であることを前提にしている．さらに，マイクロホン数が信号源数より少ない場合では機能しない．すなわち，DOAが推定できる信号数はマイクロホン数より少ない個数に限定される．この問題は，マイクロホン数を増やせば解決できると考えられるが，計算コストの増加やシステム全体の巨大化といった問題を新たに生じさせることになる．したがって，使用するマイクロホン数は，可能な限り少数が好ましい．

マイクロホン数が信号源数より少ない場合でも，広帯域信号のDOA推定ができる方法として，信号のスパース性を利用する方法がある[19][20]．スパース性は，“ W-Disjoint Orthogonality ”とも称され，すべての時間周波数で支配的な信号は高々１個しか存在しないことを意味している[21][22][23][24][25][26][27][28]．Rickardらは，DUET（Degenerate Unmixing and Estimation Technique）法[29][30]を提唱し，これに基づくDOA推定法を提案している[31]．DUET原理に基づくDOA推定[32][33]は，無響室下であれば，クラスタリングが良好に行えるため，信号源数が未知の場合でも有効に機能する．しかし，残響下では，信号源数を既知としてクラスタリングを行う必要がある[34][35]．DUET原理に基づく場合，このように多数のデータをため込んでクラスタリングを行うことが前提となる．しかし，残響下でクラスタリングを正確に行うのは容易でない．また，現実的には，信号源数が既知である場合は少ない．信号源数の推定に関する研究も行われているが，信号源数が時間とともに増減する場合は難しいと考えられる．これは，多数のデータを必要とすることに起因している．すなわち，バッチ処理的な方法では，突発音などの瞬間的な音や，移動音源への対応は難しい．移動音源の追跡については，パーティクルフィルタなど[13]による研究が知られているが，状態遷移関数と尤度関数の適

切な設計が必須となり，複雑な動きや急激な変化に対応できるまでには至っていない．

以上の観点から，本論文では，広く普及しているステレオICレコーダ等の利用を念頭に，DUET原理に基づいてフレーム単位に音源のDOAを推定する方法を提案する．データをため込むことなく，フレーム毎にDOAが推定できれば，話者が揺れ動きながら話すときや，マイクロホン内蔵の携帯端末の対面角度が変化するときでも，話者のDOAをリアルタイムに追跡することが期待できる。

以降の内容について順を追って説明し，本論文の構成を明らかにする．

第**2**章では，音源のDOA推定について概説する．具体的には，代表的なDOA推定法であるDSB法，MUSIC法，スパース性に基づいたDUET法の基本原理を説明し，各方法の長所や短所を明らかにして，実際への適用に関しての課題を述べる．特に，クラスタリング等のバッチ処理を必要とする方法では，突発音などの瞬間的な音や，移動音源のDOAを推定することは難しいことを指摘する．言い換えると，突発音などの瞬間的な音や，移動音源への適用を行うには，フレーム単位にDOAを推定することが必須である．

第**3**章では，音声のDOA推定に関して予備的な検討を行う．具体的には，まず，音声は音声区間と無音区間を繰り返す断続波であり，音声のDOAに関する情報は，音声区間にだけ含まれ，無音区間には含まれないが，現実には無関係な雑音成分が無音区間に重畳することを述べる．次に，フレーム数が少ない場合，無音区間が雑音や残響の影響を受けることから，音声区間内の連続する20フレームでの局所DOAをもとに，DUET法に基づく方法とMUSIC法によりDOA推定した場合，どの程度の推定精度が得られるかをシミュレーションにより検討する．そして，シミュレーション結果とCramer-Rao Boundの観点から，DUET法に基づくDOA推定法のMUSIC法に対する優位性を明らかにする．

第**4**章では，DUET法に基づくフレーム単位のDOA推定法を提案する．ここでは，まず，マイクロホンで観測された混合信号は，無音源区間，１音源区間，複数音源区間に分けられることを述べる．次に，混合信号を短時間離散フーリエ変換して得られる複素スペクトルの位相差をもとに，フレーム単位に時間周波数における局所DOAを求めて，その頻度分布をとれば３種類の形状に分類されることを指摘する．すなわち，１音源フレームでは１つのピークをもつ単峰的な分布，２音源フレームでは２つのピークをもつ双峰的な分布，無音源フレームではピークのない比較的平坦な分布となることを明らかにする．さらに，１音源フレームではピークが明確な単峰的な分布となるこ

とを利用して，１音源フレームを選別した後，その分布のピークを探索することでDOAを推定するという提案アプローチの流れを説明する．この場合，１音源フレームの選別が重要である．そこで，分布の形状を測る尺度として，Hoyer尺度などのスパース尺度を採択し，その尺度に適切な閾値を設けることで１音源フレームが選別できることをシミュレーションにより示す．最後に，Hoyer尺度による１音源フレームの検出結果と実際の音声区間を比較して，Hoyer尺度の閾値を0.5とすれば選別されたフレームは全て音声区間内に収まることを確認する．

第**5**章では，提案法の有効性をシミュレーションにより検証する．まず，残響時間が200[msec]，SN比(SNR: Signal-Noise Ratio)が20[dB]の環境下で，単一音源のDOA推定を対象にしたシミュレーションを行い，提案法はDUET法に基づく方法とほぼ同等の精度でDOA推定できることを確認する．具体的には，提案法は，目的音源の方位をブロードサイド(マイクロホン正面)方位から±30°の範囲に絞り込めば，推定誤差は2°未満，標準偏差は3°程度の精度でフレーム単位にDOA推定できることを確認する．次に，残響時間とSN比を変えて，提案法が有効に機能する範囲をシミュレーションにより調べ，提案法を有効に機能させるには，残響時間が250[msec] 以下でSN比が15[dB]以上の環境で使用する必要があることを明らかにする．以上の準備のもとで，残響時間が250[msec] 以下でSN比が15[dB]以上として，目的音源の方位を±30°の範囲に絞り込み，３音源が存在する環境で，提案法により目的音源のDOAを推定するシミュレーションを行う．その結果，残響時間が200[msec]以下でSN比が15[dB]以上あれば，目的音源のDOA推定値の誤差が3°未満，標準偏差が3°程度の精度で推定できることを述べる．

第**6**章では，以上の内容を整理してまとめるとともに，今後の研究課題や展開を述べて，本論文を総括する．

第2章

DOA推定の概要

音源の到来方向（DOA：Direction of arrival）を推定する代表的な方法として，DSB法，MUSIC法，スパース性に基づくDUET法がある．ここでは，それぞれの根本原理やアルゴリズムを説明するとともに長所や短所を以下のように明らかにする．すなわち，DSB法はアルゴリズムが簡単であるがDOA推定精度は低く単一音源のときにしか機能しないことを述べる．また，MUSIC法はDOA推定精度が高いが，マイクロホン数が音源数より少ない状況では機能しないことを述べる．一方，DUET法に基づくDOA推定の場合，マイクロホン数が音源数より少なくても機能するが，DSB法やMUSIC法と同様にバッチ処理による方法であるため，突発音などの瞬間的な音や移動音源のDOA推定に適用することは難しいことを指摘する．

以上のことから，突発音などの瞬間的な音や，移動音源のDOAを推定するためには，フレーム単位でのDOA推定が必須であることを明らかにする．

2.1 遅延和ビームフォーマ法

M個の無指向性マイクロホンがd[m]の等間隔で直線上に並んだ線形マイクロホンアレイに，１つの音源$s(t)$がブロードサイド方向(線形マイクロホンアレイの正面方向)から測って角度θ_sの方位から平面波として到来する場合を考える（Fig. 2.1）．音源からマイクロホンまでの伝搬波は，音源からマイクロホンアレイの中心までの距離γ[m]がマイクロホン間隔dに比べて十分長い場合（遠方場：Far field），平面波として取り扱われ，

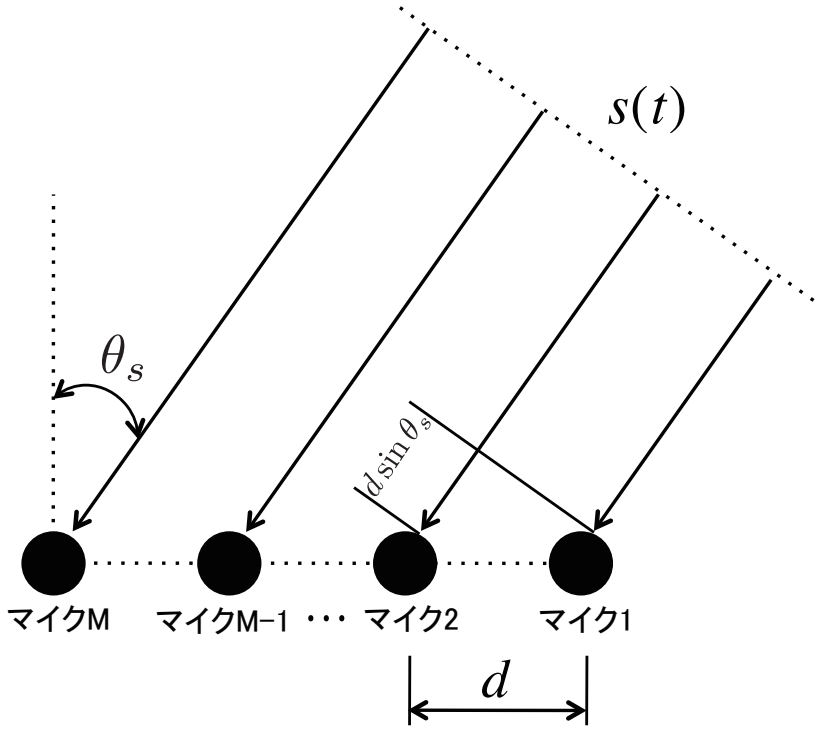


図 2.1 遠方場における直線上等間隔アレイ

そうでない場合（近傍場：Near field），球面波として取り扱われる．近傍場の目安は大まかに$\gamma < 2D^2/\lambda$[m]と与えられる[13]．ここに，$D = (M-1)d$[m]はアレイの最大幅，λ[m]は伝搬波の波長である．本節と次節では，遠方場の立場から，伝搬波を平面波と見なして議論する．この場合，m番目のマイクロホンでは，音源からの入射波が1番目のマイクロホンに比べて$(m-1)\tau$遅れて

$$x_m(t) \;=\; s(t-(m-1)\tau) \quad m = 1, 2, \cdots, M \tag{2.1}$$

のように観測される．ここに，τは，音速をc[m/sec]として，

$$\tau \;=\; d\sin\theta_s/c \tag{2.2}$$

と定義される隣接マイクロホン間の到達時間差（遅延時間）である．

上式の時間領域表現をフーリエ変換して周波数領域に直すと，

$$x_m(\omega) \;=\; e^{-j\omega(m-1)\tau}s(\omega) \quad m = 1, 2, \cdots, M \tag{2.3}$$

となる．ここに，$x_m(\omega)$と$s(\omega)$はそれぞれ$\{x_m(t)\}$と$\{s(t)\}$をフーリエ変換したもので，ωは

(角)周波数である．また，式(2.3)をベクトル表記すれば，

$$\boldsymbol{x}(\omega) \quad = \quad \boldsymbol{a}s(\omega) \tag{2.4}$$

となる．ここに，$\boldsymbol{x}(\omega) = [x_1(\omega), x_2(\omega), \cdots, x_M(\omega)]^T$，$\boldsymbol{a} = [1, e^{-j\omega\tau}, \cdots, e^{-j\omega_(M-1)\tau}]^T$，$\boldsymbol{T}$は転置記号である．遅延時間$\tau$は，式(2.2)から分かるように，音源の到来方向$\theta_s$に依存して決まる．それゆえ，$\boldsymbol{a}$も$\theta_s$に依存して決まる．そこで，このことを明示すため，$\boldsymbol{a}$を$\boldsymbol{a}(\theta_s)$と書改めて，方位$\theta_s$からの音源に対する到来方向ベクトルと呼ぶことにする．

以上の準備のもとで，各マイクロホンでの観測値$\boldsymbol{x}(\omega)$を荷重$\boldsymbol{w} = [w_1, w_2, \cdots, w_M]^T$で重みづけた総和をアレイ出力

$$y(\omega) \quad = \quad \boldsymbol{w}^H\boldsymbol{x}(\omega) \tag{2.5}$$

と定義し，アレイ応答関数を

$$G(\omega) \quad = \quad y(\omega)/s(\omega) = \boldsymbol{w}^H\boldsymbol{a}(\theta_s) \tag{2.6}$$

と定義する．ここに，Hは共役転置を表す．このとき，$\|\boldsymbol{w}\|^2 = M$なる拘束のもとでアレイ応答関数$G(\omega)$が最大となる$\boldsymbol{w}$，すなわち，評価関数

$$J(\omega) \quad = \quad |\boldsymbol{w}^H\boldsymbol{a}(\theta_s)|^2 + \epsilon(\|\boldsymbol{w}\|^2 - M) \tag{2.7}$$

を最大にする$\boldsymbol{w}$を求めると，

$$\boldsymbol{w} \quad = \quad \boldsymbol{a}(\theta_s) \tag{2.8}$$

なる最適解が得られる．ここに，ϵはラグランジュの未定定数である．

この場合，式(2.5)より，アレイ出力は

$$\begin{aligned} y(\omega) \quad &= \quad \boldsymbol{a}(\theta_s)^H\boldsymbol{a}(\theta_s)s(\omega) \\ &= \quad [1, e^{j\omega\tau}, \cdots, e^{j\omega_(M-1)\tau}][1, e^{-j\omega\tau}, \cdots, e^{-j\omega_(M-1)\tau}]^T s(\omega) \\ &= \quad Ms(\omega) \end{aligned} \tag{2.9}$$

と展開されることから，各マイクロホンで生じた時間遅れ$(m-1)\tau$（等価的に位相遅れ$e^{-j\omega(m-1)\tau}$）はその共役を乗じることにより同相化されることになる．その結果，θ_s方向

からの到来波は，同相化されて加算され，M倍の大きさに強調される．一方，その他の方向からの到来波は，位相がずれて加算されることになり弱められる．言い換えると，アレイ全体の指向特性（ビーム）がθ_s方向に対して形成される．このように各マイクロホンでの観測値を同相化して源信号を強調する方法を遅延和ビームフォーマ（DSB：Delay-and-Sum Beamformer）法と云う．

いま，試行的に仮定した方位（試行方位）をθと表記し，試行方位θに対して，

$$\boldsymbol{w}(\theta) \quad = \quad [1, e^{-j\omega d\sin\theta/\mathrm{c}}, \cdots, e^{-j\omega(M-1)d\sin\theta/\mathrm{c}}]^T \tag{2.10}$$

なるベクトルを定義する．$\boldsymbol{w}(\theta)$は，試行方位θを–90°から90°の範囲で少しずつ変え（回転させ）ながら逐次生成していくことから，ステアリング・ベクトルと呼ばれる．便宜的に，$\boldsymbol{w}(\theta)$と$\boldsymbol{a}(\theta_s)$は両方ともステアリング・ベクトルと呼ばれることがあるが，ここでは，論旨を明確にするため，前者をステアリング・ベクトル，後者を到来方向ベクトルと区別する．

このとき，DSB法に基づいて，未知の到来方向θ_sを–90°から90°の範囲で探索して推定する手順をまとめると，次のようになる．

i) 試行方位θを–90°から90°の範囲で逐次変えながら，θに対するステアリング・ベクトル$\boldsymbol{w}(\theta)$を生成し，アレイ出力を$y(\theta,\omega) = \boldsymbol{w}(\theta)^H\boldsymbol{x}(\omega)$と求める．

ii) アレイ出力$y(\theta,\omega)$ $(-90° < \theta < 90°)$が最大となるときのθを探索して，その方位θを音源の到来方向θ_sの推定値として採択する．

2.2 MUSIC法

DSB法は，上述のように音源が単一の場合，有効に機能するが，複数の音源が同時にマイクロホンアレイに入射する場合，機能しない．しかし，複数の音源が同時に入射する環境下でも，音源の個数よりマイクロホンの個数が多ければ，個々の音源のDOAは推定できる．その代表的な方法として，MUSIC(MUltiple SIgnal Clasification)法があり，DSB法に比べて，計算量は増えるが，推定精度（空間分解能）は高い[14][17].

MUSIC法を説明するため，Fig. 2.1のアレイマイクロホンに，N（$< M$）個の音源$s_n(t)$ $(n = 1, 2, \cdots, N)$からの平面波がそれぞれθ_nの方向から入射する場合を考える．この場

合，m番目のマイクロホンでの観測値は

$$x_m(t) = \sum_{n=1}^{N} s_n(t-(m-1)\,\tau_n) + v_m(t) \quad for \quad m = 1, 2, \cdots, M \tag{2.11}$$

と与えられる．ここに，τ_n (=$d\sin\theta_n/c$)は方位θ_nからの到来波に対する隣接マイクロホン間での到達時間差，$v_m(t)$は雑音である．また，この時間領域表現を周波数領域表現に

$$x_m(\omega) = \sum_{n=1}^{N} s_n(\omega)e^{-j\omega(m-1)\tau_n} + v_m(\omega) \quad for \quad m = 1, 2, \cdots, M \tag{2.12}$$

と直して，さらにベクトル表記すると

$$\boldsymbol{x}(\omega) = \boldsymbol{A}\boldsymbol{s}(\omega) + \boldsymbol{v}(\omega) \tag{2.13}$$

となる．ここに，$\boldsymbol{s}(\omega) = [s_1(\omega), s_2(\omega), \cdots, s_N(\omega)]^T$，$\boldsymbol{A} = [\boldsymbol{a}(\theta_1), \boldsymbol{a}(\theta_2), \cdots, \boldsymbol{a}(\theta_N)]$，$\boldsymbol{a}(\theta_n)$ (=$[1, e^{-j\omega d\sin\theta_n/c}, \cdots, e^{-j\omega(M-1)d\sin\theta_n/c}]^T$) は方位が$\theta_n$の音源$s_n(t)$に対する到来方向ベクトル，$\boldsymbol{v}(\omega) = [v_1(\omega), v_2(\omega), \cdots, v_M(\omega)]^T$である．

以上のもとで，音源$s_n(t)$ ($n = 1, 2, \cdots, N$)と雑音$v_m(t)$ ($m = 1, 2, \cdots, M$)について，

i) $s_n(t)$は互いに無相関

ii) $s_n(t)$と$v_m(t)$は無相関

iii) $v_m(t)$はすべて分散が同一(σ^2)の白色雑音

と仮定する．このとき，観測値$\boldsymbol{x}(\omega)$の相関行列を$\boldsymbol{R} = E[\boldsymbol{x}(\omega)\boldsymbol{x}(\omega)^H]$と定義すると，$\boldsymbol{R}$は，式(2.13)より，

$$\boldsymbol{R} = \boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^H + \sigma^2 I \tag{2.14}$$

と与えられることが導かれる．ここに，$\boldsymbol{S} = E[\boldsymbol{s}(\omega)\boldsymbol{s}(\omega)^H]$，$E[\cdot]$は期待値，$I$は単位行列を表す.

また，$\boldsymbol{R}$の固有値と固有ベクトルを

$$\boldsymbol{R}\boldsymbol{g}_{m'} = \lambda_{m'}\boldsymbol{g}_{m'} \quad m' = 1, 2, \cdots, M \tag{2.15}$$

と定義して(便宜のため，固有ベクトルは正規直交化されたものとする．)，固有値$\lambda_{m'}$ $(m' = 1, 2, \cdots, M)$を大きい順に並べ替えると，

$$\lambda_1 \geq \cdots \geq \lambda_N > \lambda_{N+1} = \cdots = \lambda_M = \sigma^2 \tag{2.16}$$

となって，$\boldsymbol{R}$の固有値は雑音の分散σ^2に等しいグループ$\{\lambda_m | m = N+1, N+2, \cdots, M\}$と$\sigma^2$より大きなグループ$\{\lambda_m | m = 1, 2, \cdots, N\}$に区分されることが分かる[13]．この場合，各音源のパワーを大きい順にρ_n^2 $(n = 1, 2, \cdots, N)$とすると，値の大きな方のN個の固有値は，$\lambda_n = \rho_n^2 + \sigma^2$ $(n = 1, 2, \cdots, N)$ のように音源のパワーと雑音のパワー（分散）の和として与えられる．固有値$\{\lambda_m | m = N+1, N+2, \cdots, M\}$に対応する固有ベクトル$\{\boldsymbol{g}_m | m = N+1, N+2, \cdots, M\}$で張られる部分空間は雑音部分空間$\mathcal{V}$，固有値$\{\lambda_m | m = 1, 2, \cdots, N\}$に対応する固有ベクトル$\{\boldsymbol{g}_m | m = 1, 2, \cdots, N\}$で張られる部分空間は信号部分空間$\mathcal{S}$と呼ばれ，両者は直交補空間（$\mathcal{S}$=$\mathcal{V}^{\perp}$）をなす．そこで，$\boldsymbol{R}$を次のように固有値分解する．

$$\boldsymbol{R} = \sum_{m=1}^{N} \lambda_m \boldsymbol{g}_m \boldsymbol{g}_m^H + \sum_{m=N+1}^{M} \lambda_m \boldsymbol{g}_m \boldsymbol{g}_m^H \tag{2.17}$$

このとき，式(2.14)と式(2.17)のそれぞれに$\boldsymbol{g}_m \in \mathcal{V}$を左右からかけると，

$$\begin{aligned} \boldsymbol{g}_m^H \boldsymbol{R} \boldsymbol{g}_m &= \boldsymbol{g}_m^H \boldsymbol{A} \boldsymbol{S} \boldsymbol{A}^H \boldsymbol{g}_m + \sigma^2 \quad for \quad m = N+1, N+2, \cdots, M \\ \boldsymbol{g}_m^H \boldsymbol{R} \boldsymbol{g}_m &= \lambda_m = \sigma^2 \quad for \quad m = N+1, N+2, \cdots, M \end{aligned}$$

となって，$\boldsymbol{g}_m^H \boldsymbol{A} \boldsymbol{S} \boldsymbol{A}^H \boldsymbol{g}_m = 0$，すなわち，

$$\boldsymbol{a}^H(\theta_n) \boldsymbol{g}_m = 0 \quad for \quad \begin{cases} n = 1, 2, \cdots, N \\ m = N+1, N+2, \cdots, M \end{cases}$$

なる関係が導かれる．この関係式は，到来方向ベクトル$\boldsymbol{a}(\theta_n)$ $(n = 1, 2, \cdots, N)$は$\mathcal{V}$を張る固有ベクトル$\{\boldsymbol{g}_m | m = N+1, N+2, \cdots, M\}$と直交することを示している．したがって，このことと$\mathcal{S}$=$\mathcal{V}^{\perp}$なる事実より，$\{\boldsymbol{a}(\theta_n) | n = 1, 2, \cdots, N\}$は$\mathcal{S}$を張ることが分かる．

以上のことから，試行方位θに対するステアリング・ベクトルを式(2.10)と同様に

$$\boldsymbol{w}(\theta) = [1, e^{-j\omega d \sin\theta / c}, \cdots, e^{-j\omega(M-1) d \sin\theta / c}]^T$$

と生成し，

$$U(\theta) = \sum_{m=N+1}^{M} |\boldsymbol{w}^H(\theta) \boldsymbol{g}_m|^2 \tag{2.18}$$

のように定義される評価関数の値を求めることで，試行方位θが音源の真の到来方向θ_nと等しいか否かを判定できる．すなわち，$\theta=\theta_n$のとき$U(\theta)=0$，$\theta\neq\theta_n$のとき$U(\theta)>0$となることから，$U(\theta)$が極小値をとるときのθをθ_nの推定値とすることができる．ただし，極値探索を効果的に行う観点から，MUSIC法では，式(2.18)の代わりに

$$U_{MU}(\theta) = \frac{\|\boldsymbol{w}(\theta)\|^2}{\sum_{m=N+1}^{M}|\boldsymbol{w}^H(\theta)\boldsymbol{g}_m|^2} \tag{2.19}$$

と定義されるMUSICスペクトルを評価関数として用いるのが一般的である．この場合，$U_{MU}(\theta)$は，試行方位θが音源の真の到来方向θ_n $(n=1,2,\cdots,N)$と一致したとき，ピークをとる．したがって，$U_{MU}(\theta)$がピークとなるときのθをθ_n $(n=1,2,\cdots,N)$のいずれか1つの推定値として採択することになる．

したがって，複数の音源が存在するもとで，未知の音源到来方向θ_nをMUSIC法に基づいて推定する手順をまとめると，次のようになる．

i) マイクロホンアレイの観測値$\boldsymbol{x}(\omega)$をもとに相関行列$\boldsymbol{R}$を作成し，$\boldsymbol{R}$の固有値と固有ベクトルを求めて，その固有値を式(2.16)のように大きい順にソートする．

ii) ソートされた固有値の中で，値の小さな$(M-N)$個の固有値$\{\lambda_m|m=N+1,N+2,\cdots,M\}$に対応する固有ベクトル$\{\boldsymbol{g}_m|m=N+1,N+2,\cdots,M\}$を保存する．

iii) 試行方位θに対するステアリング・ベクトル$\boldsymbol{w}(\theta)$を作成して，式(2.19)の評価関数$U_{MU}(\theta)$の値を算出する．

iv) 試行方位θを–90°から90°まで適度な刻みで逐次更新しながら，iii) を繰り返す．

v) 以上で得られた$U_{MU}(\theta)$の系列値からN個のピークを探索して，ピークをとるときのθを音源の到来方向θ_n $(n=1,2,\cdots,N)$の1つの推定値として採択する．

前節のビームフォーム原理に基づくDSB法では，試行方位θに向け形成したビームを空間的に回転走査することで音源の到来方向（DOA）を推定している．一方，MUSIC法では，式(2.18)に基づいて形成される死角（Null）を空間的に回転走査することでDOAを推定している．この場合，ビームと死角の特性はマイクロホン数Mが増えるほど鋭くなるが，死角の方がビームに比べてより鋭く形成される．そのため，MUSIC法による場合，DSB法に比べて，空間分解能が高くなって，DOAの推定精度が高くなる．

しかし，固有値や固有ベクトルを求める必要があるため，DSB法より計算量は多く，その計算量はマイクロホンの個数とともに増える．また，実際の適用においては，音源以外の方向から相関の高い反射波が到来したり，残響による影響のため，前述のi)からiii)の仮定が崩れてしまう．そのため，Rの固有値を大きい順にソートしても，$\mathcal{S}$に関与する固有値$\{\lambda_m | m = 1, 2, \cdots, N\}$と$\mathcal{V}$に関与する固有値$\{\lambda_m | m = N+1, N+2, \cdots, M\}$の境目が不明瞭になる．したがって，MUSIC法の適用に際し，音源数Nは既知であることが，絶対的な前提条件となる．また，式(2.17)から式(2.18)までの議論から分かるように，音源数Nがマイクロホン数Mより少ないことも前提要件である．

2.3 DUET法の概要

ここでは，音源からの直接波だけでなく反射波も入力信号としてモデル化してDOAを推定する方法であるDUET法について述べる．この手法は，信号のスパース性を利用するため，その適用はスパース的な信号源に限られるが，マイクロホン数が信号源数より少ない場合でも上手く機能する．

未知の信号源$s_n^*(t)$ $(n = 1, 2, \cdots, N)$から到来する伝搬波をマイクロホンで観測した場合，各マイクでは

$$x_m(t) = \sum_{n=1}^{N} \sum_{t'=0}^{T'-1} h_{mn}(t') s_n^*(t - t') + v_m(t) \tag{2.20}$$

と信号源からの直接波に加えて反射波が畳み込まれて観測される．ここに，$x_m(t)$はm $(= 1, 2, \cdots M)$番目のマイクでの観測値，$h_{mn}(t')$はn番目の信号源からm番目のマイクまでのインパルス応答，t'は遅れ時間，T'はインパルス応答長，$v_m(t)$は雑音である．

式(2.20)の時間領域畳込みモデルは現実的な観測モデルである．しかし，応用目的により問題を解析的に取り扱い易くするため，何からの仮定を設けて簡略化したモデルに立脚して議論を展開することが多い．DUET法でも幾つかの仮定を設ける．以下ではこれらの仮定について説明する．

2.3.1 DUET法のデータモデルに関する仮定

DUET法では，無響室での観測を前提に，信号源からの直接波のみが$M = 2$個のマイクに到達すると仮定して，観測データを

$$x_1(t) = \sum_{n=1}^{N} s_n(t) \tag{2.21}$$

$$x_2(t) = \sum_{n=1}^{N} a_n s_n(t - \delta_n) \tag{2.22}$$

とモデル化する．ここに，$s_n(t)$は，n番目の信号源の直接波がマイク1に到達したときの値で，a_nとδ_nはマイク２でのマイク1に対する相対的な減衰係数と時間遅れである．また，$a_n \neq a_{n'}$か$\delta_n \neq \delta_{n'}$であれば，$s_n(t)$と$s_{n'}(t)$は異なる信号とする．

DUET法では，さらに，信号性質やマイク間隔について以下のような仮定を設けて，観測値$x_m(t)$ $(m = 1, 2)$のみを用いて信号$s_n(t)$ $(n = 1, 2, \cdots, N)$を分離する．

2.3.2 W-Disjoint Orthogonality

信号$s_n(t)$を窓関数$W(t)$で掛けした短時間フーリエ変換(STFT: short-time Fourier transform)を

$$F^W[s_n(t)](\tau, \omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} W(t - \tau) s_n(t) e^{-J\omega t} dt \tag{2.23}$$

と定義し，さらに便宜上，$s_n(\tau, \omega) = F^W[s_n](\tau, \omega)$と略記する．ここに，$J = \sqrt{-1}$である．このとき，

$$s_n(\tau, \omega) s_{n'}(\tau, \omega) = 0 \quad \forall(\tau, \omega) \quad \forall n \neq n' \tag{2.24}$$

ならば，信号$s_n(t)$と$s_{n'}(t)$はW-Disjoint Orthogonality（WDO）であると云われる．WDOは時間周波数領域での直交性を述べたもので，各時間周波数(τ, ω)で支配的な信号は高々１個しか存在しないことを意味している．このWDOをDUET法では仮定する．

2.3.3 局所定常性に関する仮定

式(2.23)は$W(t) = 1$のときフーリエ変換となって移動定理が成り立つ．$W(t)$の台がハミング窓等のように有限なSTFTの場合，移動定理は厳密には成り立たないが，シフト幅

が窓幅に比べて小さければ，

$$F^W[s_n(t-\delta_n)](\tau,\omega) \approx e^{-j\omega\delta_n} F^W[s_n(t)](\tau,\omega) \tag{2.25}$$

のように近似的に成り立つ．このように移動定理が近似的に成り立つ場合，信号は局所定常的であると云われ[36]，時間遅れδ_n[sec]と位相遅れ$\omega\delta_n$[rad]が等価的に扱える．この局所定常性は，狭帯域信号に対して置かれる狭帯域仮定（Narrowband Assumption）を広帯域信号に対して拡張した概念と位置づけられる．中心周波数ω_cが既知で振幅$\eta(t)$や位相$\phi(t)$が未知の信号$s(t)=\eta(t)\cos\{\omega_c t+\phi(t)\}$を狭帯域信号と云う．その時間遅れは$s(t-\delta)$=$\eta(t-\delta_n)\cos\{\omega_c(t-\delta_n)+\phi(t-\delta_n)\}$と表現されるが，振幅や位相の変化が$\eta(t-\delta)\approx\eta(t)$や$\phi(t-\delta)\approx\phi(t)$と緩やかであれば（狭帯域仮定），$s(t-\delta)$=$\eta(t)$ $\Re\{e^{j(\omega_c(t-\delta)+\phi(t))}\}$= $\Re\{e^{-j\omega_c\delta}\tilde{s}(t)\}$のように時間遅れと位相遅れを等価的に扱える．ここに，$\Re$は実数部，$\tilde{s}(t)=\Re\{\eta(t)e^{j(\omega_c t+\phi(t))}\}$ である．

2.3.4 センサ間隔に関する仮定

式(2.25)の複素指数関数は$e^{-j\omega\delta_n}$=$e^{-j(\omega\delta_n+2\pi)}$ と同じ値を周期的にとる多義的な関数である．このように多義的な場合，位相ラップが生じて，結果的に空間的エイリアシングが起きる．これを回避するには，位相遅れ$\omega\delta_n$[rad]の範囲を絞り込む必要がある．位相遅れは，マイク対正面の左からの到来波に対して正，右からの到来波に対して負の値をとる．そこで，$|\omega\delta_n|<\pi$と絞り込むことにする．

時間遅れδ_n[sec]については，マイク間隔をd[m]，伝播速度をc[m]とすると，マイク対で生じる最大の時間遅れは$\delta_{max}=d/c$[sec]となる．一方，周波数ω[rad/sec]については，信号の最大周波数をω_{max}とすると，サンプリング周波数ω_sは$2\omega_{max}$以上に定める必要がある．そこで，ω_s=$2\omega_{max}$と置く．この場合，$|\omega\delta_n|<\pi$を満たすには，マイク間隔を$d<c/\omega_s$とする必要がある．これが空間的エイリアシングを起こさないために課される制約である．

2.4 DUET法による信号分離原理とDOA推定

マイク間隔dは制約の範囲内にあり，信号$s_n(t)$は局所定常的であるとして，式(2.21)(2.22)をSTFTすると，

$$\begin{bmatrix} x_1(\tau,\omega) \\ x_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ a_1 e^{-j\omega\delta_1} & \cdots & a_N e^{-j\omega\delta_N} \end{bmatrix} \begin{bmatrix} s_1(\tau,\omega) \\ \vdots \\ s_N(\tau,\omega) \end{bmatrix} \tag{2.26}$$

のような時間周波数表現が得られる．ここに，$x_m(\tau,\omega)$は$x_m(t)$をSTFTして得られる複素スペクトルである．

2.4.1 DUET法による信号分離原理

ここで，さらにWDOを仮定，すなわち，各時間周波数(τ,ω)では高々１つの信号$s_n(\tau,\omega)$のみが支配的で他の信号と重ならないと仮定すると，式(2.26)は

$$\begin{bmatrix} x_1(\tau,\omega) \\ x_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 \\ a_n e^{-j\omega\delta_n} \end{bmatrix} s_n(\tau,\omega) \quad \exists n \tag{2.27}$$

と変形される．この場合，$x_1(\tau,\omega)$と$x_2(\tau,\omega)$の比は，

$$x_2(\tau,\omega)/x_1(\tau,\omega) = a_n e^{-j\omega\delta_n} \tag{2.28}$$

となって減衰比a_nと時間遅れδ_nのみに依存して決まる．それゆえ，a_nとδ_nは，各時間周波数(τ,ω)で

$$\tilde{a}_n(\tau,\omega) = |x_2(\tau,\omega)/x_1(\tau,\omega)| \tag{2.29}$$

$$\tilde{\delta}_n(\tau,\omega) = (-1/\omega)\angle(x_2(\tau,\omega)/x_1(\tau,\omega)) \tag{2.30}$$

と未知の信号$s_n(\tau,\omega)$に依存することなく推定できる．ここに，$\angle$は偏角を表す．以降では，便宜のため，$(\tilde{a}_n(\tau,\omega),\tilde{\delta}_n(\tau,\omega))$を局所推定値と呼ぶことにする．

局所推定値は，真値(a_n,δ_n)の回りに集まって分布し，信号数に等しいN個のクラスタ$\{C_n | n = 1,2,\cdots,N\}$に分類される．それゆえ，マスク関数を

$$\mathcal{M}_n(\tau,\omega) = \begin{cases} 1 & (\tilde{a}_n(\tau,\omega),\tilde{\delta}_n(\tau,\omega)) \in C_n \\ 0 & otherwise \end{cases} \tag{2.31}$$

と作成すれば，クラスタC_nの信号$s_n(\tau,\omega)$が

$$\hat{s}_n(\tau,\omega) = \mathcal{M}_n(\tau,\omega)x_1(\tau,\omega) \quad \forall(\tau,\omega) \tag{2.32}$$

と分離できる．以上がDUET法の信号分離原理である．

2.4.2 DUET法に基づくDOA推定

上述のことから，局所推定値の2次元ヒストグラムを作成した場合，そのヒストグラムは信号源数と等しい個数のピークをもち，そのピークは真値(a_n,δ_n)にほぼ等しい点で得られることになる．そこでRickardらは，局所推定値$\{(\tilde{a}_n(\tau,\omega),\tilde{\delta}_n(\tau,\omega))\}$を$N$個のクラスタにクラスタリングして，時間遅れ$\delta_n$を

$$\hat{\delta}_n \;=\; \frac{1}{|C_n|}\sum_{(\tau,\omega)\in C_n}\tilde{\delta}_n(\tau,\omega) \tag{2.33}$$

と求めて，信号$s_n(t)$のDOAを

$$\hat{\theta}_n(\tau,\omega) \;=\; \sin^{-1}\left\{\frac{c}{2\pi d}\hat{\delta}_n(\tau,\omega)\right\} \tag{2.34}$$

と推定することを提案している[31]．ここに，$|C_n|$はクラスタC_nに含まれる(τ,ω)の個数である．

そして，一定の時間間隔で$\beta = 0,1,\cdots,59$の60値のいずれかを独立に採る10系列の広帯域FSK（Frequency Shift Keying）信号

$$s_n(t) = \sin 2\pi(10^9 + 16\times 10^3\beta)\,t \quad n = 1,\cdots,10 \tag{2.35}$$

を対象に，これらのDOAを$-60°$から$60°$の範囲でランダムに変えてシミュレーションを行い，提案法の妥当性を検証している．具体的には，式(2.21)(2.22)で与えられる$x_m(t)$に雑音$v_m(t)$を付加して，$x_m(t) \leftarrow x_m(t) + v_m(t)$ $(m = 1,2)$と生成される値を２つのマイクでの観測値としてDOA推定を行い，最大推定誤差が$0.5°$未満となる割合がSNR=15[dB]のとき97.3%，SNR=20[dB]のとき99.4%となることを示して，MUSIC等の既存法に対する優位性を述べている．

しかし，クラスタリング等のバッチ処理を必要とする方法では，突発音などの瞬間的な音や，移動音源のDOAを推定することは難しい．突発音などの瞬間的な音や，移動

音源への適用を行うには，データを溜め込むことなく，フレーム単位にDOAを推定することが必須である．

2.5 短時間離散フーリエ変換と局所DOA

現実的な環境でDOA推定を行うには，式(2.20)のように観測される$x_m(t)$ $(m = 1, 2)$を連続時間形式から離散時間形式に直して短時間離散フーリエ変換(STDFT)する必要がある．すなわち，サンプル周期をT_s[sec]として得られる観測データ$\{x_m[j] = x_m(jT_s) \,|\, j = 0, 1, \cdots, J-1\}$を$L$個ずつ切出した$k$番目のフレームデータ$\{x_m[l + kF] \,|\, l = 0, 1, \cdots, L-1\}$を

$$x_m[k, \omega_l] \;=\; \sum_{l=0}^{L-1} x_m[l + kF]\, W[l]\, e^{-j2\pi kl/L} \tag{2.36}$$

とSTDFTして，時間周波数$[k, \omega_l]$ $(k = 0, 1, \cdots, K-1;\ l = 0, 1, \cdots, L-1)$での複素スペクトル$x_m[k, \omega_l]$を求めておく必要がある．ここに，$\omega_l = 2\pi l/L$で，$F$はフレーム周期，$W[l]$は窓関数である．

以降では，式(2.33)の時間遅れ$\tilde{\delta}_n$の替わりに，位相遅れ$\tilde{\varphi}_n = \omega_l \tilde{\delta}_n$を各時間周波数$[k, \omega_l]$で

$$\tilde{\varphi}[k, \omega_l] \;=\; -\angle(x_2[k, \omega_l]\bar{x}_1[k, \omega_l]) \tag{2.37}$$

と求めて，次式のように算出される$\tilde{\theta}[k, \omega_l]$を局所DOAと呼ぶことにする．ここに，$\bar{}$は複素共役を表す.

$$\tilde{\theta}[k, \omega_l] \;=\; \sin^{-1}\left\{\frac{c}{2\pi d}\tilde{\varphi}_n[k, \omega_l]\right\} \tag{2.38}$$

2.6 まとめ

本章では，音源のDOA推定について概説した．具体的には，DSB法，MUSIC法，スパース性に基づいたDUET法の基本原理を説明し，各方法の長所や短所を明らかにして，実際への適用に関しての課題を述べた．

DSB法は，試行方位θ(-90° ~ 90°)に対するステアリング・ベクトルを生成し，出力されたアレイ入力が最大となるθを探索することで音源のDOAを推定するという原理になって

いることを説明した．しかし，この場合，DOAは比較的簡単に推定できるが，複数の音源が同時にマイクロホンに入射するときは機能せず，空間分解能が低いことを述べた．

MUSIC法は，信号部分空間を張る到来方向ベクトルと，雑音部分空間を張る固有ベクトルが直交する性質を利用して，MUSICスペクトルのピークから音源のDOAを推定する方法であることを説明した．この方法は，複数の音源が同時にマイクロホンに入射する環境下でも，音源の個数よりマイクロホンの個数が多ければ，個々の音源のDOAは推定できて，空間分解能も高いことを述べた．しかし，固有値や固有ベクトルを求める必要があるため，DSB法より計算量は多く，その計算量はマイクロホンの個数とともに増えることや，信号数が既知でなければ適用できないことを述べた．

音源のスパース性に基づいたDUET法は，クラスタリングにより局所DOAのヒストグラムを作成し，そのピークから個々の音源のDOAを推定する方法であることを説明した．この方法は，マイクロホンの個数は音源の個数より少なくてもDOAが推定できるという特色があり，無響室下でクラスタリングが良好に行えれば，信号源数が未知の場合でも有効に機能することを述べた．しかし，クラスタリング等のバッチ処理を必要とするため，突発音などの瞬間的な音や，移動音源のDOAを推定することは難しいことを指摘した．

以上のことから，突発音などの瞬間的な音や，移動音源への適用を行うには，フレーム単位でのDOA推定が必須であることを明らかにした．

第3章

音声のDOA推定に関する予備的検討

前節でDOA推定対象となった広帯域FSK信号は，周波数が1000[MHz]から1009.44[MHz]に渡る振幅が一定の連続波である．また，MUSIC法は，狭帯域の連続波のDOA推定を目的とした手法である．一方，音声は100〜4000[Hz]の広帯域信号で，音声は音声区間と無音区間を繰り返す断続波である．この場合，音声のDOAに関する情報は，音声区間には含まれるものの，無音区間には含まれない．しかし，現実の環境では，無音区間に無関係な暗騒音や残響が入り込んでしまい，そのことが音声のDOA推定に悪影響を及ぼす．これは，フレーム数が少ない場合，特に問題となる．すなわち，該当するフレームがすべて無音区間内にある場合，これらのフレームにおける局所DOAは求めるべき音声のDOAとは無関係な値をとることになる．

そこで本章では，音声区間内の連続する20フレームでの局所DOAをもとに目的音声のDOAを推定し，どの程度の推定精度が得られるかを，シミュレーションにより検討する．

3.1 シミュレーション環境

シミュレーションは，１つの音源$s_1^*(t)$からの到来波を２つのマイクロホンで観測する場合を考え，新聞記事読み上げ音声コーパス[37]から選んだ6.5秒程度のソース音声$s_1^*(t)$（男女各3発話の6パターン）をもとに式(2.20)のように観測される$x_m(t)$をマイクロホン収録音声として行った．すなわち，Habets[38] による室内インパルス応答生成モデル[39]に基づいて，残響時間RT_{60}が200[msec]のインパルス応答を生成し，その応答にソー

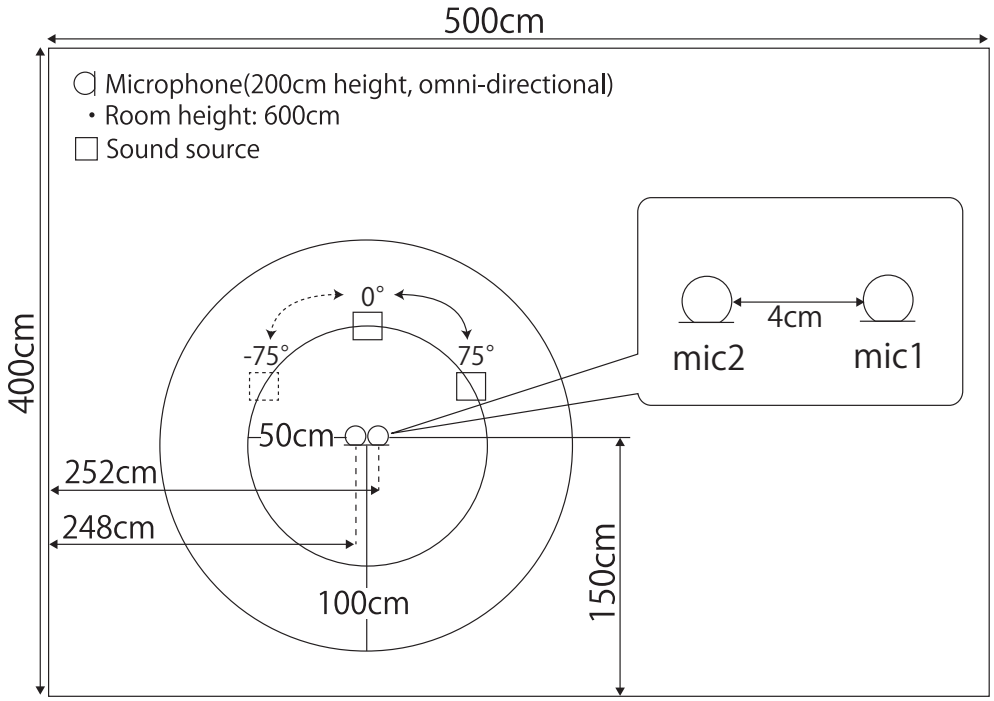


図 3.1 シミュレーション環境(単一音源)

ス音声を畳み込んだ信号に正規性白色雑音を付加したものを収録音声とした．収録音声は，サンプリング周波数を8000[Hz]，分解能を16[bit]とし，分析フレーム長を64[msec]，フレームシフト幅を16[msec]，窓関数をHamming窓として，短時間離散フーリエ変換した．また，２つのマイクロホンの間隔dを空間的エイリアシングが起きないように4[cm]として，その中心から音源を50[cm]離し，方位θを0°（=ブロードサイド方向=マイクロホン対の正面方向）から75°まで15°ずつ変えて配置した．

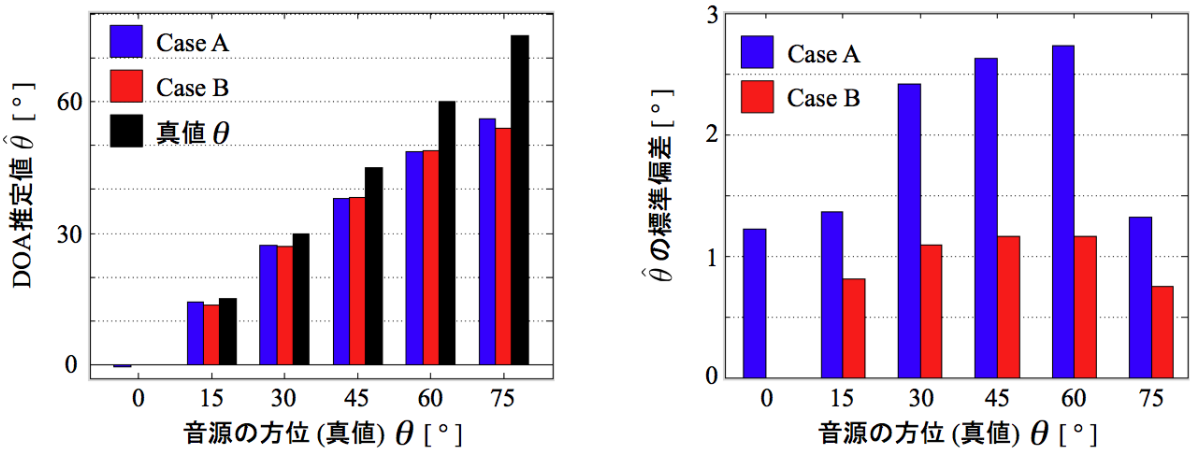
3.2 DUET法に基づく単一音源のDOA推定

前述のシミュレーション環境下でSNRを20[dB]として，音声区間で連続する20フレーム（368[msec]）を用いて単一音源のDOA（θ）を推定した．また，比較のため，発話の全フレーム（約6.5秒の概ね400 フレーム）を用いた場合についても，同様にDOAを推定した． 頻度分布は，式(2.38)の局所DOA $\tilde{\theta}[k,\omega_l]$をもとに次のように作成した．

すなわち，$-90.5° \sim 90.5°$をθ_q°=$[q° - 91.5°, q° - 90.5°)$ $(q = 1, 2, \cdots, 181)$のように1度単位に分け，局所DOAがθ_q°に入る頻度をp_qとして，頻度分布を$P(\tilde{\theta})$=$[p_1, p_2, \cdots, p_{181}]$と作成

表 3.1 DUET法に基づくDOA推定値

Case	$\theta = 0°$	$\theta = 15°$	$\theta = 30°$	$\theta = 45°$	$\theta = 60°$	$\theta = 75°$
A	-0.50° (1.22°)	14.33° (1.36°)	27.33° (2.42°)	37.83° (2.63°)	48.50° (2.73°)	56.16° (1.32°)
B	0.00° (0.00°)	13.66° (0.81°)	27.00° (1.09°)	38.16° (1.16°)	48.83° (1.16°)	53.83° (0.75°)

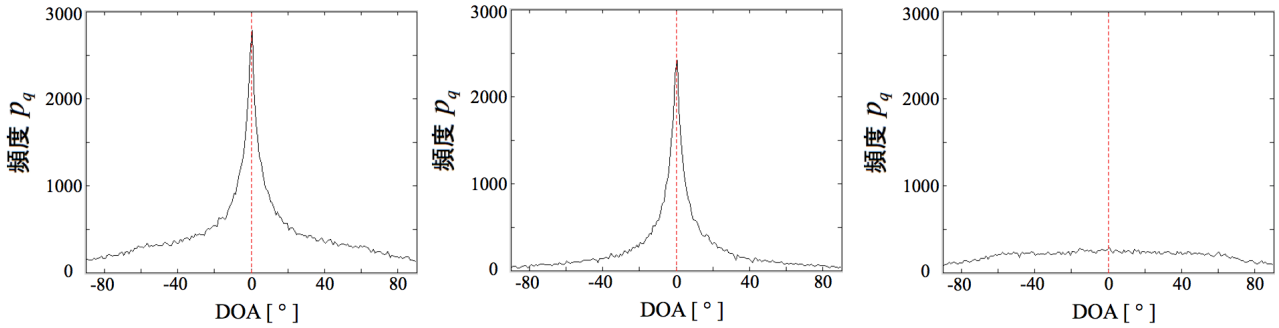


(a) DOA推定値の平均　　(b) DOA推定値の標準偏差

図 3.2 DUET法に基づくDOA推定値の棒グラフ

した．また，その頻度分布のピークを探索して，ピークを採るときの方位（$\hat{\theta}$）をDOAの推定値とした．具体的には，頻度分布$P(\tilde{\theta})$で頻度p_qが最大となるときの角度番号qを$\hat{q}$として，$\hat{\theta} = (\hat{q} - 91)°$と算出される$\hat{\theta}$を音源のDOA推定値とした．

このときのDOA推定値の平均と標準偏差（（）の数値）を表3.1に示す．表中，Case Aは音声区間の連続する20フレームを用いたときの結果で，Case Bは全フレームを用いたときの結果である．また，図3.2(a)にCase AとCase BのDOA推定値の平均，図3.2(b)にその標準偏差を棒グラフで示す．青の棒グラフがCase A，赤の棒グラフがCase B，黒の棒グラフが音源の方位θである．Case Aの場合，推定誤差はθが0°から外れるにつれて大きくなり，30°までは3°未満に収まるが，θ=45°を越えると7°以上となって推定精度は急激に劣化する．同様のことが，全フレームを用いたCase Bのときの推定誤差についても云える．また，標準偏差はCase Bの方がCase Aに比べて小さいが，その差は概ね1°程度で

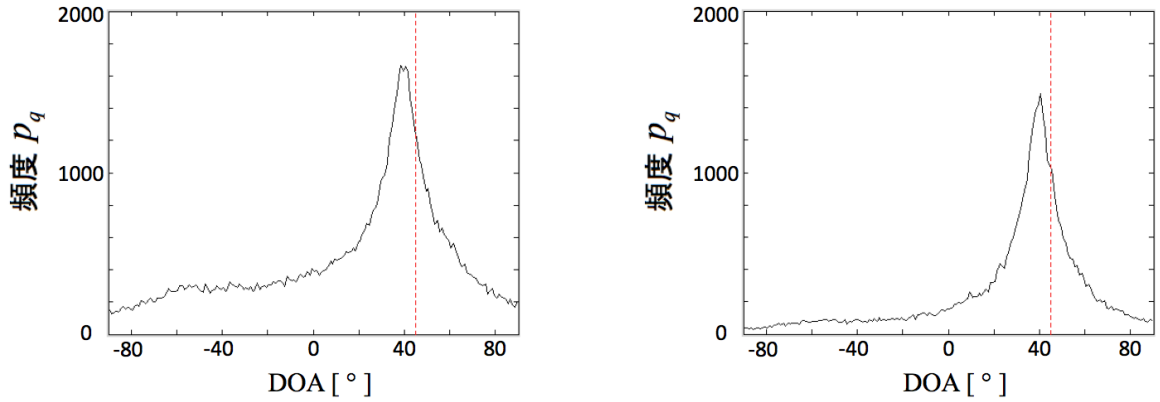


(a) all frames (b) voiced frames (c) ambient noise frames

図 3.3 $\theta = 0°$ のときの局所DOA頻度分布

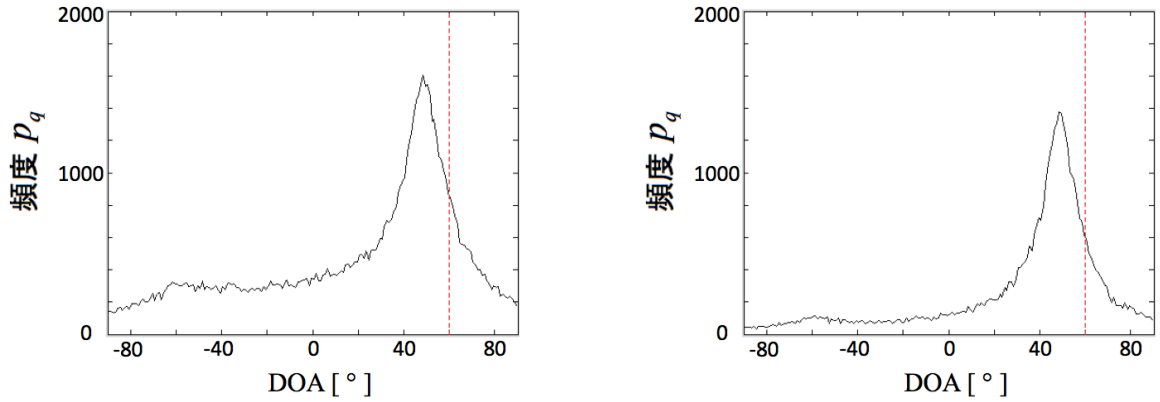
ある．つまり，推定値のバラツキは，全フレームを用いた場合と音声区間の20フレームを用いた場合とで，大きく変わらない．

上述でθ=45°を越えると推定精度が急激に劣化する原因を探るため，各発話を音声区間と無音区間に分けて，それぞれの区間における局所DOAの頻度分布を調べた．まず，6.60秒の女性発話（音声区間計3.48秒，無音区間計3.12秒）を$\theta = 0°$の方向から流したときに得られる局所DOAの頻度分布を図3.3に示す．図の縦軸は頻度p_q，横軸はDOA[°]，赤の破線はDOAの真値($\theta = 0°$)である．音声区間は，ソース音声を対象に，各サンプル時刻でフレーム（＝150サンプル）毎に平均パワーを求め，これらをその最大値で割って0～1の範囲になるように規格化し，その値が0.01を超えたところを音声の始端，0.0005以下になるところを音声の終端とした．図中，(a)は全フレームでの局所DOAの分布，(b)は音源区間フレームでの局所DOAの分布，(c)は無音区間フレームでの局所DOAの分布である．これらの図から，(a)と(b)の分布は0 °近傍でピークを採り，DOAの真値$\theta = 0°$とほぼ一致することがみてとれる．一方，(c)は比較的平坦な分布となって際だったピークがなく，DOA推定に関する手がかりもない．さらに，(a)と(b)の分布について，平均（標準偏差）を求めてみた．その結果，分布(a)で0.07°（37.43°），分布(b)で−0.13°（28.61°）となって，いずれの場合も平均はDOAの真値0°にほぼ一致することが分かる．



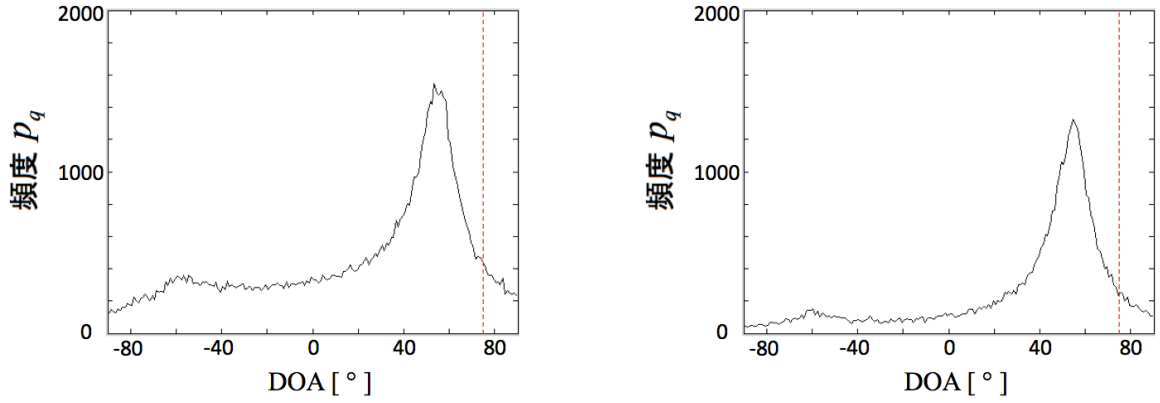
(a) all frames　　(b) voiced frames

図 3.4 $\theta = 45°$ のときの局所DOA頻度分布



(a) all frames　　(b) voiced frames

図 3.5 $\theta = 60°$ のときの局所DOA頻度分布



(a) all frames　　(b) voiced frames

図 3.6 $\theta = 75°$ のときの局所DOA頻度分布

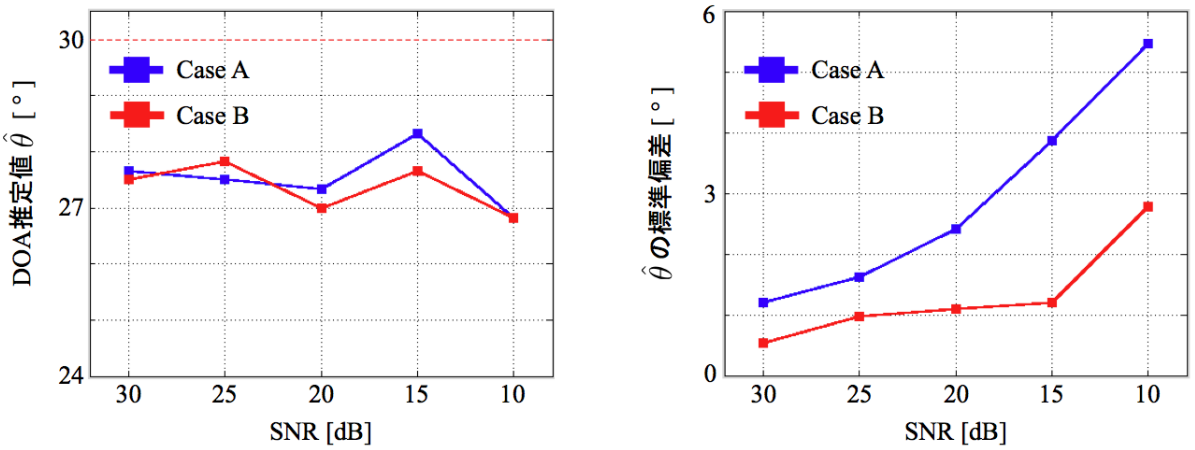
次に，同一発話を$\theta = 45°$の方向から流したときの局所DOAの頻度分布を図3.4に，$\theta = 60°$の方向から流したときの局所DOAの頻度分布を図3.5に，$\theta = 75°$の方向から流したときの局所DOAの頻度分布を図3.6に示す．各図の(a)は全フレームでの分布，(b)は音源区間フレームでの分布である．まず，図3.4の場合，(a)と(b)の分布のピークは40 °付近にあることが見てとれる．しかし，この値は，真値45°から約5°外れており，表3.1で$\theta = 45°$のとき推定誤差が約7°となることと符合する．さらに，平均（標準偏差）を計算してみたところ，その値は分布(a)で16.84°（42.29°）となって，真値45°から約30°と大きく外れる結果となった．また，分布(b)の平均（標準偏差）は27.68°（33.98°）となって，真値45°とは約17°乖離する結果となった．次に，図3.5の場合，(a)と(b)の分布のピークは50 °付近にあることが見てとれる．この値は，真値60°から約10°外れており，表3.1で$\theta = 60°$のとき推定誤差が約12°となることと符合する．このときの平均（標準偏差）は分布(a)で19.53°（44.67°）となって，ここでも真値60°から約40°と大きく外れる結果となった．また，分布(b)の平均（標準偏差）は32.51°（37.14°）となって，真値60°とは約27°乖離する結果となった．最後に，図3.6の場合，(a)と(b)の分布のピークは55 °付近にあることが見てとれる．この値は，真値75°から約20°外れており，表3.1で$\theta = 75°$のとき推定誤差が約20°となることと符合する．このときの平均（標準偏差）は分布(a)で20.43°（46.34°）となって，ここでも真値75°から約55°と大きく外れる結果となった．また，分布(b)の平均（標準偏差）は34.26°（39.79°）となって，真値75°とは約40°乖離する結果となった．

このような真値との乖離は，局所DOAが音源区間内のフレームから得られたものであっても，その分布は図3.4(b)，図3.5(b)，図3.6(b)のようにピーク左側の頻度がより高く，左の裾が重い分布となることに起因している．$\theta > 0$のときの局所DOAの分布は左の裾が重く，逆に$\theta < 0$のときの分布は右の裾が重くなる．また，分布のピーク$\tilde{\theta}_{peak}$はθより内側，つまり$(\theta - \tilde{\theta}_{peak}) \gtrless 0$ $(\theta \gtrless 0)$に位置し，ピークと真値の差$|\theta - \tilde{\theta}_{peak}|$は$|\theta|$が大きくなる程拡がる傾向にあった．

局所DOA頻度分布のピークが，$\theta = 0°$のときを除いて，真値から乖離するのは，調波構造をもつ音声の場合，音声区間内の時間周波数$[k, \omega_l]$のすべてが必ずしも音声の直接波に由来するものではないから，と考えられる．言い換えると，式(2.33)のように，$[k, \omega_l]$を音源からの直接波に由来する時間周波数のみを$\{[k, \omega_l] \in C_n\}$と正確にクラスタリングできれば，局所DOAの分布は歪むことなく左右対称となって真値に近い推定値が得られるは

表 3.2 SNRに対するDUET法に基づくDOA推定値(θ=30°，RT_{60}=200[msec])

Case	SNR=30[dB]	SNR=25[dB]	SNR=20[dB]	SNR=15[dB]	SNR=10[dB]
A	27.66° (1.21°)	27.50° (1.63°)	27.33° (2.42°)	28.33° (3.86°)	26.83° (5.46°)
B	27.50° (0.54°)	27.83° (0.98°)	27.00° (1.09°)	27.66° (1.21°)	26.83° (2.78°)



(a) DOA推定値　　(b) DOA推定値の標準偏差

図 3.7 SNRに対するDUET法に基づくDOA推定値の折れ線グラフ(θ=30°，RT_{60}=200[msec])

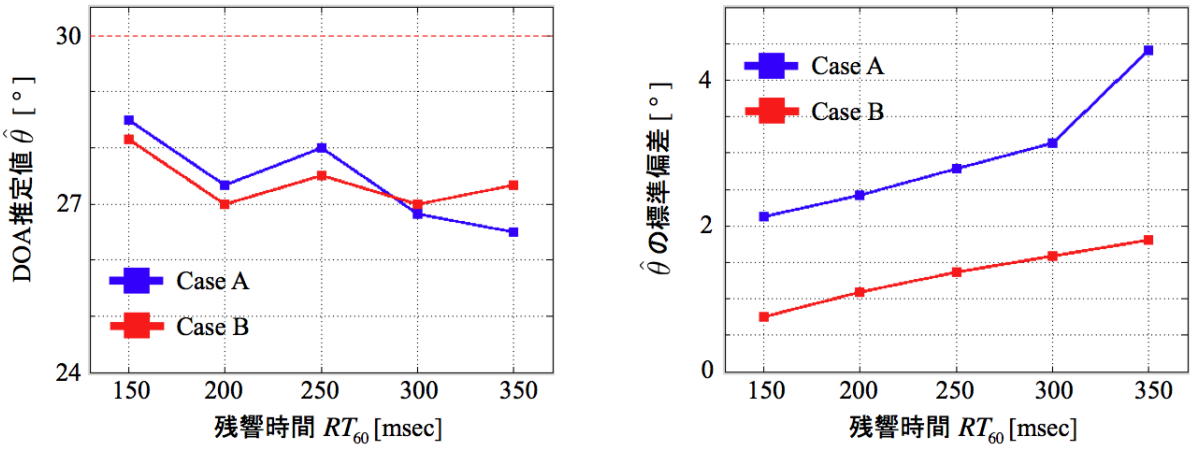
ずである．しかし，残響のある環境下でクラスタリングを正確に行うのは容易でない．

次に，DUET法に基づく方法によりDOA推定を行う場合，暗騒音や残響がどの程度の影響を及ぼすか調べるため，SNRと残響時間RT_{60}に対するDOA推定精度をシミュレーションによりそれぞれ検証した．ここでは，音源の方位を$\theta = 30°$とした．まず，SNRがDUET法に基づくDOA推定に及ぼす影響を調べるため，残響時間RT_{60}を200[msec]として，SNRを30[dB]から10[dB]まで5[dB]ずつ変えてシミュレーションを行った．このときのDOA推定値の平均と標準偏差（()の数値）を表3.2に示す．表中，Case Aは音声区間の連続する20フレームを用いたときの結果で，Case Bは全フレームを用いたときの結果である．また，図3.7(a)にCase AとCase BのDOA推定値の平均，図3.7(b)に標準偏差を折れ線グラフで示す．図中，縦軸はDOA推定値$\hat{\theta}$とその標準偏差，横軸はSNR[dB]，青の折れ線グラフはCase A，赤の折れ線グラフはCase B，赤の破線は音源方位の真値(θ =

30°)である．表3.2と図3.7(a)から，SNRに対するDOA推定誤差はCase AでもCase Bでも，概ね変わらないことが見てとれる．具体的には，推定誤差はSNRに関わらず全体的に3°程度に収まっており，両者の差も1°未満であることがわかる．一方，標準偏差についてはCase AとCase Bは共に，SNRが低いほど大きくなることが表3.2と図3.7(b)から見てとれる．また，Case Aの標準偏差はCase Bに比べ全てのSNRで大きいことがわかる．具体的には，Case Bの標準偏差は，SNRが30[dB]のとき0.54°，SNRが10[dB]のとき2.78°となって約2°増加するが，全体的に3°未満に収まっている．そして，Case Aの標準偏差は，SNRが30[dB]のとき1.21°，SNRが10[dB]のとき5.46°となって約4°増加する．しかし，両者の差は最大でも3°未満に収まることから，DUET法に基づくDOA推定の場合，使用するフレームが少なくてもSNRの影響は少ないといえる．

表 3.3 RT_{60} に対するDUET法に基づくDOA推定値(θ=30°，SNR=20[dB])

Case	RT_{60}=150[msec]	RT_{60}=200[msec]	RT_{60}=250[msec]	RT_{60}=300[msec]	RT_{60}=350[msec]
A	28.50° (2.12°)	27.33° (2.42°)	28.00° (2.78°)	26.83° (3.14°)	26.50° (4.41°)
B	28.16° (0.75°)	27.00° (1.09°)	27.50° (1.37°)	27.00° (1.59°)	27.33° (1.80°)



(a) DOA推定値　　(b) DOA推定値の標準偏差

図 3.8 RT_{60} に対するDUET法に基づくDOA推定値の折れ線グラフ(θ=30°，SNR=20[dB])

次に，残響時間RT_{60}がDUET法に基づくDOA推定に及ぼす影響を調べるため，SNRを20[dB]として，RT_{60}を150[msec]から350[msec]まで50[msec]ずつ変えてシミュレーションを行った．このときのDOA推定値の平均と標準偏差（()の数値）を表3.3に示す．また，図3.8(a)にCase AとCase BのDOA推定値の平均，図3.8(b)に標準偏差を折れ線グラフで示す．表3.3と図3.8(a)から，RT_{60}に対するDOA推定誤差はCase AでもCase Bでも，概ね変わらないことが見てとれる．具体的には，推定誤差は全てのRT_{60}に対して2 ~ 3°程度に収まっており，両者の差も1°未満であることがわかる．一方，標準偏差に関してはCase AとCase Bの両者ともに，RT_{60}に比例して大きくなることが表3.3と図3.8(b)から見てとれる．また，Case Aの標準偏差はCase Bに比べ全てのRT_{60}で大きいことがわかる．具体的には，Case Bの標準偏差は，RT_{60}が150[msec]のとき0.75°，RT_{60}が350[msec]のとき1.80°となって約1°増加するが，全体的に2°未満に収まっている．そして，Case Aの標準偏差は，RT_{60}が150[msec]のとき2.12°，RT_{60}が350[msec]のとき4.41°となって約2°増加する．しかし，両者の差は最大でも3°未満に収まっていることから，DUET法に基づくDOA推定の場合，使用するフレームが少なくてもRT_{60}の影響は少ないといえる．

以上のことから，DUET法に基づくDOA推定の場合，SNRとRT_{60}に関わらず，少数フレームでも全フレームを用いたときと同等の精度が得られることが確認できた．

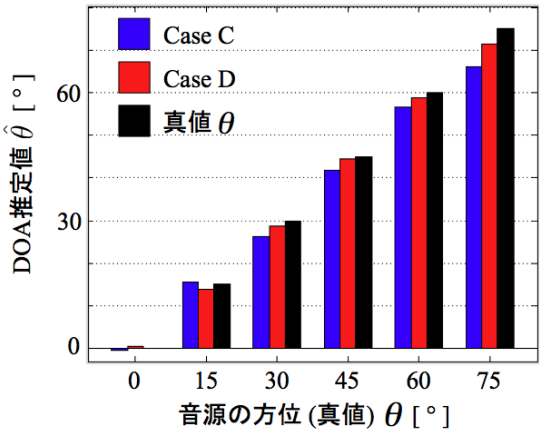
3.3 MUSIC法による単一音源のDOA推定

狭帯域信号の代表的なDOA推定法であるMUSIC法を各周波数ビンω_l ($l = 0, 1, \cdots, 255$)に適用してDOAを求めたときの平均と標準偏差を表3.4に示す．表中，Case Cは音声区間の連続する20フレームを用いたときの結果，Case Dは全フレームを用いたときの結果である．また，図3.9(a)にCase CとCase DのDOA推定値の平均，図3.9(b)に標準偏差をそれぞれ棒グラフで示す．青の棒グラフがCase C，赤の棒グラフがCase D，黒の棒グラフが音源の方位θである．部屋の大きさや壁・床等の吸音率に関係して共振周波数が変わることから，残響の影響は周波数によって異なるが，一般に低周波域でより強く現れる[38]．そのため，MUSIC法では低周波数域を外して推定値を求めることが多い[40]．

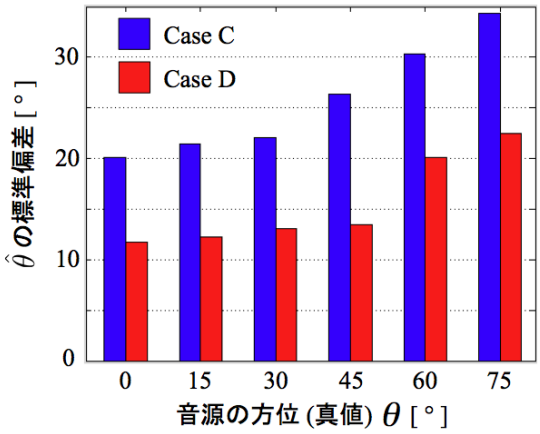
そこで，Case CとCase Dにおいて500Hz未満を除外した周波数ビンω_l ($l = 32, 33, \cdots, 255$)での局所DOAから得られた結果をそれぞれCase C†とCase D†として提示する．また，図3.9(c)にCase C†とCase D†のDOA推定値の平均，図3.9(d)に標準偏差を棒グラフで示す．青の棒グラフがCase C†，赤の棒グラフがCase D†，黒の棒グラフが音源の方位θである．まず，Case D†の場合，推定誤差はθに比例して大きくなることが読み取れる．この比例関係はCase C†，Case C，Case Dでも概ね成り立っている．しかし，DUET法に基づくDOA推定法と比較すると，推定誤差は概ね小さいことがわかる．具体的には，Case D†の場合，推定誤差は$\theta = 0°$から$60°$の範囲で$1°$未満となり，$\theta = 75°$のときでも$2°$未満に収まっている．同様に，Case Dの場合，推定誤差は$\theta = 0°$から$60°$の範囲で$2°$未満となり，$\theta = 75°$のときでも$4°$未満に収まっている．また，音声区間20フレームを用いたCase C†の場合でも，推定誤差は$\theta = 30°$と$\theta = 75°$のときを除いて$2°$未満に収まり，$\theta = 30°$のとき約$3°$，$\theta = 75°$のとき約$5°$である．最後に，音声区間20フレームを用いたCase Cの場合，推定誤差は$\theta = 0°$と$\theta = 15°$のとき$1°$未満，$\theta = 30°$と$\theta = 45°$のとき$4°$未満，$\theta = 75°$のとき約$9°$である．これらのことから，500Hz未満を除外することなく全周波数ビンでの局所DOAをもとに推定したCase DやCase Cの場合，推定誤差はCase D†やCase C†に比べて大きくなることがわかる．さらに，音声区間20フレームを用いたCase C†やCase Cの場合，推定誤差はCase D†やCase Dに比べて概ね大きくなることがわかる．

表 3.4 MUSIC法によるDOA推定値

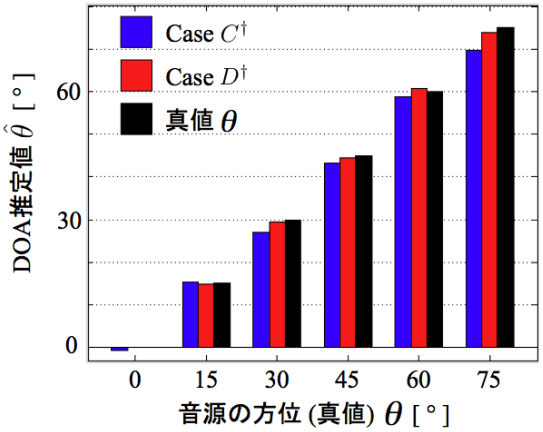
Case	$\theta = 0°$	$\theta = 15°$	$\theta = 30°$	$\theta = 45°$	$\theta = 60°$	$\theta = 75°$
C	-0.51° (20.11°)	15.65° (21.40°)	26.32° (22.06°)	41.85° (26.37°)	56.60° (30.33°)	65.98° (34.32°)
D	0.45° (11.76°)	13.84° (12.22°)	28.62° (13.06°)	44.34° (13.44°)	58.69° (20.07°)	71.46° (22.47°)
$C^{\dagger}$	-0.66° (11.74°)	15.30° (13.14°)	26.89° (14.70°)	43.32° (18.96°)	58.67° (21.00°)	69.80° (22.16°)
$D^{\dagger}$	0.03° (1.11°)	14.76° (3.45°)	29.51° (4.56°)	44.49° (5.86°)	60.73° (10.61°)	73.77° (12.31°)



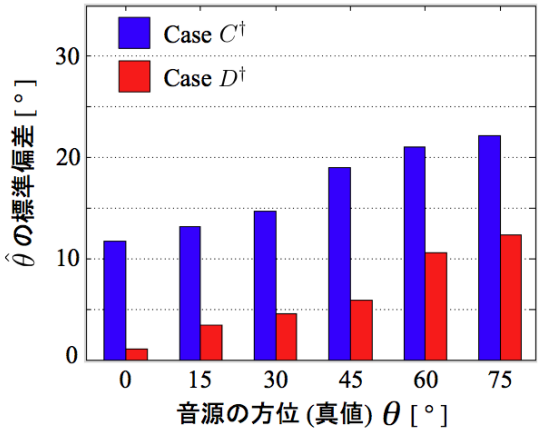
(a) DOA推定値の平均(CとD)



(b) DOA推定値の標準偏差(CとD)

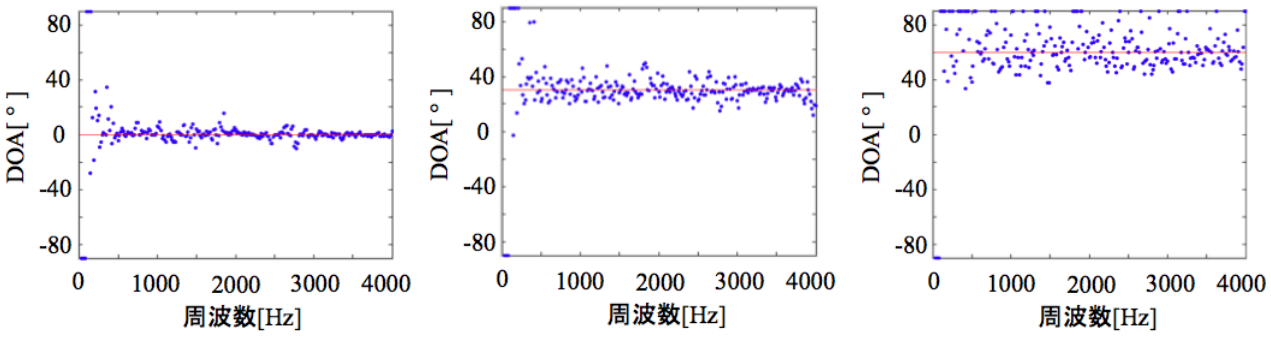


(c) DOA推定値の平均($C^{\dagger}$と$D^{\dagger}$)



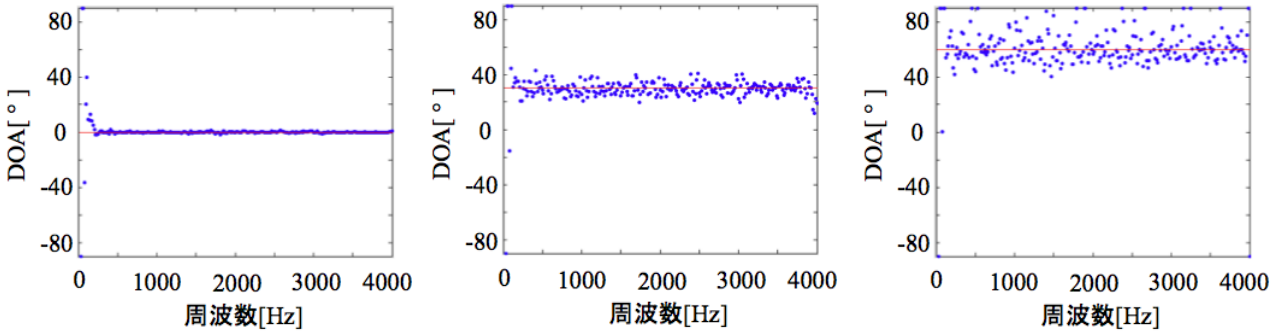
(d) DOA推定値の標準偏差($C^{\dagger}$と$D^{\dagger}$)

図 3.9 MUSIC法によるDOA推定値の棒グラフ



(a) $\theta = 0°$　(b) $\theta = 30°$　(c) $\theta = 60°$

図 3.10 MUSIC法による周波数毎のDOA推定結果(Case C)



(a) $\theta = 0°$　(b) $\theta = 30°$　(c) $\theta = 60°$

図 3.11 MUSIC法による周波数毎のDOA推定結果(Case D)

次に，MUSIC法によるDOA推定値の標準偏差について述べる．表3.4の（）内の数値と図3.9の(b)と(d)から，MUSIC法によるDOA推定値の標準偏差は，θに比例して急激に大きくなることが読み取れる．まず，Case Dの場合，標準偏差は$\theta = 0°$のとき11.76°，$\theta = 75°$のとき22.47°となり，θが大きくなるに伴い10°以上も増加する．500Hz未満の周波数を除外したCase D†の場合，Case Dに比べて標準偏差は小さくなるが，$\theta = 0°$のとき1.11°，$\theta = 75°$のとき12.31°となり，θに比例して標準偏差が大きくなる問題は解決しない．また，最も標準偏差が小さいCase D†の場合でも，DUET法に基づくDOA推定法の標準偏差(表3.1 Case A)より，全てのθで標準偏差が大きいことがわかる．さらに，音声区間20フレームを用いたCase CやCase C†の場合，全フレームを用いたCase DやCase D†に比べ，標準偏差が全てのθにおいて10°程度大きくなることが見てとれる．

MUSIC法によるDOA推定についてより詳しく検証するため，Case CとCase Dにおける周波数毎のDOA推定値を調査してみた．その結果を，図3.10と図3.11にそれぞれ示

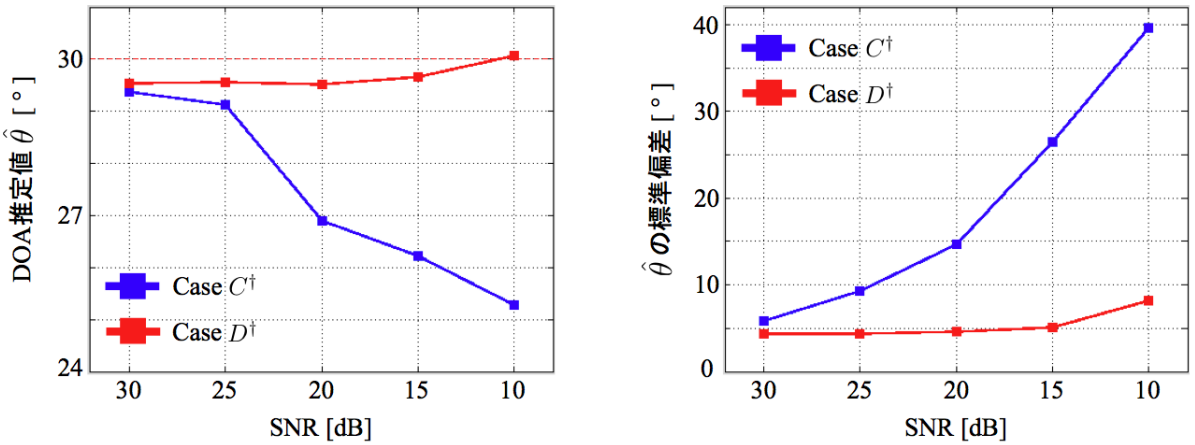
す．図中，縦軸はDOA[°]，横軸は周波数[Hz]，赤の実線はDOAの真値θ，青点はDOA推定値である．図3.10と図3.11より，500Hz未満でのDOA推定値は特に真値θと大きくかけ離れていることが見てとれる．このことは，特に，音声区間の連続する20フレームを用いたCase C (図3.10)のとき顕著にみられる．さらに，Case CではCase Dに比べて，DOA推定値のバラツキが全周波数帯域で大きいことがわかる．また，Case CとCase Dは共に，θに比例してDOA推定値のバラツキが大きくなることが見てとれる．これらのことは，表3.4や図3.9の結果と符合する．以上の結果から，MUSIC法を音声のような広帯域信号に適用する場合，使用する周波数帯域を限定する必要があることが改めて確認できた．また，使用するフレーム数が少ないほど，θが大きいほど，DOA推定値のバラツキが大きくなることも確認できた．これらの原因については後述する．

次に，MUSIC法によるDOA推定を行う場合，暗騒音や残響がどの程度の影響を及ぼすか調べるため，SNRと残響時間RT_{60}に対するDOA推定精度をシミュレーションによりそれぞれ検証した．ここでは，500Hz未満の周波数を除外し，音源の方位を$\theta = 30°$とした．まず，SNRがMUSIC法によるDOA推定に及ぼす影響を調べるため，残響時間RT_{60}を200[msec]として，SNRを30[dB]から10[dB]まで5[dB]ずつ変えてシミュレーションを行った．このときのDOA推定値の平均と標準偏差（( ) の数値）を表3.5に示す．表中，Case C†は音声区間の連続する20フレームを用いたときの結果で，Case D†は全フレームを用いたときの結果である．また，図3.12(a)にCase C†とCase D†のDOA推定値の平均，図3.12(b)に標準偏差を折れ線グラフで示す．図中，縦軸はDOA推定値$\hat{\theta}$とその標準偏差，横軸はSNR[dB]，青の折れ線グラフはCase C†，赤の折れ線グラフはCase D†，赤の破線は音源方位の真値($\theta = 30°$)である．表3.5と図3.12(a)から，SNRに対するDOA推定誤差はCase D†のとき概ね変わらないことが見てとれる．具体的には，推定誤差はSNRに関わらず全体的に1°程度に収まっている．一方，Case C†の場合，推定誤差はSNRが30[dB]と25[dB]のとき1°未満だが，SNRが20[dB]と15[dB]のとき約3° ~4°，SNRが10[dB]のとき約5°となり，SNRが低いほど大きくなる．

標準偏差についてはCase C†とCase D†は共に，SNRが低いほど大きくなることが表3.5と図3.12(b)から見てとれる．また，Case C†の標準偏差はCase D†に比べ全てのSNRで大きく，SNRに対する標準偏差の増加率も大きいことがわかる．具体的には，Case D†の標準偏差は，SNRが30[dB]から10[dB]と低くなるに伴い，4.30°から8.14°と約4°程度の増加

表 3.5 SNRに対するMUSIC法に基づくDOA推定値(θ=30°，RT_{60}=200[msec]，500Hz以下除外)

Case	SNR=30[dB]	SNR=25[dB]	SNR=20[dB]	SNR=15[dB]	SNR=10[dB]
C†	29.37° (5.88°)	29.12° (9.26°)	26.89° (14.70°)	26.22° (26.52°)	25.27° (39.56°)
D†	29.54° (4.30°)	29.56° (4.34°)	29.51° (4.56°)	29.66° (5.12°)	30.06° (8.14°)



(a) DOA推定値　　(b) DOA推定値の標準偏差

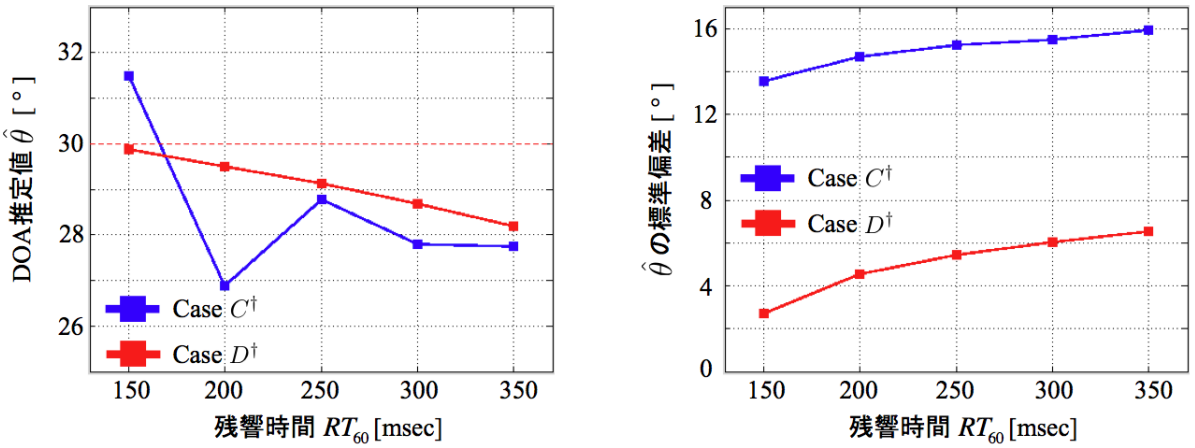
図 3.12 SNRに対するMUSIC法によるDOA推定値の折れ線グラフ(θ=30°，RT_{60}=200[msec]，500Hz以下除外)

に止まっているが，Case C†の標準偏差は，SNRが30[dB]から10[dB]と低くなるに伴い，5.88°から39.56°と約34°ほど急増する．これらの結果から，MUSIC法によるDOA推定の場合，使用するフレームが少ないときSNRの影響を大きく受けることが確認できた．

次に，残響時間RT_{60}がMUSIC法によるDOA推定に及ぼす影響を調べるため，SNRを20[dB]として，RT_{60}を150[msec]から350[msec]まで50[msec]ずつ変えてシミュレーションを行った．このときのDOA推定値の平均と標準偏差（()の数値）を表3.6に示す．また，図3.13(a)にCase C†とCase D†のDOA推定値の平均，図3.13(b)に標準偏差を折れ線グラフで示す．表3.6と図3.13((a)から，Case D†の場合，DOA推定誤差は残響時間RT_{60}に比例して大きくなることがわかる．具体的には，RT_{60}が150[msec]から350[msec]と長くなるの

表 3.6 RT_{60} に対するMUSIC法によるDOA推定値(θ=30°，SNR=20[dB]，500Hz以下除外)

Case	RT_{60}=150[msec]	RT_{60}=200[msec]	RT_{60}=250[msec]	RT_{60}=300[msec]	RT_{60}=350[msec]
C†	31.49° (13.53°)	26.89° (14.70°)	28.77° (15.23°)	27.79° (15.50°)	27.76° (15.93°)
D†	29.88° (2.73°)	29.51° (4.56°)	29.13° (5.46°)	28.69° (6.03°)	28.19° (6.54°)



(a) DOA推定値　　(b) DOA推定値の標準偏差

図 3.13 RT_{60} に対するMUSIC法によるDOA推定値の折れ線グラフ(θ=30°，SNR=20[dB]，500Hz以下除外)

に対し，DOA推定誤差は約1.5°ほど大きくなる．しかし，全体的なDOA推定誤差は2°以内に収まっている．Case C†の場合も同様に，DOA推定誤差は残響時間RT_{60}に比例して大きくなる傾向が見られる．また，DOA推定誤差は概ね2~ 3°程度であり，Case D†に比べて大きいことがわかる．

標準偏差についてはCase C†とCase D†は共に，残響時間RT_{60}に比例して大きくなることが表3.6と図3.13(b)から見てとれる．また，Case C†の標準偏差はCase D†に比べ全てのRT_{60}において，概ね10°ほど大きいことがわかる．しかし，残響時間RT_{60}に対しての標準偏差の増加率はCase C†もCase D†でも変わらない．具体的には，Case D†の標準偏差は，RT_{60}が150mseのとき2.73°，RT_{60}が350mseのとき6.54°となって約4°大きくなるのに

対し，Case C†の標準偏差は，RT_{60}が150mseのとき13.53°，RT_{60}が350mseのとき15.93°となって約2°大きくなる．これらの結果から，MUSIC法によるDOA推定の場合，使用するフレームが少なくても残響時間RT_{60}の影響は少ないといえる．以下では，MUSIC法において，使用するフレーム数が少なくθが大きいほどDOA推定値のバラツキが大きくなる原因について考察する．

MUSIC法よる場合，推定値のバラツキ，つまり推定値の分散はCramer-Rao Bound

$$CRB \approx \frac{6\sigma^2}{\rho^2 KM(M^2-1)d^2\cos^2\theta} \tag{3.1}$$

に従うことが知られている[41][42][43][44]．ここに，ρ^2とσ^2はそれぞれ信号と雑音の平均パワー，dは隣接センサ間の間隔である．式(3.1)は，方位θがブロードサイド方向（0°）から外れるほど，またSN比が低くデータ数（K）やマイクロホン数（M）が少ないほど，隣接センサ間の間隔が狭いほど，推定値がバラツクことを意味している．表3.4と表3.5の標準偏差の結果から，推定値のバラツキはいずれのCaseでもCramer-Rao Boundに従う結果となっていることが読み取れる．すなわち，Case CとCase Dの比較，あるいはCase C†とCase D†の比較から分かるように，フレーム数が少なくなると標準偏差は急激に大きくなることが確認できる．そして，SNRが低いほど，θが大きいほど，標準偏差が大きくなることも確認できる．また，標準偏差が一番小さいCase D†の場合でも，DUET法に基づくDOA推定値の標準偏差よりも大きいことが，表3.1との対比から分かる．これとは対照的に，DUET法に基づくDOA推定の場合，前述のように，推定値のバラツキはフレーム数が少なくなってもあまり変わらない．このことは，DUET法に基づくDOA推定法のMUSIC法に対する優位な点と考えられる．

以上のDOA推定値やそのバラツキに関する考察結果から，小数のフレームでDOAを推定しようとする場合，目的音源の方位を–30° ~ 30°程度の範囲に絞り込めば，DUET法に基づくDOA推定法の方がMUSIC法より有利と考えられる．

3.4 まとめ

本章では，少数フレームによる音声のDOA推定に関する予備的検討を行った．音声は音声区間と無音区間を繰り返す断続波であり，音声のDOAに関する情報は，音声区間

にだけ含まれ，無音区間には含まれないが，現実には無関係な雑音成分が無音区間に重畳することを指摘した．したがって，フレーム数が少ない場合，無音区間が雑音や残響の影響を受けることから，音声区間内の連続する20フレームでの局所DOAをもとに，DUET法に基づく方法とMUSIC法によりDOA推定した場合，どの程度の推定精度が得られるかをシミュレーションにより調べた．

DUET法に基づく方法により，音声区間の連続する20フレームを用いてDOA推定した場合，対象音源のDOAの真値(θ)が30°以内ならば，推定誤差は3°未満に収まるが，θ=45°を越えると7°以上となって推定精度は急激に劣化する結果となった．このことは，全フレームを用いた場合でも同様であった．そこで，θが0°から外れるにつれて推定誤差が大きくなる原因について考察し，真値(θ)の0°からの変位に比べて，局所DOA頻度分布のピークの変位は小さいため，結果として得られるDOA推定値は大きな誤差をもつことを明らかにした．ただし，標準偏差は全体的に3°未満に収まり，全フレームを用いた場合と音声区間の20フレームを用いた場合とで相違ない結果が得られた．

一方，MUSIC法により，音声区間の連続する20フレームを用いた場合，DOA推定誤差は$\theta \leq 60°$までは3°程度に収まるが，標準偏差は20°から30°とかなり大きくなることが判明した．この場合，全フレームを用いたり低周波数ビンを除外すればある程度は改善されるが，それでもDUET法に基づくDOA推定値の標準偏差を大きく超えていた．そこで，その原因について考察し，MUSIC法による場合，推定値の分散はCramer-Rao Boundに従うことを確認した．すなわち，MUSIC法の場合，真値θがブロードサイド方向から外れるほど，またSN比が小さくデータ数やマイクロホン数が少ないほど，推定値がバラツクことから，フレーム数が少ないときのDOA推定には不向きであることを指摘した．一方，DUET法に基づくDOA推定の場合，推定値のバラツキはフレーム数が少なくてもあまり変わらず，目的音源の方位を$-30° \sim 30°$程度に絞り込めば，MUSIC法より良好にDOAを推定できることを述べた．

第4章

フレーム単位のDOA推定

人の発話は音声区間と無音区間に分けられ，その比率は7対3程度とされている[45]. また，会話や会議では暗黙のマナーとして他人の発言中に口を挟むことは少ない. そのため，多数の人が集まる会議の場でも，同時にマイクロホンに入る音声は多くて２～３発話であろうと考えられる. 現実に，２つの音源(音源１(図4.1上段)と音源２(図4.1中段))が同時に発話したときに観測されるマイクロホン収録音(図4.1下段)は，図4.1中に示される区間 (a)のように定常雑音と１つの音声が存在する区間，図4.1中の区間 (b)のように定常雑音と複数の音声が存在する区間，図4.1中の区間 (c)のように方向性のない定常雑音(暗騒音)のみが存在する区間に分けられる.

そこで，目的音源の方位がブロードサイド方向からみて±30°の範囲にある場合を対象に，マイクロホン収録音声を，１つの音声が存在する区間（１音源区間），２つ以上の音声が存在する区間（複数音源区間），無音源区間（暗騒音区間）の３つの区間に分けて，フレーム単位に1音源区間を選別し，その１音源フレームから目的音源のDOAを推定する手法を提案する.

4.1 単一フレームにおける局所DOA頻度分布

3.1節と同じシミュレーション環境で，２つの音源$s_1^*(t)$と$s_2^*(t)$がマイクロホン対の中心から50cm離れて$\theta_1 = 30°$と$\theta_2 = -30°$の方位にある例を考える．このとき，局所DOA$\{\tilde{\theta}[k,\omega_l] \,|\, l = 0,1,\cdots,255\}$がフレーム単位に$\theta_q^\circ$=$[q° - 91.5°, q° - 90.5°)$に入る頻度を

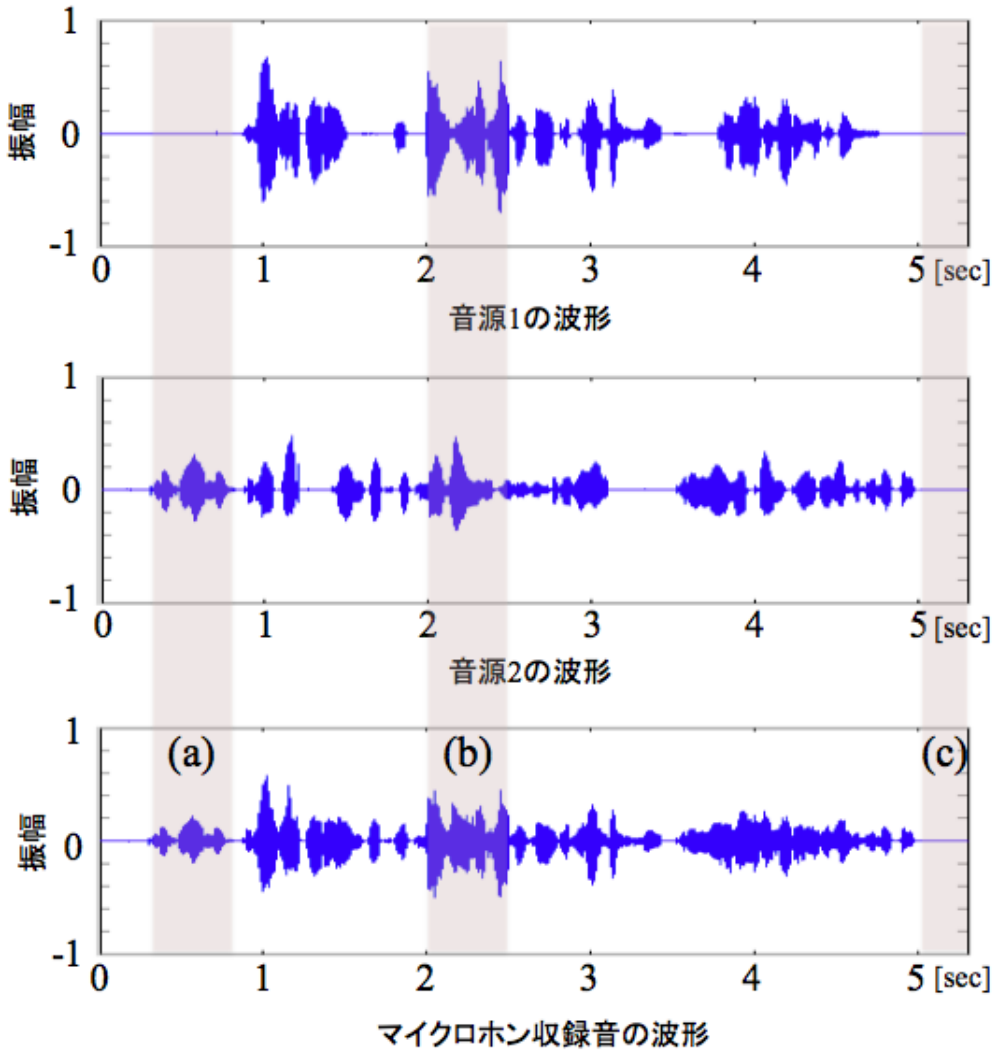


図 4.1 マイクロホン収録音の区分

$p_q(k)$として，その分布を$P_k(\tilde{\theta})$=[$p_1(k), p_2(k), \cdots, p_{181}(k)$]と求めてみた．このときの大まかな流れの概観を図4.2に示す.

その場合，各フレームでの分布形状は，サンプル数が256個に減った分，図3.3や図3.4に比べて荒々しくはなるが，図4.3のように３種類に分類される．すなわち，図4.3 (a)のようにθ_1かθ_2に１つのピークをもつ単峰的な分布（１音源フレーム），図4.3 (b)のようにθ_1とθ_2に２つのピークをもつ双峰的な分布（2音原フレーム），図4.3 (c)のように際立ったピークのない比較的平坦な分布（暗騒音フレーム）の３種類に分類される．したがって，この3種類が弁別できれば，１音源フレームや２音源フレームでは，分布の

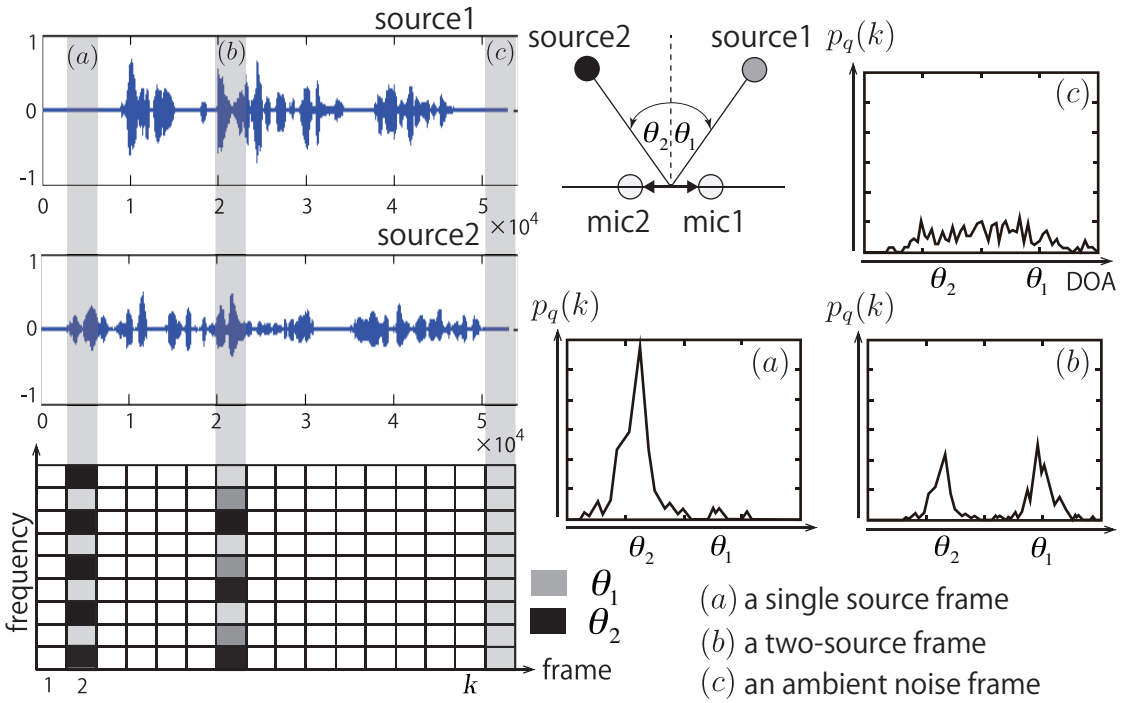


図 4.2 単一フレームでの局所DOA頻度分布についての概観図

ピークを探索することにより，音源のDOAを推定できる．

4.2 スパース尺度による１音源フレームの選別

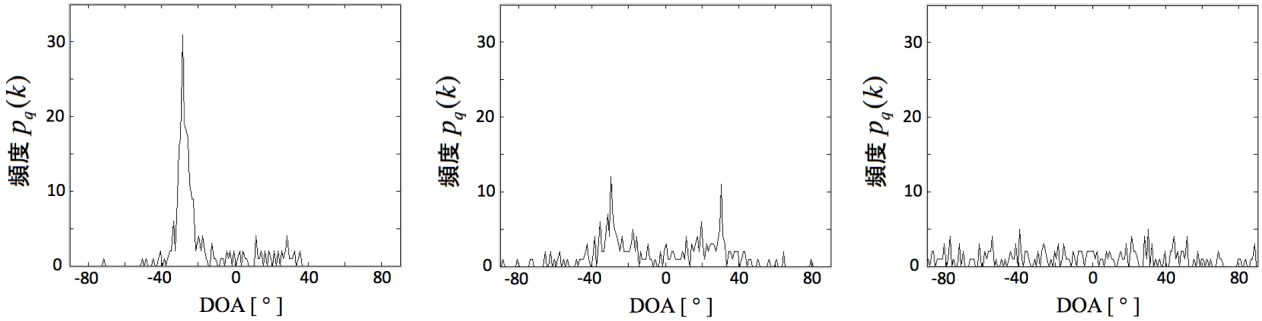
スパース性（WDO）を測る尺度として，次式で定義されるKurtosisやGini係数，Hoyer尺度などがある[46].

$$\mathscr{K}(P_k(\tilde{\theta})) = \frac{\sum_{q=1}^{Q} p_q^4(k)}{\left(\sum_{q=1}^{Q} p_q^2(k)\right)^2} \tag{4.1}$$

$$\mathscr{G}(P_k(\tilde{\theta})) = 1 - 2\sum_{q=1}^{Q} \frac{p_q(k)}{\sum_{q=1}^{Q} p_q(k)}\left(\frac{Q-q+\frac{1}{2}}{Q}\right) \tag{4.2}$$

$$\mathscr{H}(P_k(\tilde{\theta})) = \left(\sqrt{Q} - \frac{\sum_{q=1}^{Q} p_q(k)}{\sqrt{\sum_{q=1}^{Q} p_q^2(k)}}\right)\left(\sqrt{Q}-1\right)^{-1} \tag{4.3}$$

これらは，図4.3 (a)のように単峰的な分布，図4.3 (b)のように双峰的な分布，図4.3 (c)のように比較的平坦な分布の順に，高い値をとる．



(a) a single source frame (b) a two-source frame (c) an ambient noise frame

図 4.3 単一フレームでの局所DOA頻度分布

表 4.1 スパース尺度の値

	(a)	(b)	(c)
Kurtosis	0.051	0.021	0.018
(規格化後の値)	(1)	(0.41)	(0.35)
Gini	0.74	0.55	0.48
(規格化後の値)	(1)	(0.74)	(0.65)
Hoyer	0.54	0.35	0.29
(規格化後の値)	(1)	(0.65)	(0.53)

図4.3の3つの分布に対して，各スパース尺度(Kurtosis $\mathcal{K}(P_k(\tilde{\theta}))$，Gini係数 $\mathcal{G}(P_k(\tilde{\theta}))$，Hoyer尺度 $\mathcal{H}(P_k(\tilde{\theta}))$)の値を求めた．その結果を表4.1に示す．表中の数値は，各分布に対するそれぞれのスパース尺度の値であり，()内の数値は，それぞれの最大値((a)のときの値)で規格化した数値である．Kurtosis $\mathcal{K}(P_k(\tilde{\theta}))$は，分布(a)のとき 0.051，分布(b)のとき 0.021，分布(c)のとき 0.018となり，(b)と(c)の差は一番小さいが，(a)と(c)の差は一番大きく 3 倍近い差となった．Gini係数 $\mathcal{G}(P_k(\tilde{\theta}))$は，分布(a)のとき 0.74，分布(b)のとき 0.55，分布(c)のとき 0.48となり，(b)と(c)の差はKurtosisよりも大きいが，(a)と(c)の差は一番小さい結果となった．Hoyer尺度 $\mathcal{H}(P_k(\tilde{\theta}))$は，分布(a)のとき 0.54，分布(b)のとき 0.35，分布(c)のとき 0.29となり，(b)と(c)の差は一番大きく，(a)と(c)にも倍近い差が得られた．以上の結果から，それぞれに若干の特色があるが，スパース尺度により，1 音源フレーム，2音源フレーム，暗騒音フレームが識別できることが確認された[48][49].

次に，スパース尺度の閾値（$\mathscr{K}(P_k(\tilde{\theta})) > \eta$, $\mathscr{G}(P_k(\tilde{\theta})) > \eta$, $\mathscr{H}(P_k(\tilde{\theta})) > \eta$）をどのように定めれば１音源フレームが取り出せるかを検討するため，**3.1**節と同じ環境のもとで，音源方向を$\theta = 30°$として，シミュレーションを行った．Kurtosisの閾値と１音源フレーム選別の関係を図4.4に，Gini係数の閾値と１音源フレーム選別の関係を図4.5に，Hoyer尺度の閾値と１音源フレーム選別の関係を図4.6にそれぞれ示す．各図は，横軸をフレーム番号（k）として，局所DOA頻度分布$P_k(\tilde{\theta})$のピークを採る方位ϕ_k $(k = 1, 2, \cdots, K)$ を縦軸にプロットしたものである．スパース尺度は常に正の値をとることから，閾値を$\eta = 0$とした場合，全フレームでϕ_kが得られる．そのため，図4.4(a)，図4.5(a)，図4.6(a)のように，前半と後半の暗騒音フレームでのϕ_kは広くばらついてプロットされることになる．まず，図4.4に示すKurtosis$\mathscr{K}(P_k(\tilde{\theta}))$の閾値と１音源フレームの関係について考える．$\mathscr{K}(P_k(\tilde{\theta})) > 0.03$とした場合(図4.4(b))，暗騒音フレームでのプロット点はほぼ消滅する．さらに，$\mathscr{K}(P_k(\tilde{\theta})) > 0.04$ (図4.4(c))，$\mathscr{K}(P_k(\tilde{\theta})) > 0.05$ (図4.4(d))とした場合，ほぼ１音源フレームのみが選別されると考えられ，そこでのϕ_kは音源の方位$\theta = 30°$にほぼ一致している．次に，図4.5に示すGini係数$\mathscr{K}(P_k(\tilde{\theta}))$の閾値と１音源フレームの関係について考える． $\mathscr{G}(P_k(\tilde{\theta})) > 0.5$とした場合(図4.5(b))，暗騒音フレームでのプロット点は消滅せずにϕ_kは広くばらついてプロットされている．一方，$\mathscr{G}(P_k(\tilde{\theta})) > 0.6$ (図4.5(c))の場合，暗騒音フレームでのプロット点はほぼ消滅する．さらに，$\mathscr{G}(P_k(\tilde{\theta})) > 0.7$ (図4.5(d))の場合，ほぼ１音源フレームのみが選別されると考えられ，そこでのϕ_kは音源の方位$\theta = 30°$にほぼ一致している．最後に，図4.6に示すHoyer尺度$\mathscr{H}(P_k(\tilde{\theta}))$の閾値と１音源フレームの関係について考える．$\mathscr{H}(P_k(\tilde{\theta})) > 0.3$とした場合(図4.6(b))，暗騒音フレームでのプロット点は消滅せずにϕ_kは広くばらついてプロットされている．一方，閾値を$\eta = 0.4$とした場合(図4.6(c))，暗騒音フレームでのプロット点はほぼ消滅する．さらに，閾値を$\eta = 0.5$とした場合，ほぼ１音源フレームのみが選別されると考えられ，そこでのϕ_kは図4.6(d) のように音源の方位$\theta = 30°$にほぼ一致している．

以上の結果から，スパース尺度の閾値を適切に設定することにより，ほぼ１音源フレームのみ選別でき，そのときのϕ_kは音源の方位θにほぼ一致することが確認できた．Kurtosisは，2音源フレームの選別を視野に入れた場合，暗騒音フレームとの識別が難しい．さらに，異常値に対して脆弱であることの報告もある[47]．また，Gini係数は頻度$p_q(k)$を小さい順に並べ替えなければならない制約がある．そこで，以下ではHoyer尺度

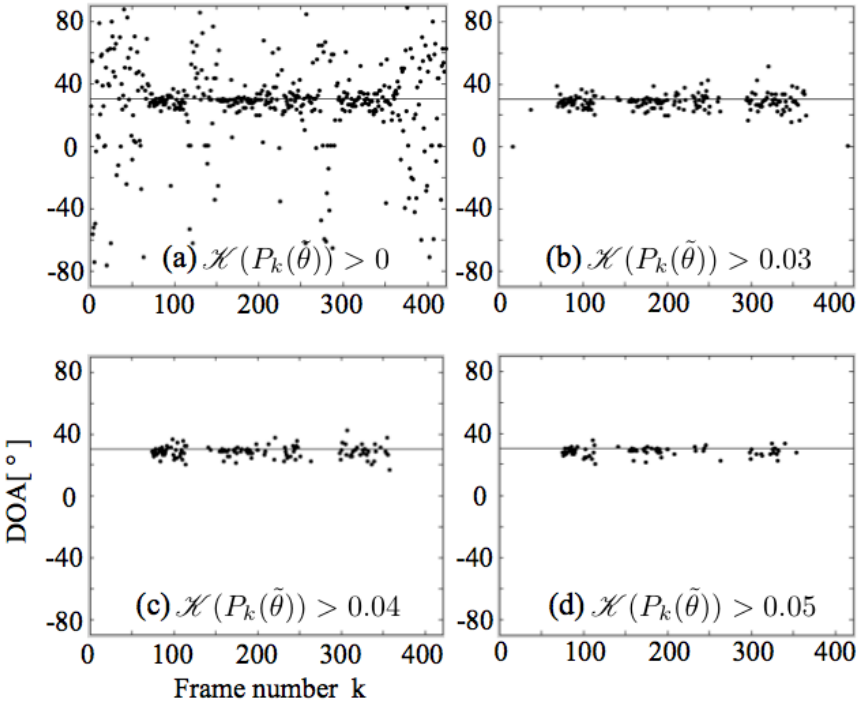


図 4.4 Kurtosisの閾値と1音源フレーム選別の関係

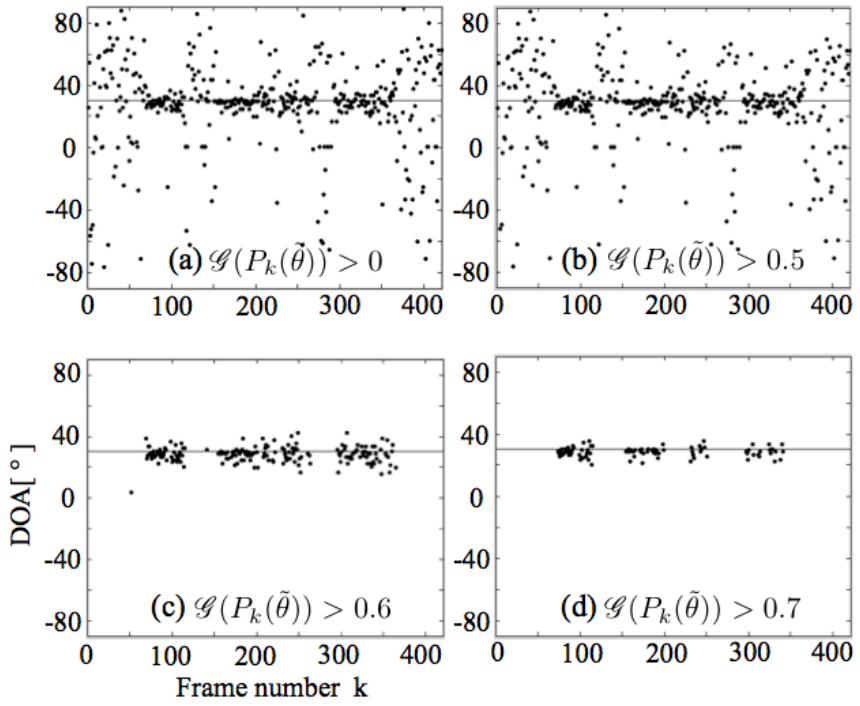


図 4.5 Gini係数の閾値と1音源フレーム選別の関係

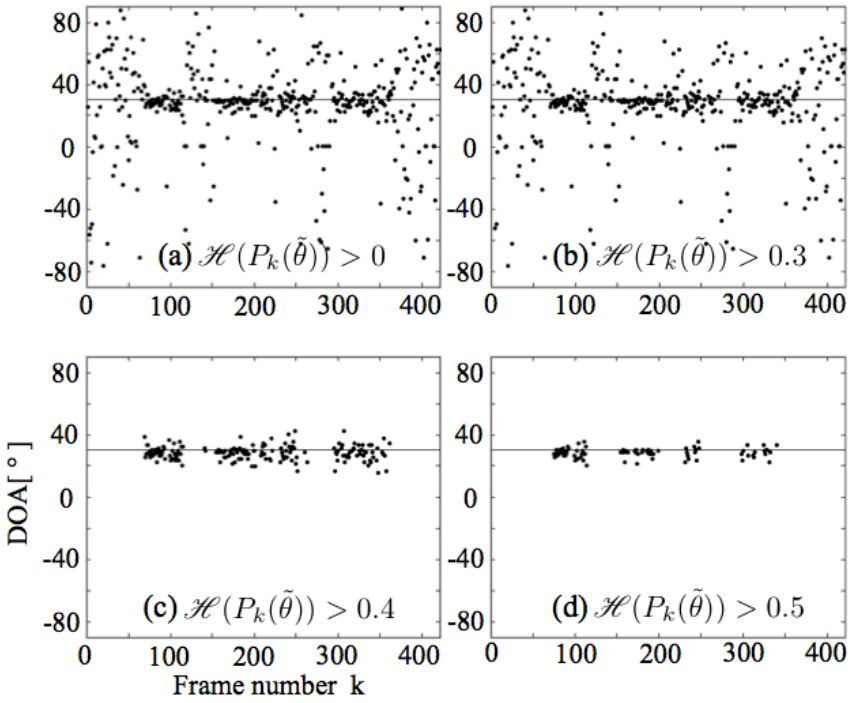


図 4.6 Hoyer尺度の閾値と1音源フレーム選別の関係

を用いる．

以上の見解が妥当であるか否かを確認するため，$\mathscr{H}(P_k(\tilde{\theta}))$による 1 音源フレームの検出結果と実際の音声区間を比較した．各フレームに対する$\mathscr{H}(P_k(\tilde{\theta}))$の値をプロットしたものを図4.7に示す．図の横軸はフレーム番号，縦軸は$\mathscr{H}(P_k(\tilde{\theta}))$の値である．また，赤の実線は音声の始端，赤の破線は音声の終端を表し，上部横方向に並ぶ黒点は，閾値を0.5として，そのフレームが 1 音源フレームと判定されたことを示している．この場合，図の結果から$\mathscr{H}(P_k(\tilde{\theta}))$により選別された 1 音源フレームは，完全に実際の音声区間内に収まっていることがみてとれる．一方，閾値を0.4とした場合，図4.6のと同様に，暗騒音区間で 1 音源フレームと判定される例はほぼ削除されるが，それでも誤って 1 音源フレームと判定された例が散見された．以上のことから，$\mathscr{H}(P_k(\tilde{\theta}))$の閾値を0.5とすれば，1 音源区間内のフレームのみが選別されることが確認できる．

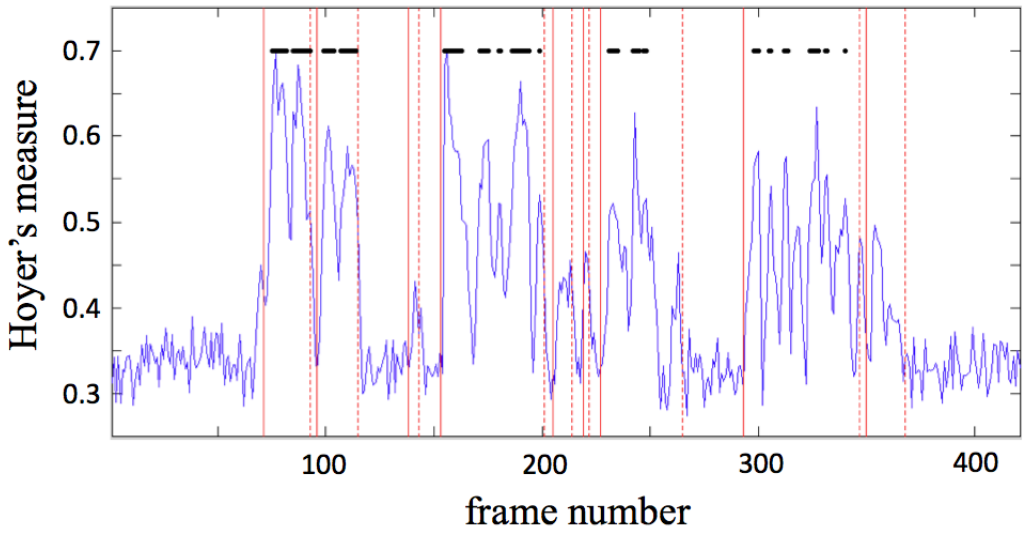


図 4.7 各フレームに対するHoyer尺度の値

4.3 フレーム単位のDOA推定

上述のことから分かるように，局所DOAの頻度分布$P_k(\tilde{\theta})$でHoyer尺度が0.5以上となるフレームを１音源フレームと判定し，そのフレームでの頻度分布のピークを探索することにより音源のDOAが推定できる．したがって，１音源フレームと判定されたフレームの番号を$\check{k}$と表記し，そのフレーム$\check{k}$のみについて，頻度分布$P_{\check{k}}(\tilde{\theta})$=$[p_1(\check{k}), p_2(\check{k}), \cdots, p_{181}(\check{k})]$の頻度$p_q(\check{k})$が最大となる角度番号$q$を

$$\hat{q} = \arg\max_{q} \{p_q(\check{k}) \mid q = 1, 2, \cdots, 181\} \tag{4.4}$$

のように探索して，音源のDOAを$\hat{\theta} = (\hat{q} - 91)^\circ$と推定することにする．以上のように１音源フレームを検出して，DOAをフレーム単位に推定する方法を提案法としてまとめると次のようになる．また，提案法の流れを図4.8に示す.

1. ２つのマイクロホンでの観測値をフレーム（k）毎にSTDFTして複素スペクトル$x_m[k, \omega_l]$ $(m = 1, 2; l = 0, 1, \cdots, 255)$ を求める．

2. $x_2[k, \omega_l]\bar{x}_1[k, \omega_l]$をもとに式(2.37)(2.38)のように局所DOA $\tilde{\theta}[k, \omega_l]$を求める．

3. 局所DOAがθ_q°=$[q^\circ - 91.5^\circ, q^\circ - 90.5^\circ)$に入る頻度を$p_q(k)$として頻度分布$P_k(\tilde{\theta})$=$[p_1(k), \cdots, p_{181}(k)]$を求める．

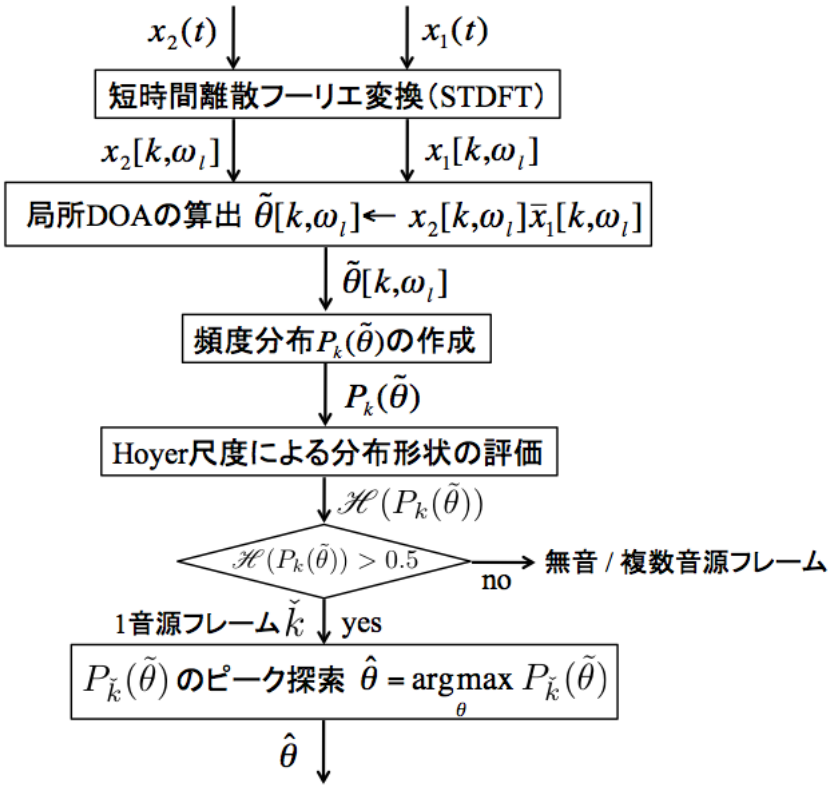


図 4.8 フレーム単位のDOA推定の流れ

4. $P_k(\tilde{\theta})$を式(4.3)のHoyer尺度$\mathscr{H}(P_k(\tilde{\theta}))$で測り，その値が0.5以上となるときを 1 音源フレーム（$\check{k}$）と判定し，そうでなければ 1. に戻る．

5. 1音源フレームの頻度分布$P_{\check{k}}(\tilde{\theta})$で頻度$p_q(\check{k})$が最大となるときの$q$を$\hat{q}$として，$\hat{\theta}=(\hat{q}-91)^\circ$を音源のDOA推定値とする．

4.4 まとめ

本章では，DUET法に基づくフレーム単位のDOA推定法を提案した．具体的には，まず，マイクロホンで観測された混合信号は，無音源区間，1 音源区間，複数音源区間に分けられることを言及した．次に，混合信号を短時間離散フーリエ変換して得られる複素スペクトルの位相差をもとに，フレーム単位に時間周波数における局所DOAを求めて，その頻度分布をとれば 3 種類の形状に分類されることを述べた．すなわち，1 音源フレームでは 1 つのピークをもつ単峰的な分布，2 音源フレームでは 2 つのピーク

をもつ双峰的な分布，無音源フレームではピークのない比較的平坦な分布となることを述べた．そして，1 音源フレームではピークが明確な単峰的な分布となるため，1 音源フレームが選別できればその分布のピークを探索することでDOA推定ができることを指摘した．以上のことから，フレーム単位のDOA推定を行うためには，1 音源フレームの選別が重要であることを明らかにした．分布の形状を測る尺度として，Hoyer尺度などのスパース尺度があり，適切な閾値を設けることで 1 音源フレームを選別できることをシミュレーションにより確認した．また，このとき選別された 1 音源フレームのピークは音源の方位にほぼ一致した．最後に，Hoyer尺度による 1 音源フレームの検出結果と実際の音声区間を比較して，Hoyer尺度の閾値を0.5とすれば選別されたフレームは全て音声区間内に収まることを確認した．

第5章

シミュレーションおよび考察

前章では，フレーム毎に局所DOA頻度分布を求めて，Hoyer尺度を評価基準として1音源フレームを選別した後，分布のピークを探索してピーク時の方位をDOA推定値として採択する方法を提案した．ここでは，提案法の有効性をシミュレーションにより検証する．

最初に，単一音源を対象にしたシミュレーションを行い，音源の方位，残響時間，SN比に対する提案法のDOA推定精度を検証する．その中で，音源がブロードサイド方向から±30°の範囲にあり，残響時間が250[msec]以下，SN比が15[dB]以上の環境下で，提案法によるDOAの推定誤差は2°未満，標準偏差は概ね3°未満に収まることを明らかにする．また，残響時間とSN比に対する1音源フレームの検出個数の差異を明らかにする．

次に，目的音源の方位をブロードサイド方向から±30°の範囲に絞り込み，目的音源1個，妨害音源2個の3音源環境下でシミュレーションを行う．その中で，残響時間が200[msec]以下，SN比が15[dB]以上であれば，提案法によるDOA推定誤差は3°未満，標準偏差は概ね3°未満に収まることを述べる．

5.1 提案法による単一音源のDOA推定

提案法の有効性を確認するため，**3.1**節と同じシミュレーション環境下で，提案法を適用してフレーム単位にDOAを推定した．このときの結果を表5.1に示す．表中の数値はDOA推定値の平均，()内の数値は標準偏差を表す．表5.1より，推定誤差は$\theta \leq 30°$までは

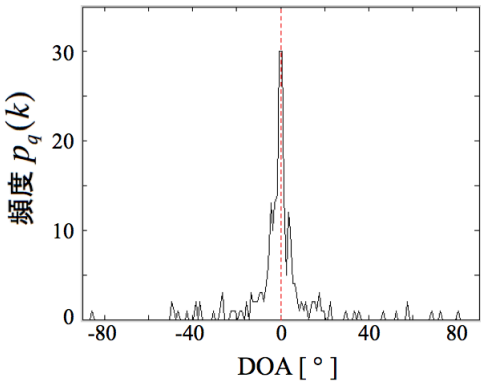
表 5.1 提案法による単一音源のDOA推定値(RT_{60}=200[msec]，SNR=20[dB])

SNR	$\theta = 0°$	$\theta = 15°$	$\theta = 30°$	$\theta = 45°$	$\theta = 60°$	$\theta = 75°$
20[dB]	0.27° (1.71°)	14.95° (2.62°)	28.59° (2.82°)	39.86° (2.66°)	49.15° (3.02°)	54.71° (3.00°)

2°未満に収まっているが，$\theta = 45°$では約5°程度，$\theta = 60°$では約10°程度，$\theta = 75°$では約20°程度と，θが30°を超えると急激に増えることがわかる．この傾向は，全フレームや音声区間の連続する20フレームを用いてDOA推定したときの表3.1の結果と同様である．

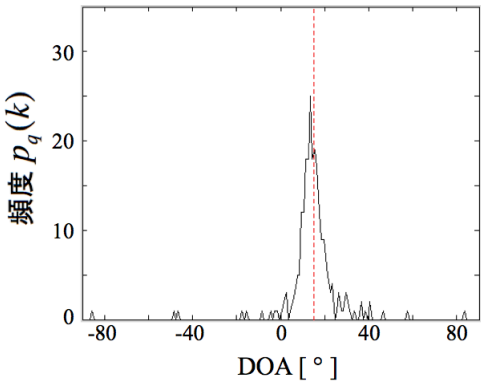
そこで，$\theta = 45°$を越えると推定精度が急激に劣化する原因を探るため，方位θを0°から75°まで15°ずつ変えて配置したときに得られる1音源フレームの局所DOAの頻度分布を調べた．このときの局所DOAの頻度分布を図5.1に示す．図の縦軸は頻度$p_q(k)$，横軸はDOA[°]，赤の破線はDOAの真値($\theta = 0 \sim 75°$)である．また，図中，(a)は$\theta = 0°$での局所DOAの頻度分布，(b)は$\theta = 15°$での局所DOAの頻度分布，(c)は$\theta = 30°$での局所DOAの頻度分布，(d)は$\theta = 45°$での局所DOAの頻度分布，(e)は$\theta = 60°$での局所DOAの頻度分布，(f)は$\theta = 75°$での局所DOAの頻度分布である．分布(a)，分布(b)，分布(c)はそれぞれ0°，15°，30°近傍でピークを採り，それぞれのDOAの真値θとほぼ一致することが見てとれる．一方，分布(d)，分布(e)，分布(f)のピークはそれぞれのDOAの真値θと大きくかけ離れていることがわかる．具体的には，分布(d)のピークは40°付近にあることが見てとれ，表5.1で$\theta = 45°$のとき推定誤差が約5°となることと符合する．同様に，分布(e)のピークは50°付近にあることが見てとれ，5.1で$\theta = 60°$のとき推定誤差が約10°となることと符合する．分布(f)に関しても，上述と同じことが確認できる．さらに，平均（標準偏差）を計算してみたところ，その値は分布(a)で-0.52°（17.13°），分布(b)で14.44°（12.68°），分布(c)で27.76°（14.63°）となり，それぞれのDOAの真値θとほぼ一致する結果となった．また，分布(d)の平均（標準偏差）は38.99°（16.02°），分布(e)で43.19°（22.81°），分布(f)で52.50°（19.58°）となり，それぞれのDOAの真値θとは大きく乖離する結果となった．このような真値との乖離は，3章と同様の理由であると考えられる．

一方，標準偏差は方位θに関わらず全体的に３°程度に収まっている．このことと表3.1を比較すると，提案法の標準偏差は，音声区間の連続する20フレームを用いた場合とほ



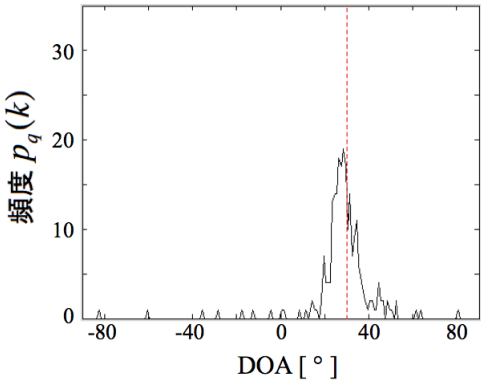
(a) $\theta = 0°$

平均：-0.52°，標準偏差：17.13°



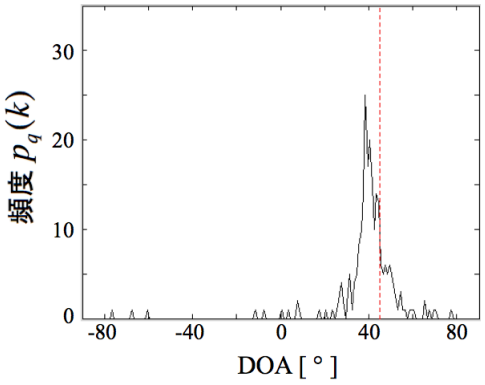
(b) $\theta = 15°$

平均：14.44°，標準偏差：12.68°



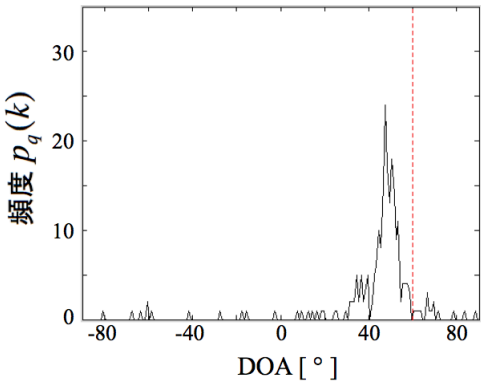
(c) $\theta = 30°$

平均：27.76°，標準偏差：14.63°



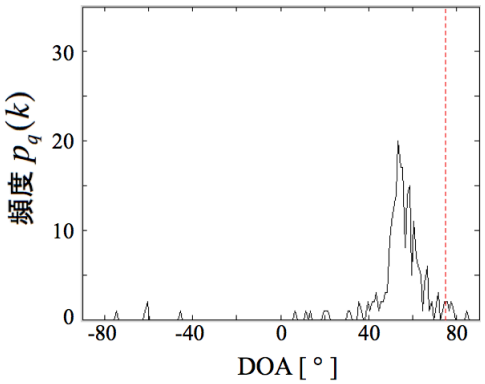
(d) $\theta = 45°$

平均：38.99°，標準偏差：16.02°



(e) $\theta = 60°$

平均：43.19°，標準偏差：22.81°



(f) $\theta = 75°$

平均：52.50°，標準偏差：19.58°

図 5.1 1 音源フレームの局所DOA頻度分布(RT_{60}=200[msec]，SNR=20[dB])

ぼ同程度であり，全フレームを用いた場合と比べても概ね２°程度の差であることがわかる．すなわち，提案法は多数のフレームを用いた場合と遜色なく機能している．以上の結果から，DOA推定方位を$-30° \leq \theta \leq 30°$の範囲に絞り込めば，提案法によりフレーム単位でも推定誤差が２°未満で，標準偏差が３°程度の精度でDOA推定できることが確認できる．このように，音源がマイクロホンの正面付近($\theta = \pm 30°$の範囲)から到来するという設定は，ステレオマイク付ICレコーダに音声を録音する際，マイクロホン対を話者の方向に向ける習慣等とも合致しているため，多くの状況で成り立つと考えられる．

次に，提案法が有効に機能する環境の範囲を確認するため，上述と同じシミュレーション環境下で，残響時間RT_{60}とSNRを変えてシミュレーションを行った．具体的には，残響時間RT_{60}が150[msec]，200[msec]，250[msec]，300[msec]の４つの場合について，SNRをそれぞれ30[dB]，20[dB]，15[dB]，10[dB]と変えてシミュレーションを行った．このときの結果を表5.2，表5.3，表5.4，表5.5に示す．表中の数値はDOA推定値の平均，()内の数値は標準偏差を表す．また，各表におけるDOA推定値の平均と標準偏差を棒グラフ化したものを図5.2，図5.3，図5.4，図5.5に示す．青の棒グラフがSNR = 30[dB]のときの結果，赤の棒グラフがSNR = 20[dB]のときの結果，緑の棒グラフがSNR = 15[dB]のときの結果，黄色の棒グラフがSNR = 10[dB]のときの結果，黒の棒グラフが音源の方位θである．

以上の結果から分かるように，残響時間RT_{60}(150[msec]，200[msec]，250[msec]，300[msec])やSNR(30[dB]，20[dB]，15[dB]，10[dB])の値に関わらず，表5.1と同様に，推定誤差はDOA推定方位が$\theta \leq 30°$ならば２°未満，標準偏差は全方位($\theta = 0 \sim 75°$)で概ね３°程度に収まっていることがわかる．このことは，提案法で１音源フレームさえ検出できれば，残響時間やSN比に依存することなく，DOAを高精度に推定できることを示唆している．しかし，細かく見てみると，残響時間RT_{60}が長くてSNRが低いときの方が，僅かではあるが，推定誤差や標準偏差が小さくなっている．具体的には，表5.2において，SNR=15[dB]に対するθ=30°のときのDOA推定値は28.96°，標準偏差は3.05°であるが，表5.3において同一条件のときの結果をみてみると，DOA推定値は28.80°，標準偏差は2.67°となって，残響時間RT_{60}が200[msec]から250[msec]と長くなったにもかかわらず，標準偏差は小さくなる．また，表5.3において，SNR=10[dB]に対するθ=30°のときのDOA推定値は29.35°，標準偏差は2.20°となり，SNRが低くなったにもかかわらず，推定誤差と標準偏差はともに小さくなる．上述と同様のことが，図5.2，図5.3，図5.4，図5.5

からも云える．

この原因は，以下のように１音源フレームの検出個数の違いに起因していると考えられる．このことを説明するため，残響時間RT_{60}とSNRを変えて各組み合わせに対して６回の試行を行った状況で検出された１音源フレームの平均個数を表5.6に示す．表5.6中の数値は，１音源フレームの平均検出個数で,（）内の数値は１音源フレームが検出できた試行の割合である．表5.6より，１音源フレームの検出個数は残響時間が長くSNRが低いほど減少することがわかる．この原因は，残響や暗騒音の影響で局所DOA頻度分布$P_k(\tilde{\theta})$の形状がゆるやかになるからと考えられる．また，音源の方位を$\theta = 30°$，残響時間RT_{60}を200[msec]として，SNRをそれぞれ30[dB]，20[dB]，15[dB]，10[dB]と変えたときに得られた同一フレームでの局所DOA頻度分布の例を図5.6に示す．図中，(a)はSNR = 30[dB]での局所DOAの頻度分布，(b)はSNR = 20[dB]での局所DOAの頻度分布，(c)はSNR = 15[dB]での局所DOAの頻度分布，(d)はSNR = 10[dB]での局所DOAの頻度分布である．図5.6より，SNRが低いほど局所DOA頻度分布$P_k(\tilde{\theta})$の形状がゆるやかになることが確認できる．表5.6において，残響時間RT_{60}が150[msec]から300[msec]の場合に着目すると，１音源フレームの検出個数の減少率は，SNR=30[dB]のとき18%，SNR=20[dB]のとき34%，SNR=15[dB]のとき36%，SNR=10[dB]のときはほぼ変化なし，という結果である．次に，表5.6において，SNRを30[dB]から10[dB]とした場合に着目すると，１音源フレームの検出個数の減少率は，RT_{60}=150[msec]のとき95%，RT_{60}=200[msec]のとき96%，RT_{60}=250[msec]のとき95%，RT_{60}=300[msec]のとき95%となっている．また，SNR=10[dB]の場合，１音源フレームの検出できる割合は55%~77%となり，DOAが推定できない例が多発している．さらに，SNR=15[dB]の場合でも，RT_{60}=300[msec]のときに限り，１音源フレームが検出できない例が3%見られる．したがって，提案法は，１音源フレームが100%検出できる，残響時間RT_{60}が250[msec]以下，SN比が15[dB]以上で有効と判断される．

以上のことから，目的方位の範囲を$\theta = \pm 30°$に絞り込み，残響時間が250[msec]以下でSNRが15[dB]以上あれば，提案法によりフレーム単位で音源のDOAを推定誤差が２°未満，標準偏差が３°程度の精度で推定できることを確認した．次節では，３音源下で上記の条件を適用し，提案法の有効性を検証する．

表 5.2 提案法による単一音源のDOA推定値(RT_{60}=150[msec])

SNR	$\theta = 0°$	$\theta = 15°$	$\theta = 30°$	$\theta = 45°$	$\theta = 60°$	$\theta = 75°$
30[dB]	0.16° (1.27°)	14.79° (1.86°)	28.55° (1.90°)	40.26° (2.04°)	49.30° (2.21°)	54.92° (2.14°)
20[dB]	0.15° (1.92°)	14.96° (2.30°)	28.83° (2.52°)	40.40° (2.25°)	49.42° (2.96°)	55.13° (2.39°)
15[dB]	0.28° (1.95°)	14.96° (2.28°)	28.96° (3.05°)	40.00° (2.48°)	49.58° (2.83°)	55.46° (2.16°)
10[dB]	-0.66° (1.10°)	14.09° (1.57°)	29.02° (2.90°)	39.73° (2.57°)	48.56° (2.68°)	54.62° (2.86°)

表 5.3 提案法による単一音源のDOA推定値(RT_{60}=200[msec])

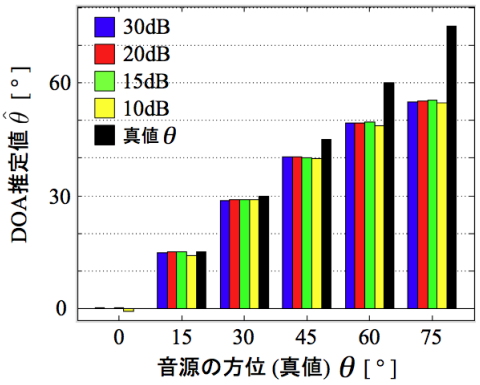
SNR	$\theta = 0°$	$\theta = 15°$	$\theta = 30°$	$\theta = 45°$	$\theta = 60°$	$\theta = 75°$
30[dB]	0.16° (1.27°)	14.73° (2.21°)	28.38° (2.45°)	39.91° (2.56°)	48.85° (2.82°)	54.68° (2.84°)
20[dB]	0.27° (1.71°)	14.95° (2.62°)	28.59° (2.82°)	39.86° (2.66°)	49.15° (3.02°)	54.71° (3.00°)
15[dB]	0.23° (1.89°)	14.96° (2.40°)	28.80° (2.67°)	40.09° (2.78°)	49.30° (3.08°)	54.77° (3.48°)
10[dB]	0.52° (2.01°)	14.90° (2.36°)	29.35° (2.20°)	38.92° (1.98°)	49.33° (2.48°)	54.96° (1.58°)

表 5.4 提案法による単一音源のDOA推定値(RT_{60}=250[msec])

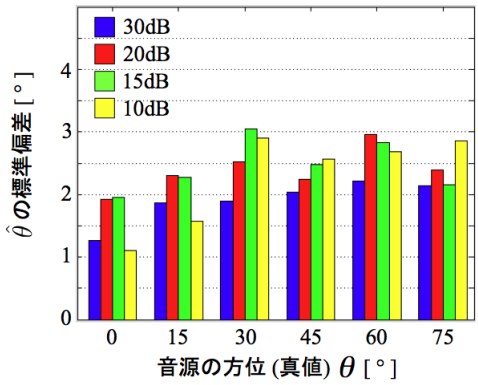
SNR	$\theta = 0°$	$\theta = 15°$	$\theta = 30°$	$\theta = 45°$	$\theta = 60°$	$\theta = 75°$
30[dB]	0.15° (1.16°)	14.78° (2.42°)	28.23° (2.70°)	40.01° (2.86°)	48.58° (3.21°)	54.53° (3.00°)
20[dB]	0.11° (1.99°)	14.63° (2.71°)	28.56° (3.08°)	39.74° (2.90°)	49.02° (3.37°)	54.70° (3.32°)
15[dB]	0.17° (2.11°)	15.07° (3.32°)	28.99° (3.61°)	39.78° (3.11°)	49.66° (3.34°)	56.07° (2.61°)
10[dB]	0.51° (1.99°)	15.19° (1.86°)	29.14° (1.52°)	39.66° (2.13°)	48.07° (2.40°)	56.50° (1.59°)

表 5.5 提案法による単一音源のDOA推定値(RT_{60}=300[msec])

SNR	$\theta = 0°$	$\theta = 15°$	$\theta = 30°$	$\theta = 45°$	$\theta = 60°$	$\theta = 75°$
30[dB]	0.11° (1.23°)	14.64° (2.58°)	28.06° (3.00°)	39.64° (2.86°)	48.69° (3.42°)	54.43° (3.17°)
20[dB]	0.16° (1.84°)	14.80° (2.71°)	28.36° (2.97°)	39.90° (3.32°)	49.10° (3.30°)	55.04° (3.27°)
15[dB]	0.40° (1.98°)	15.39° (3.13°)	28.99° (3.17°)	39.71° (2.79°)	49.67° (3.54°)	55.36° (2.81°)
10[dB]	-0.02° (1.96°)	15.56° (1.79°)	30.02° (2.63°)	39.84° (1.38°)	47.04° (3.49°)	56.09° (2.23°)

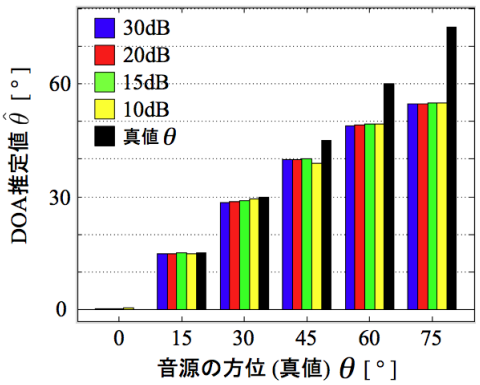


(a) DOA推定値の平均値

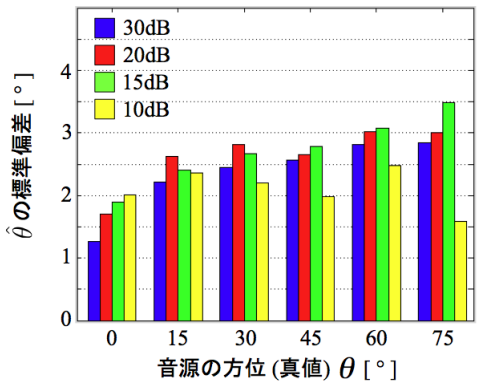


(b) DOA推定値の標準偏差

図 5.2 提案法による単一音源のDOA推定値の棒グラフ(RT_{60}=150[msec])

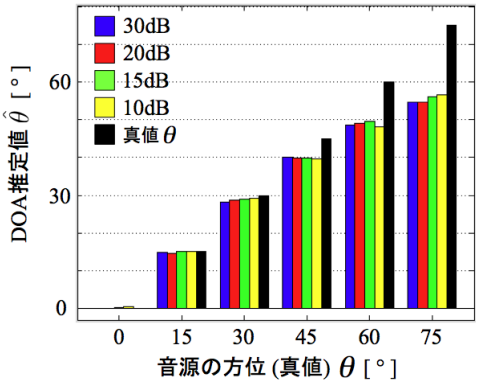


(a) DOA推定値の平均値

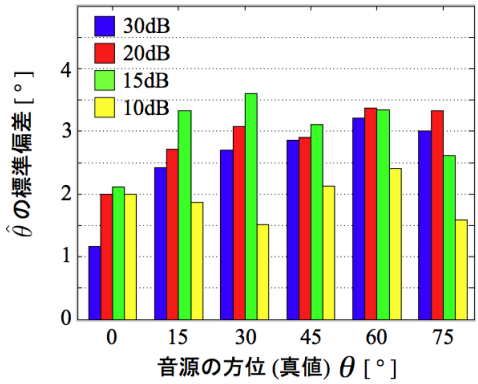


(b) DOA推定値の標準偏差

図 5.3 提案法による単一音源のDOA推定値の棒グラフ(RT_{60}=200[msec])

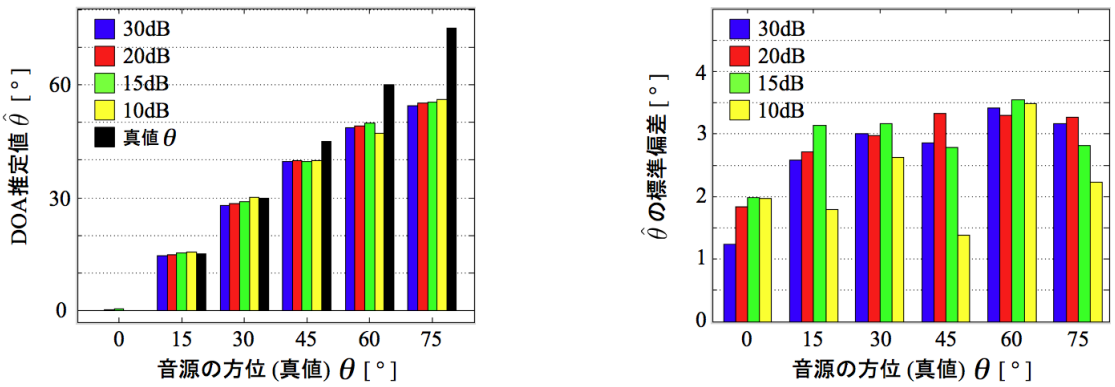


(a) DOA推定値の平均値



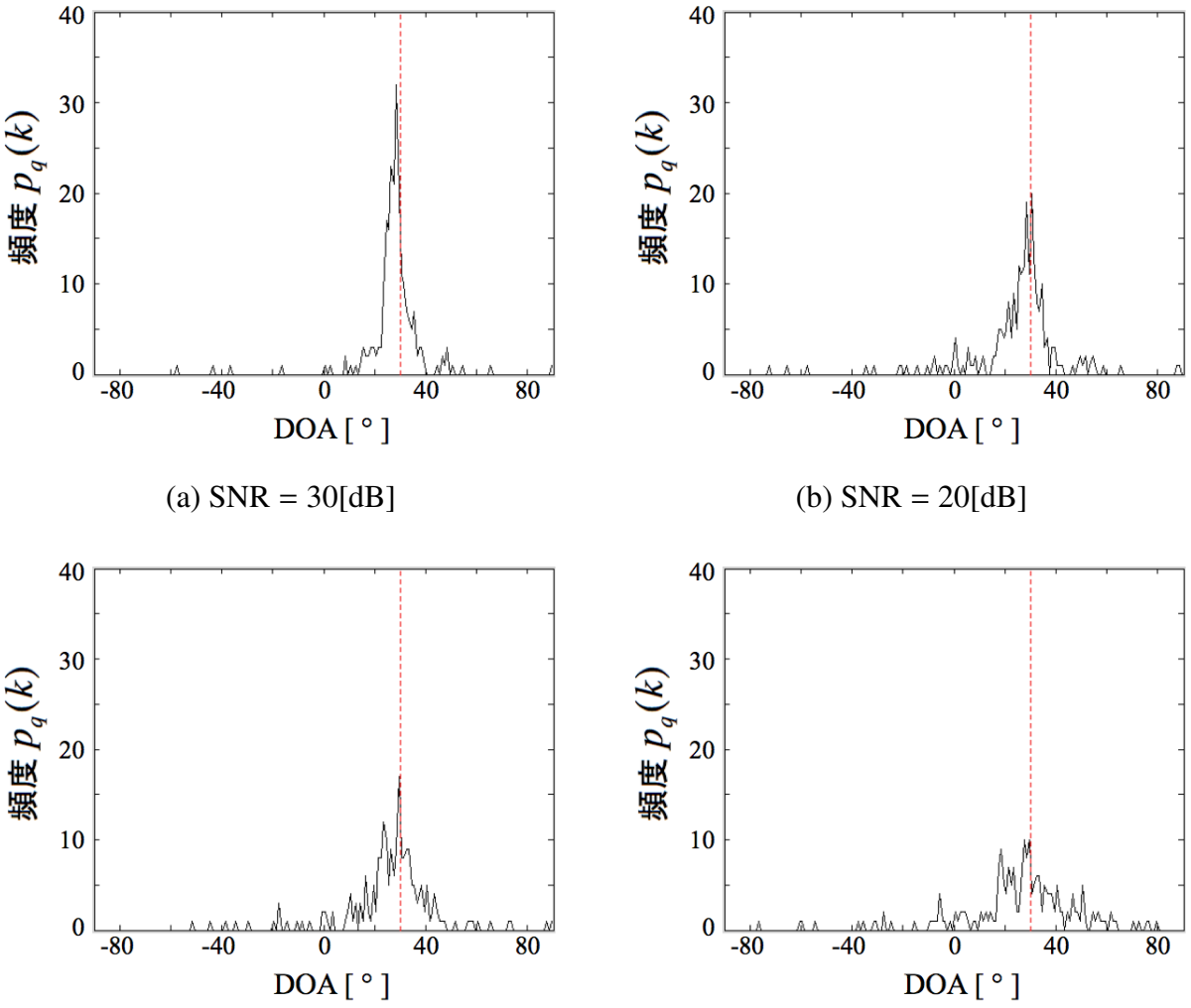
(b) DOA推定値の標準偏差

図 5.4 提案法による単一音源のDOA推定値の棒グラフ(RT_{60}=250[msec])



(a) DOA推定値の平均値　　(b) DOA推定値の標準偏差

図 5.5 提案法による単一音源のDOA推定値の棒グラフ(RT_{60}=300[msec])



(a) SNR = 30[dB]　　(b) SNR = 20[dB]

(c) SNR = 15[dB]　　(d) SNR = 10[dB]

図 5.6 SNRに対する局所DOA頻度分布の変化(RT_{60}=200[msec], $\theta = 30°$)

表 5.6 1 音源フレームの検出個数(単一音源)

RT_{60}	SNR=30[dB]	SNR=20[dB]	SNR=15[dB]	SNR=10[dB]
150[msec]	200.86 (100%)	71.11 (100%)	25.00 (100%)	9.35 (77.77%)
200[msec]	188.55 (100%)	59.36 (100%)	20.16 (100%)	8.23 (72.22%)
250[msec]	172.44 (100%)	52.83 (100%)	17.86 (100%)	7.90 (61.11%)
300[msec]	165.08 (100%)	47.44 (100%)	16.48 (97.22%)	8.20 (55.55%)

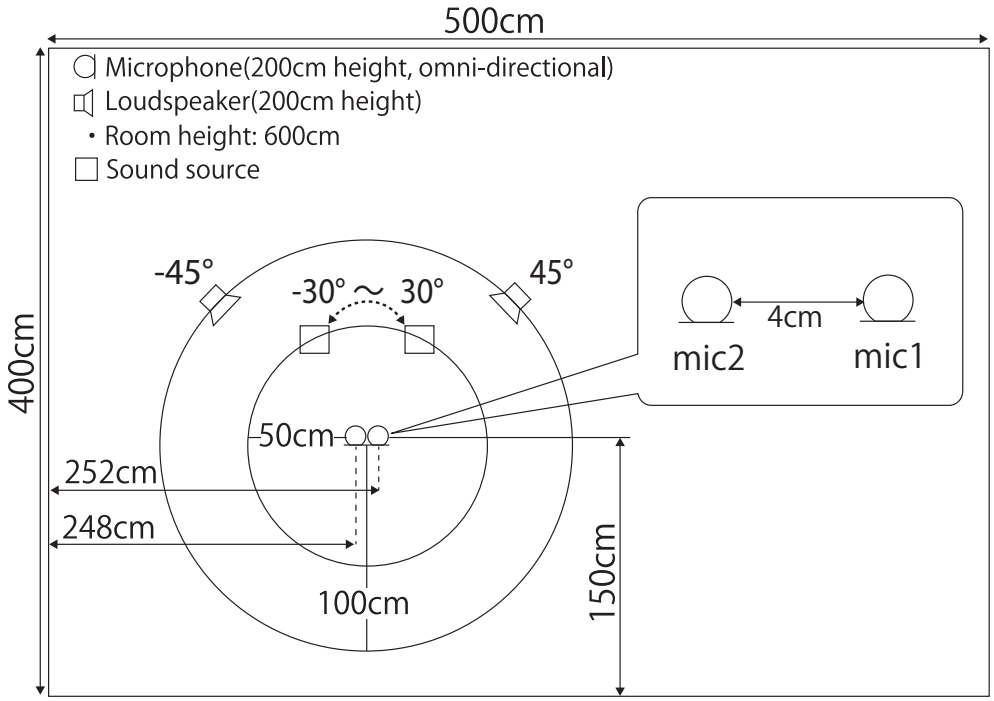


図 5.7 シミュレーション環境(3 音源)

5.2 提案法による３音源環境でのDOA推定

３音源を配置した図5.7のシミュレーション環境下で，残響時間をRT_{60}=150[msec], 200[msec]，250[msec]として，SNRをそれぞれ30[dB]，20[dB]，15[dB]と変えてシミュレーションを行った．具体的には，$s_1^*(t)$を目的音源，$s_2^*(t)$と$s_3^*(t)$を妨害音源として，妨害音源がマイクロホン対の中心から1m離れて$\theta_2 = -45°$および$\theta_3 = 45°$の方向から流れるもとで，50cm離れた目的音源の方向θ_1をフレーム単位に推定した．その際，前述と同じ男女各３発話[37]を目的音源や妨害音源のソース音声として，120回の試行を行った．このとき，RT_{60}を150[msec]として，提案法により３音源下でDOA推定を行ったときの結果を図5.8に示す．なお，この場合に限って，３つの音源以外の暗騒音は無いものとしてシミュレーションを行った．図の縦軸はDOA[°]，横軸はフレーム番号(k)，赤の実線はDOAの真値(θ = -45°，0°，45°)，青点はDOA推定値である．図5.8に示すように，１音源フレームが検出されたとしても，それが３つのうちどの音源に依るものか判らない．そこで，$|\hat{\theta}| < 35°$となる$\hat{\theta}$を目的音源$s_1^*(t)$のDOA推定値$\hat{\theta}_1$とした．複数音源下における提案法の流れを図5.9に示す．

このときのDOA推定値の平均と標準偏差を表5.7，表5.8，表5.9に示す．また，これらの

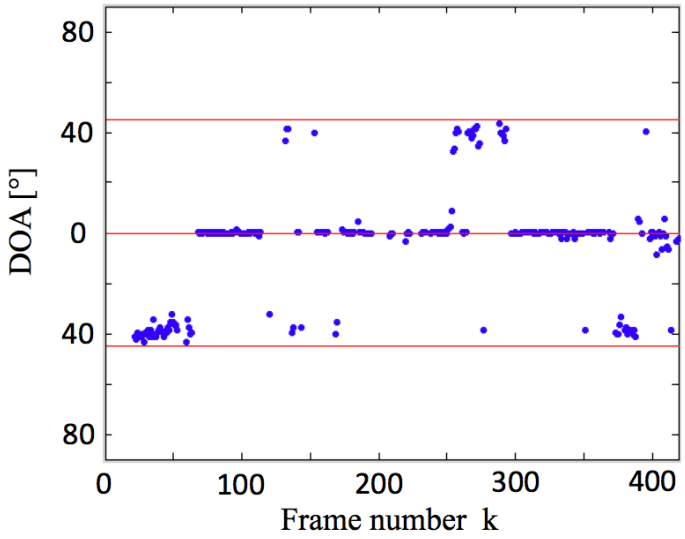


図 5.8 提案法による 3 音源環境でのDOA推定の例(RT$_{60}$=150[msec], θ = -45°, 0°, 45°)

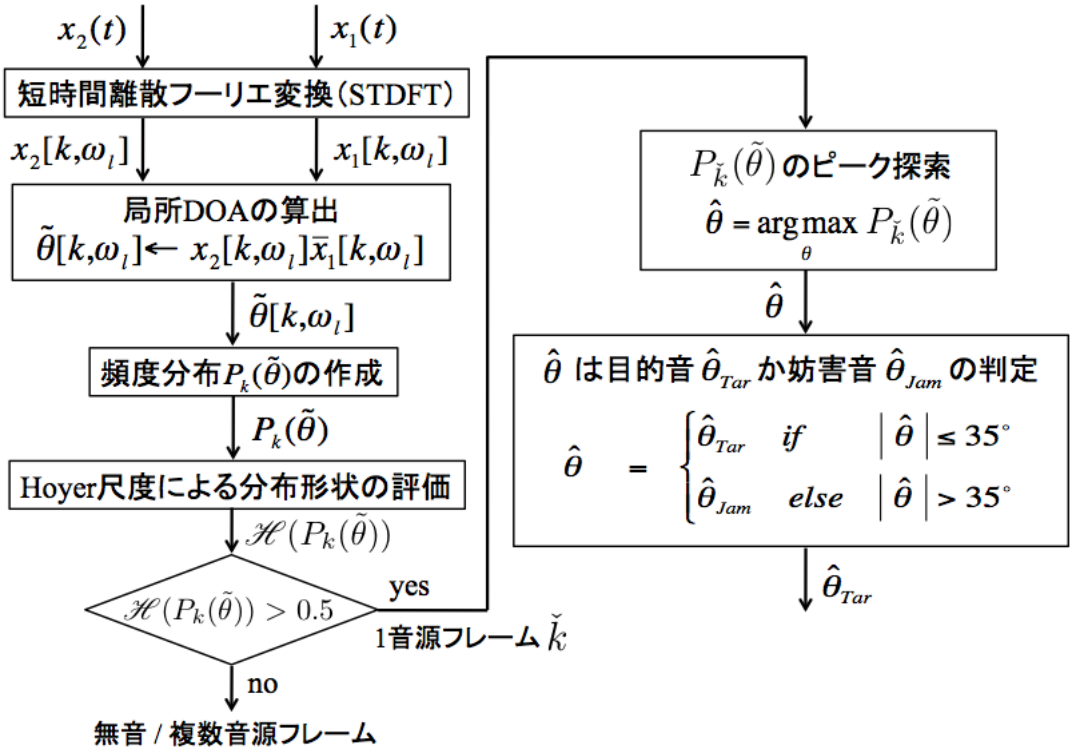


図 5.9 複数音源環境における提案法の流れ

値を棒グラフ化したものを図5.10，図5.11，図5.12に示す．これらの結果から，θ_1が±30°の範囲で，RT$_{60}$≤250[msec]，SNR≥15[dB]のとき，推定誤差は概ね2°未満，標準偏差は

表 5.7 提案法による 3 音源環境での DOA 推定値(RT_{60}=150[msec])

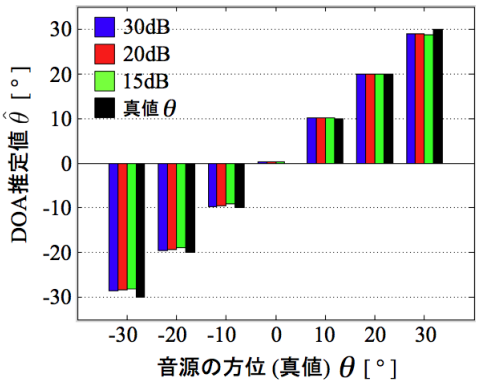
SNR	θ =-30°	θ =-20°	θ =-10°	θ =0°	θ =10°	θ =20°	θ =30°
30[dB]	-28.51° (2.17°)	-19.51° (2.52°)	-9.68° (2.18°)	0.21° (1.41°)	10.16° (2.19°)	20.03° (2.54°)	28.95° (2.11°)
20[dB]	-28.25° (2.44°)	-19.32° (2.51°)	-9.54° (2.27°)	0.29° (1.59°)	10.11° (2.31°)	19.92° (2.48°)	28.91° (2.22°)
15[dB]	-28.22° (2.40°)	-19.00° (2.46°)	-9.19° (2.43°)	0.22° (1.91°)	10.06° (2.26°)	19.95° (2.42°)	28.81° (2.25°)

表 5.8 提案法による 3 音源環境での DOA 推定値(RT_{60}=200[msec])

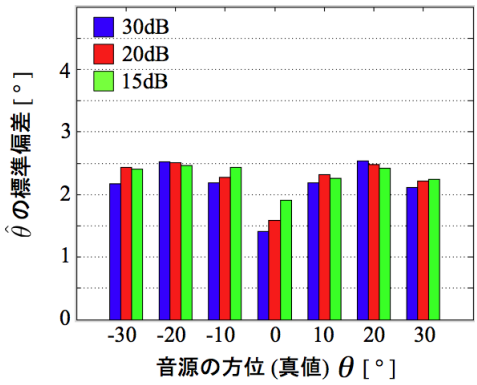
SNR	θ =-30°	θ =-20°	θ =-10°	θ =0°	θ =10°	θ =20°	θ =30°
30[dB]	-28.03° (2.82°)	-19.26° (3.28°)	-9.51° (2.71°)	0.25° (1.74°)	10.11° (2.73°)	19.91° (3.24°)	28.64° (2.93°)
20[dB]	-27.91° (2.91°)	-19.09° (3.00°)	-9.40° (2.79°)	0.32° (1.62°)	10.17° (2.72°)	19.92° (2.99°)	28.78° (2.59°)
15[dB]	-27.84° (2.57°)	-19.36° (2.89°)	-9.28° (2.63°)	0.36° (1.82°)	10.61° (2.47°)	20.12° (2.67°)	28.64° (2.60°)

表 5.9 提案法による 3 音源環境での DOA 推定値(RT_{60}=250[msec])

SNR	θ =-30°	θ =-20°	θ =-10°	θ =0°	θ =10°	θ =20°	θ =30°
30[dB]	-27.60° (3.11°)	-18.92° (3.32°)	-9.36° (2.94°)	0.21° (1.65°)	10.12° (2.84°)	19.77° (3.60°)	28.42° (3.06°)
20[dB]	-27.50° (3.11°)	-18.88° (3.13°)	-9.19° (3.25°)	0.27° (1.64°)	10.06° (3.11°)	19.78° (3.38°)	28.25° (3.11°)
15[dB]	-27.59° (2.67°)	-18.72° (2.88°)	-9.27° (2.94°)	0.02° (2.00°)	10.20° (2.37°)	19.83° (2.47°)	28.78° (2.41°)

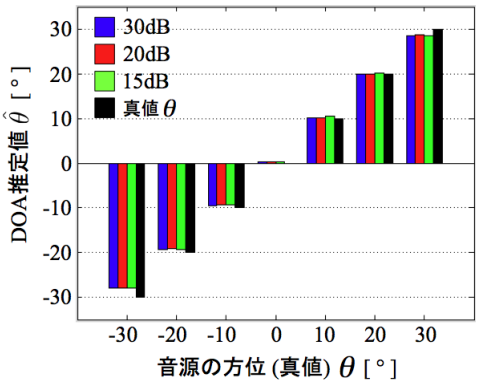


(a) DOA推定値の平均値

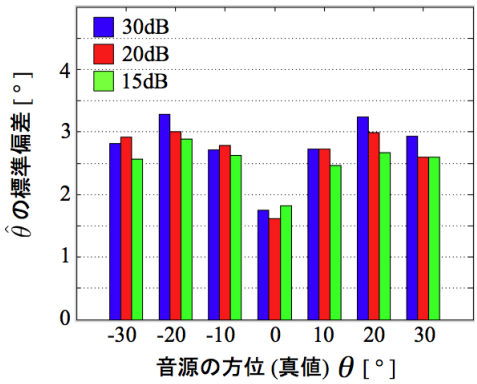


(b) DOA推定値の標準偏差

図 5.10 提案法による 3 音源環境でのDOA推定値の棒グラフ(RT$_{60}$=150[msec])

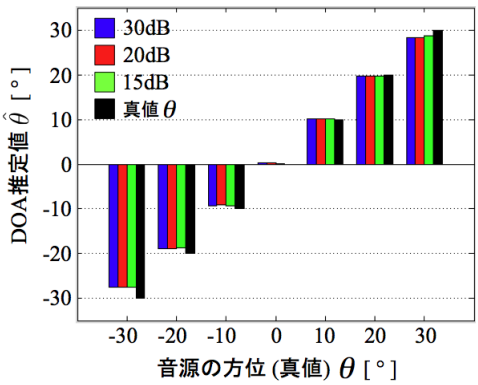


(a) DOA推定値の平均値

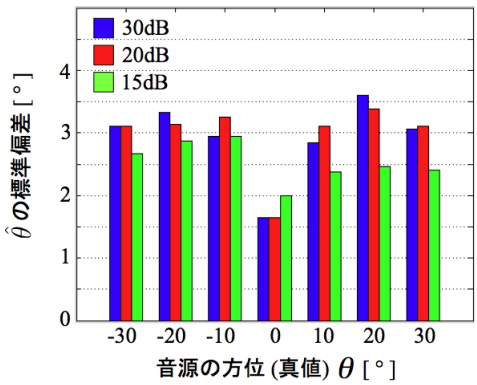


(b) DOA推定値の標準偏差

図 5.11 提案法による 3 音源環境でのDOA推定値の棒グラフ(RT$_{60}$=200[msec])



(a) DOA推定値の平均値



(b) DOA推定値の標準偏差

図 5.12 提案法による 3 音源環境でのDOA推定値の棒グラフ(RT$_{60}$=250[msec])

概ね3°未満に収まることが見てとれる．さらに，１音源を対象にしたときの表5.2~表5.5の結果との比較から，妨害音源の有無にかかわらず，ほぼ同等のDOA推定精度が得られていることがわかる．このことから，提案法はθ_1が±30°の範囲に限られるが，１音源フレームさえ検出できれば，環境だけではなく音源数にも依存しない高精度なDOA推定が可能なことを示唆している．しかし，表5.7，表5.8，表5.9を細かく見てみると，残響時間が短いときやSNRが高いときの方が，わずかではあるが，推定誤差は大きく，標準偏差は小さい場合がある．具体的には，表5.8において，θ=10°のときのSNR=15[dB]のDOA推定値は10.61°で標準偏差は2.47°であるが，表5.9において同様の条件のときの結果をみてみると，DOA推定値は10.20°で標準偏差は2.37°である．この場合，残響時間RT_{60}が200[msec]から250[msec]と長くなったにもかからわず，推定誤差と標準偏差はともに小さい．また，SNRが低くなったにもかかわらず，推定誤差と標準偏差が小さくなる例が散見される．以上と同様のことが，図5.10，図5.11，図5.12からも確認できる．

上述の逆転現象も，１音源フレームの検出個数の違いに起因していると考えられる．各RT_{60}とSNRで，１音源フレームが検出された個数を表5.10に示す．表5.10の数値は，120試行における平均検出個数で,（）内の数値は１音源フレームが検出できた試行の割合である．表5.10より，表5.6と同様に，１音源フレームの検出個数は残響時間が長くSNRが低いほど検出される割合が減少することがわかる．具体的には，残響時間RT_{60}を150[msec]から250[msec]とした場合，SNR=30[dB]のとき25%減，SNR=20[dB]のとき28%減，SNR=15[dB]のとき25%減となり，１音源フレームの検出個数は全体的に26%減少した．また，SNRを30[dB]から15[dB]とした場合，残響時間RT_{60}=150[msec]のとき81%減，残響時間RT_{60}=200[msec]のとき82%減，残響時間RT_{60}=250[msec]のとき81%減となり，１音源フレームの検出個数は全体的に81%減少した．そして，残響時間RT_{60}=250[msec]でSNR=15[dB]のとき，１音源フレームが検出できないケースが7%見られた．さらに，前節と同様に，SN比を10[dB]とした場合や残響時間をRT_{60}=300[msec]とした場合，１音源フレームの検出されないケースが多発した．以上のことから，提案法は，図５の環境の場合，残響時間が200[msec]以下，SN比が15[dB]以上で有効と判断される．

表 5.10 1音源フレームの検出個数(3音源環境)

RT_{60}	SNR=30[dB]	SNR=20[dB]	SNR=15[dB]
150[msec]	81.76 (100%)	34.32 (100%)	15.26 (100%)
200[msec]	71.11 (100%)	28.78 (100%)	13.00 (100%)
250[msec]	61.27 (100%)	24.79 (100%)	11.52 (92.97%)

5.3 まとめ

本章では，提案法の有効性をシミュレーションにより検証した．まず，残響時間RT_{60}が200[msec]，SNRが20[dB]の単一音源環境でシミュレーションを行って，提案法はDUET法に基づく方法とほぼ同等の精度でDOA推定できることを確認した．具体的には，提案法は，目的音源の方位θをブロードサイド方位から±30°に絞り込めば，推定誤差は2°未満，標準偏差は3°程度の精度でフレーム単位にDOA推定できることを確認した．次に，残響時間RT_{60}とSNRを変えて，提案法が有効に機能する範囲をシミュレーションにより検討した．その結果，提案法は，１音源フレームさえ検出できれば，残響時間RT_{60}やSNRに依存することなく，DOAを高精度に推定できることを示唆しているが，残響時間$_{60}$が長くSNRが低いほど，１音源フレームの検出個数は減少することを指摘した．したがって，提案法を有効に機能させるには，残響時間RT_{60}が250[msec] 以下でSNRが15[dB]以上の環境で使用する必要があることを明らかにした．そこで，残響時間RT_{60}が250[msec] 以下でSNRが15[dB]以上とし，目的音源の方位を±30°の範囲に絞り込み，３音源が存在する環境で，提案法により目的音源のDOAを推定するシミュレーションを行った．その結果，残響時間が200[msec]以下でSNRが15[dB]以上あれば，目的音源のDOA推定値の誤差が3°未満，標準偏差が3°程度の精度で推定できることを確認した．

第6章

結論

本論文では，音源のスパース性を利用して，２つのマイクロホンで観測された混合信号から，目的音源のDOAをフレーム単位に推定する方法を提案した．

第１章では，DOA推定に関する研究目的と背景について簡単に概説した．すなわち，まず，既存の代表的な雑音除去法をタイプ別に概観し，DOAを利用した場合，その雑音除去能力はDOAの推定精度に大きく左右されることを説明した．次に，代表的なDOA推定法であるDSB法，MUSIC法，スパース性に基づく方法の特色を対比的に説明し，スパース性に基づく方法の優位性と課題を明らかにした．具体的には，スパース性に基づく方法は，マイクロホン数が音源数より少ない場合でもDOA推定できるが，クラスタリング等のバッチ処理を必要とするため，突発音などの瞬間的な音や，移動音源のDOAをリアルタイムに推定することは難しいことを指摘した．

以上の背景をもとに，本研究では，DOAをリアルタイムに推定することの意義を明らかにして，雑音除去の中で研究を位置づけた．

第２章では，音源のDOA推定について概説した．具体的には，DSB法，MUSIC法，スパース性に基づいたDUET法の基本原理を説明し，各方法の長所や短所を明らかにして，実際への適用に関しての課題を述べた．

DSB法は，試行方位θ(-90° ~ 90°)に対するステアリング・ベクトルを生成し，出力されたアレイ入力が最大となるθを探索することで音源のDOAを推定するという原理になっていることを説明した．しかし，この場合，DOAは比較的簡単に推定できるが，複数の音源が同時にマイクロホンに入射するときは機能せず，空間分解能が低いことを述べた．

MUSIC法は，信号部分空間を張る到来方向ベクトルと，雑音部分空間を張る固有ベクトルが直交する性質を利用して，MUSICスペクトルのピークから音源のDOAを推定する方法であることを説明した．この方法は，複数の音源が同時にマイクロホンに入射する環境下でも，音源の個数よりマイクロホンの個数が多ければ，個々の音源のDOAは推定できて，空間分解能も高いことを述べた．しかし，固有値や固有ベクトルを求める必要があるため，DSB法より計算量は多く，その計算量はマイクロホンの個数とともに増えることや，信号数が既知でなければ適用できないことを述べた．

音源のスパース性に基づいたDUET法は，クラスタリングにより局所DOAのヒストグラムを作成し，そのピークから個々の音源のDOAを推定する方法であることを説明した．この方法は，マイクロホンの個数は音源の個数より少なくてもDOAが推定できるという特色があり，無響室下でクラスタリングが良好に行えれば，信号源数が未知の場合でも有効に機能することを述べた．しかし，クラスタリング等のバッチ処理を必要とするため，突発音などの瞬間的な音や，移動音源のDOAを推定することは難しいことを指摘した．以上のことから，突発音などの瞬間的な音や，移動音源への適用を行うには，フレーム単位でのDOA推定が必須であることを明らかにした．

第 3 章では，音声は音声区間と無音区間を繰り返す断続波であり，音声のDOAに関する情報は，音声区間にだけ含まれ，無音区間には含まれないが，現実には無関係な雑音成分が無音区間に重畳することを指摘した．したがって，フレーム数が少ない場合，無音区間が雑音や残響の影響を受けることから，音声区間内の連続する20フレームでの局所DOAをもとに，DUET法に基づく方法とMUSIC法によりDOA推定した場合，どの程度の推定精度が得られるかをシミュレーションにより調べた．

DUET法に基づく方法により，音声区間の連続する20フレームを用いてDOA推定した場合，対象音源のDOAの真値(θ)が30°以内ならば，推定誤差は3°未満に収まるが，θ=45°を越えると7°以上となって推定精度は急激に劣化する結果となった．このことは，全フレームを用いた場合でも同様であった．そこで，θが0°から外れるにつれて推定誤差が大きくなる原因について考察し，真値(θ)の0°からの変位に比べて，局所DOA頻度分布のピークの変位は小さいため，結果として得られるDOA推定値は大きな誤差をもつことを明らかにした．ただし，標準偏差は全体的に3°未満に収まり，全フレームを用いた場合と音声区間の20フレームを用いた場合とで相違ない結果が得られた．

一方，MUSIC法により，音声区間の連続する20フレームを用いた場合，DOA推定誤差は$\theta \leq 60°$までは3°程度に収まるが，標準偏差は20°から30°とかなり大きくなることが判明した．この場合，全フレームを用いたり低周波数ビンを除外すればある程度は改善されるが，それでもDUET法に基づくDOA推定値の標準偏差を大きく超えていた．そこで，その原因について考察し，MUSIC法による場合，推定値の分散はCramer-Rao Boundに従うことを確認した．すなわち，MUSIC法の場合，真値θがブロードサイド方向から外れるほど，またSN比が低くデータ数やマイクロホン数が少ないほど，推定値がバラツクことから，フレーム数が少ないときのDOA推定には不向きであることを指摘した．一方，DUET法に基づくDOA推定の場合，推定値のバラツキはフレーム数が少なくてもあまり変わらず，目的音源の方位を$-30° \sim 30°$の範囲に絞り込めば，MUSIC法より良好にDOAを推定できることを述べた．

第４章では，DUET法に基づくフレーム単位のDOA推定法を提案した．具体的には，まず，マイクロホンで観測された混合信号は，無音源区間，１音源区間，複数音源区間に分けられることを言及した．次に，混合信号を短時間離散フーリエ変換して得られる複素スペクトルの位相差をもとに，フレーム単位に時間周波数における局所DOAを求めて，その頻度分布をとれば３種類の形状に分類されることを述べた．すなわち，１音源フレームでは１つのピークをもつ単峰的な分布，２音源フレームでは２つのピークをもつ双峰的な分布，無音源フレームではピークのない比較的平坦な分布となることを述べた．そして，１音源フレームではピークが明確な単峰的な分布となるため，１音源フレームが選別できればその分布のピークを探索することでDOA推定ができることを指摘した．以上のことから，フレーム単位のDOA推定を行うためには，１音源フレームの選別が重要であることを明らかにした．分布の形状を測る尺度として，Hoyer尺度などのスパース尺度があり，適切な閾値を設けることで１音源フレームを選別できることをシミュレーションにより確認した．また，このとき選別された１音源フレームのピークは音源の方位にほぼ一致した．最後に，Hoyer尺度による１音源フレームの検出結果と実際の音声区間を比較した結果，Hoyer尺度の閾値を0.5とすれば選別されたフレームは全て音声区間内に収まっており，閾値の妥当性を立証した．

第５章では，提案法の有効性をシミュレーションにより検証した．まず，残響時間が200[msec]，SN比が20[dB]の環境下でシミュレーションを行って，提案法はDUET法とほ

ぼ同等の精度でDOA推定できることを確認した．具体的には，提案法は，目的音源の方位θをブロードサイド方位から±30°の範囲に絞り込めば，推定誤差は2°未満，標準偏差は3°程度の精度でフレーム単位にDOA推定できることを確認した．次に，残響時間とSN比を変えて，提案法が有効に機能する範囲をシミュレーションにより検討した．その結果，提案法は，１音源フレームさえ検出できれば，残響時間やSN比に依存することなく，DOAを高精度に推定できることを示唆しているが，残響時間が長くSN比が低いほど，１音源フレームの検出個数は減少することを指摘した．以上の結果から，提案法を有効に機能させるには，残響時間が250[msec] 以下でSN比が15[dB]以上の環境で使用する必要があることを明らかにした．そこで，残響時間が250[msec] 以下でSN比が15[dB]以上とし，目的音源の方位を±30°の範囲に絞り込み，３音源が存在する環境で，提案法により目的音源のDOAを推定するシミュレーションを行った．その結果，残響時間が200[msec]以下でSN比が15[dB]以上あれば，目的音源のDOA推定値の誤差が3°未満，標準偏差が3°程度の精度で推定できることを確認した．

以上のように，音源（話者）のDOAが±30°の範囲であれば，提案法によるDOA推定は有効に機能する．したがって，方位可動なマイクロホン対を利用すれば，DOA推定値を方位可動機構にフィードバックすることにより，音源（話者）方向を逐次追跡できる．この観点から，マイクロホン対を頭部筐体に組み込んだ対話ロボットへの応用が考えられる．また，音声認識では音声区間を正確に検出することが重要である．Hoyer尺度による１音源フレームの選別法は，音声区間の検出への応用へも期待できる．

提案法については，上記への応用を念頭に，DOA推定精度と分析フレーム長の関係や複数音源のDOA推定等に関して，今後，検討を進める予定である．

謝辞

平成22年に九州工業大学大学院情報工学府に入学し，本研究に着手して以来，今日に至るまで本学大学院情報工学府の井上勝裕教授には多くの御指導と御意見だけではなく，私が研究に取り組み易いよう多大なる御配慮まで賜わりました．長年にわたり御迷惑ばかりおかけ致しましたが，最後まで辛抱強く見守って下さったことに，心より感謝申し上げます.

近畿大学産業技術研究科の五反田博教授には，平成20年に近畿大学産業技術研究科に入学して以来，論文の書き方から人としての生き方までの幅広い御指導を頂戴致しました．私が今日まで歩んでこれたのは，全て五反田博教授のおかげです．無能な私を最後まで見捨てずに優しく御支援下さったことに，心より感謝申し上げます.

本論文をまとめるにあたり，九州工業大学大学院情報工学府の尾知博教授，古賀雅伸教授，江島俊朗教授，野田秀樹教授，前田誠助教には貴重な御意見や有益な御助言を頂きました．また，拙稿の校閲や未熟な研究発表にお付き合い下さったことに，心より感謝申し上げます．近畿大学産業技術研究科の白土浩准教授，熊本高等専門学校の石橋孝昭准教授，古屋武志博士には，研究環境の整備や研究遂行において様々な御支援や御指導を頂きました．ここに，心より感謝申し上げます．井上研究室と五反田研究室の皆様には，今日まで大変お世話になりました．皆様との出会い，一緒に過ごした日々は私の宝物であり，今後の努力の糧になるものであります．ここに，心より感謝申し上げます.

最後に，これまで自分の選んだ道を進むことに対し，温かく見守り，そして辛抱強く支援して下さった両親に心より感謝の意を表して謝辞と致します。

参考文献

[1] S.F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoustics , Speech and Signal Processing, vol.ASSP-27, no.7, pp.113–120, 1979.

[2] 岡崎雅嗣，国本利文，小林隆夫，“多段スペクトルサブトラクション法を用いた楽音の強調,” 電子情報通信学会論文誌，vol.J88-D-2，no.12，pp.2301–2310，Dec. 2005.

[3] O.L. Frost, “An algorithm for linearly constrained adaptive array processing,” Processing of IEEE, vol.60, no.8, pp.926–934, 1972.

[4] L.J. Griffiths and C.W. Jim, “An alternative approach to linear constrained adaptive beamforming,” IEEE Trans. AP, vol.AP-30, no.1, pp.27–34, Jan. 1982.

[5] Y. Kaneda and J. Ohga, “Adaptive microphone-array system for noise reduction,” IEEE Trans. ASSP, vol.34, no.6, pp.1391–1400, Dec. 1986.

[6] 大賀寿郎，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.

[7] O. Hoshuyama, A. Sugiyama, and A. Hirano, “A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters,” IEEE Trans. SP, vol.47, no.10, pp.2677–2684, 1999.

[8] J. Mayer and G.W. Elko, “A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield,” IEEE ICASSP 2002, pp.1781–1784, 2002.

[9] W. Herbordt, Sound Capture for Human/Machine Interfaces - Practical Aspects of Microphone Array Signal Processing, Springer, March 2005.

[10] C.H. Knapp and G.C. Carter, “The generalized correlation method for estimation of time delays,” IEEE Trans. Acoust. Speech Signal Process., vol.24, no.4, pp.320–327, 1976.

[11] M. Omologo and P. Svaizer, “Use of the crosspower-spectrum phase in acoustic event location,” IEEE Trans. on Speech and Audio Processing, vol.SAP-5, no.3, pp.288–292, 1997.

[12] 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-II，no.8，pp.1713–1721，2000.

[13] 浅野　太，音のアレイ信号処理ー音源の定位・追跡と分離，コロナ社，2011.

[14] R.O. Schmidt, “Multiple emitter location and signal parameter estimation,” IEEE Trans. Antennas Propag., vol.34, no.3, pp.276–280, 1986.

[15] B.D. Rao and K.V.S. Hari, “Performance analysis of root-music,” IEEE Trans. Acoust. Speech Signal Process, vol.37, no.12, pp.1939–1949, Dec. 1989.

[16] R. Roy and T. Kailath, “Esprit—estimation of signal parameters via rotational invariance techniques,” IEEE Trans. Acoust. Speech Signal Process., vol.37, no.7, pp.984–995, 1989.

[17] 菊間信良，アダプティブアンテナ技術，オーム社，2003.

[18] 片岡章俊，小林和則，日岡裕輔，“遠隔地との音声通信におけるマイクロホンアレー収音技術,” 電子情報通信学会 論文誌，vol.92，no.2，pp.118–124，2009.

[19] S. Araki, H. Sawada, R. Mukai, and S. Makino, “Doa estimation for multiple sparse source with arbitrarily arranged multiple sensors,” Journal of Signal Processing Systems, vol.63, no.3, pp.265–275, Oct. 2009.

[20] N. DING and N. HAMADA, “Doa estimation of multiple speech sources from a stereophonic mixture in underdetermined case,” IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences, vol.E95-A, no.4, pp.735–744, April 2012.

[21] L. Vielva, D. Erdogmus, C. Pantaleon, I. Santamaria, and J.C. Principe, “Underdetermined blind source separation in a time-varing enviroment,” Acoustics, Speech, and Signal Processing (ICASSP), 2002. Proceedings IEEE International Conference on, vol.III, pp.3049–3052, May 2002.

[22] A. Blin, S. Araki, and S. Makino, “A sparseness-mixing matrix estimation (smme) solving the underdetermined bss for convolutive mixtures,” Acoustics, Speech, and Signal Processing (ICASSP), 2004. Proceedings. IEEE International Conference on, vol.IV, pp.85–88, May 2004.

[23] S. Winter, H. Sawada, and S. Makino, “On real and complex valued l1-norm minimization for overcomplete blind source separation,” Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on, pp.86–89, Oct. 2005.

[24] C. Fevotte and S.J. Godsill, “A bayesian approach for blind separation of sparse sources,” Audio, Speech, and Language Processing, IEEE Transactions on, vol.14, no.6, pp.2174–2188, Nov. 2006.

[25] Y. Izumi, N. Ono, and S. Sagayama, “Sparseness-based 2ch bss using the em algorithm in reverberant environment,” Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on, pp.147–150, Oct. 2007.

[26] P.D.O. Grady and B.A. Pearlmutter, “The lost algorithm: Finding lines and separating speech mixtures,” EURASIP Journal on Advances in Signal Processing 2008, p.Article ID 2008: 784296, July 2008.

[27] M.I. Mandel, R.J. Weiss, and D.P.W. Ellis, “Model-based expectation maximization source separation and localization,” IEEE Transactions on Audio, Speech, and Language Processing, vol.18, no.2, pp.382–394, Feb. 2010.

[28] H. Sawada, S. Araki, and S. Makino, “Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment,” Audio, Speech, and Language Processing, IEEE Transactions on, vol.19, no.3, pp.516–527, March 2011.

[29] A. Jourjine, S. Rickard, and Ö. Yilmaz, “Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures,” in Proc. ICASSP2000, vol.5, pp.2985–2988, 2000.

[30] Ö. Yilmaz and S.Rickard, “Blind separation of speech mixtures via time-frequency masking,” IEEE Trans on. Signal Processing, vol.52, no.7, pp.1830–1847, 2004.

[31] S. Rickard and F. Dietrich, “Doa estimation of many w-disjoint orthogonal sources from two mixtures using duet,” Proceedings of the 10th IEEE Workshop on Statistical Signal and Array Processing (SSAP2000), pp.311–314, Aug. 2000.

[32] M. Matsuo, Y. Hioka, and N. Hamada, “Estimating doa of multiple speech signals by improved histogram mapping method,” in Proc. IWAENC2005, pp.129–132, 2005.

[33] M. Kuhne, R. Togneri, and S. Nordholm, “A novel fuzzy clustering algorithm using observation weighting and context information for reverberant blind speech separation,” Signal Process., vol.90, pp.653–669, 2010.

[34] S. Araki, H. Sawada, R. Mukai, and S. Makino, “Doa estimation for multiple sparse source with normalized observation vector clustering,” Proc. ICASSP2006, vol.V, pp.33–36, 2006.

[35] S. Araki, H. Sawada, R. Mukai, and S. Makino, “Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors,” Signal Processing, vol.87, no.8, pp.1833–1847, 2007.

[36] S. Rickard, The DUET Blind Source Separation Algorithm, T.W.L. S. Makino and H. Sawada, eds., Springer, Dordrecht, 2007.

[37] 板橋秀一, “音声情報処理研究用日本語音声データベース,” 1991.

[38] E.A.P. Habets, “Single- and multi-microphone speech dereverberation using spectral enhancement,” PhD thesis, Technische Universiteit Eindhoven, The Netherlands, 2007.

[39] R.I.R.G. forMatlab. http://home.tiscali.nl/ehabets/rir_generator.html.

[40] C.T. Ishi, O. Chatot, H. Ishiguro, and N. Hagita, “Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments,” in Proc. of the 2009 IEEE/RSJ Intl. Conf. on Intelligent Robots and System, pp.2027–2032, 2009.

[41] P. Stoica and A. Nehorai, “Music, maximum likelihood, and cramer-rao bound,” IEEE Trans on. Signal Processing, vol.37, no.5, pp.720–741, 1989.

[42] J. Li, B. Halder, P. Stoica, and M. Viberg, “Computationally efficient angle estimation for signals with known waveforms,” IEEE Trans on. Signal Processing, vol.43, no.9, pp.2154–2163, 1995.

[43] J. Li and R.T.C. Jr., “Maximum likelihood angle estimation for signals with known waveforms,” IEEE Trans on. Signal Processing, vol.41, no.9, pp.2850–2862, 1993.

[44] A. Leshem and A.J. van derVeen, “Direction-of-arrival estimation for constant modulus signals,” IEEE Trans on. Signal Processing, vol.47, no.11, pp.3125–3129, 1999.

[45] 板橋秀一，音声工学，森北出版，2005.

[46] N. Hurley and S. Rickard, “Comparing measures of sparsity,” IEEE Trans on. Information Theory, vol.55, no.10, pp.4723–4741, 2009.

[47] A. Hyvärinen and E. Oja, “Independent component analysis: algorithms and applications,” Neural Networks, vol.13, no.4-5, pp.411–430, 2000.

[48] N. Iwasaki, T. Matsuzaki, G. Hirano, H. Shiratsuchi, K. Inoue, and H. Gotanda, “Studies on real time doa estimation based on duet,” ICIC Express Letters Part B: Applications, vol.5, no.2, pp.377–386, 2014.

[49] N. Iwasaki, K. Inoue, and H. Gotanda, “A real time oriented sound source doa estimation based on sparseness,” Transactions of the Institute of Systems, Control and Information Engineers, vol.27, no.12, pp.493–500, 2014.