論　　文 43.60. −c, Jn, Gk

サブバンドピークホールド処理を用いた音源方向推定法*

鈴木　敬*1,† 金田　豊*1

[要旨] 本論文は，音源方向推定の妨害要因である反射音への対策として，サブバンドピークホールド (SBPH) 処理を用いた手法を提案する。サブバンドピークホールド処理とは，先行する直接音の振幅を保持することで，後続の反射音をマスクする処理（ピークホールド処理）をサブバンド信号に対して適用する処理である。本論文では，二つのマイクロホンで受音された信号間の時間差検出に基づく音源方向推定法を対象として SBPH 処理の有効性を検証した。虚像法を用いたシミュレーションと反射音の影響の大きい実環境での音源方向推定実験を行った結果，反射音に強いとされる従来法の PHAT 法よりも高い反射音耐性が確認された。

キーワード 室内反射音，音源方向推定，相互相関関数，マイクロホンアレー，サブバンドピークホールド
Reflected sound, DOA estimation, Cross-correlation function, Microphone array, Subband peak hold

1. はじめに

音源方向の推定は，人とロボットのコミュニケーション時にロボットが話者の方向を認識する技術 [1] や，テレビ会議システムで発言者を自動検出してカメラでクローズアップする技術 [2]，その他にも，騒音源の調査や，遠隔監視システムで異常音の発生位置を推定する技術など幅広い用途を持っている。

しかし，室内で音源方向推定の技術を用いる場合，周囲雑音や室内反射音の影響を無視することはできない。特に，屋外や室内近距離に比べ，屋内遠距離やアレーが壁際にある場合は，ある特定の方向から大きな初期反射音が到来するため，推定精度が大幅に劣化してしまう問題がある [3]。この対策として，音源方向推定の前処理として反射音を低減する手法 [4, 5] や，有声音の調波構造を利用した手法 [6] などが提案されている。しかし，これらの手法は，初期反射音の影響が大きくなる環境下において，必ずしも十分な性能は得られていない。そこで我々は，先行する直接音の振幅を保持することで，後続する反射音をマスクするピークホールド処理 [7] を，サブバンド信号に対して適用するサブバンドピークホールド処理 [8] の利用を提案する。その結果，初期反射音が大きいと予想される環境下においても，その影響を大幅に軽減できることが確認されたので報告する。

一般的に音源方向推定の代表的な方法としては，受音信号の時間差推定に基づく相互相関法 [9]，指向性ビーム走査による出力パワー推定に基づくビームフォーマ法 [10]，受音信号の空間相関行列の固有空間構造を利用した高い角度分解能を有するサブスペース法 [10] の三つが主に挙げられる。これら三つの手法はいずれも，複数のマイクロホン受音信号間の相互相関関数を利用しており，反射音の影響による相互相関関数の劣化は共通した推定誤差要因となる。そこで本論文では，三つの手法の中でも最も基本的な推定法にあたる，二つのマイクロホンを用いた相互相関法を対象として反射音耐性向上の検討を行った。

2. 音源方向推定モデル

図–1 に 2 ch マイクロホンアレーに基づく音源方向推定のモデル図を示す。このモデルでは音波を平面波と仮定している。音源の方向 θ_s から音波が到来したとき，二つのマイクロホン M1，M2 で受音された信号 $x_1(t)$，$x_2(t)$ には時間差 τ_s が生じる。この時間差 τ_s は，音波が図中の距離 ξ だけ進む時間であり，マイク間距離を d，音速を c とすれば次式のように表せる。

$$\tau_s = \xi/c = d\sin\theta_s/c \tag{1}$$

この式を，音源方向 θ_s について解くと次式のように表せる。

$$\theta_s = \sin^{-1}(c\cdot\tau_s/d) \tag{2}$$

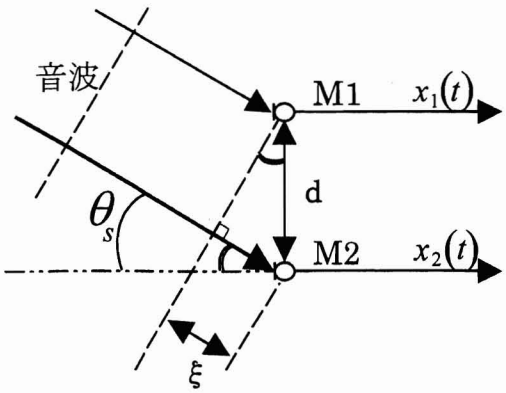
このとき，マイク間距離 d，音速 c の値は既知であるの

---

* Sound source direction estimation based on subband peak-hold processing,
by Takashi Suzuki and Yutaka Kaneda.

*1 東京電機大学大学院工学研究科情報通信工学専攻
† 現在，東京電力(株)
(問合先：金田　豊　〒101–8457 東京都千代田区神田錦町 2–2 東京電機大学大学院工学研究科情報通信工学専攻　e-mail: kaneda@c.dendai.ac.jp)

(2009年1月6日受付，2009年5月7日採録決定)



図–1 音源方向推定のためのモデル図

で，時間差 τ_s を推定すれば音源方向 θ_s が算出できる。

3. 従来の音源方向推定法

3.1 相互相関法

二つのマイクロホンを用いた音源方向推定の最も基本的な方法は，受音信号 $x_1(t)$ と $x_2(t)$ との時間差 τ_s を次式のように定義された相互相関関数 $\varphi_{12}(\tau)$ の最大値を与える τ の値として推定するものである。

$$\varphi_{12}(\tau) = E[x_1(t)x_2(t-\tau)] \tag{3}$$

ただし，$E[\cdot]$ は期待値を表す。

3.2 PHAT (Phase Transform) 法

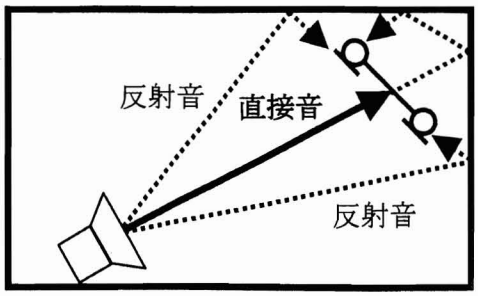
相互相関関数 $\varphi_{12}(\tau)$ は $x_1(t)$ と $x_2(t)$ のクロススペクトル $\Phi_{12}(\omega)$ を用いて，次式のように表すことができる。

$$\varphi_{12}(\tau) = \int \Phi_{12}(\omega)e^{j\omega\tau}d\omega \tag{4}$$

この時，周囲雑音や室内反射音などの環境下において性能を確保するために $\Phi_{12}(\omega)$ に様々な周波数重み $\Psi(\omega)$ を付けることが提案されている（一般化相互相関関数）[9]。その中で，クロススペクトルの振幅成分を打ち消し位相項のみを用いる方法が PHAT 法（又は白色化相互相関法 [11] や CSP: Cross power Phase 法 [12, 13] とも呼ばれる）である。PHAT 法では，次式のように定義される白色化相互相関関数 $\varphi_{\mathrm{P12}}(\tau)$ の最大値を与える τ の値として時間差 τ_s を推定する。

$$\begin{aligned}\varphi_{\mathrm{P12}}(\tau) &= \int \psi_{\mathrm{PHAT}}(\omega)\cdot\Phi_{12}(\omega)\cdot e^{j\omega\tau}d\omega \\ &= \int \frac{1}{|\Phi_{12}(\omega)|}\cdot\Phi_{12}(\omega)\cdot e^{j\omega\tau}d\omega \end{aligned} \tag{5}$$

この方法は，一般化相互相関関数の中でも高い反射音耐性を持つとされている [14] ため，本論文における提案手法との性能比較対象として用いることにする。



図–2 初期反射音の影響

4. ピークホールド処理

4.1 相互相関関数の問題点とピークホールド

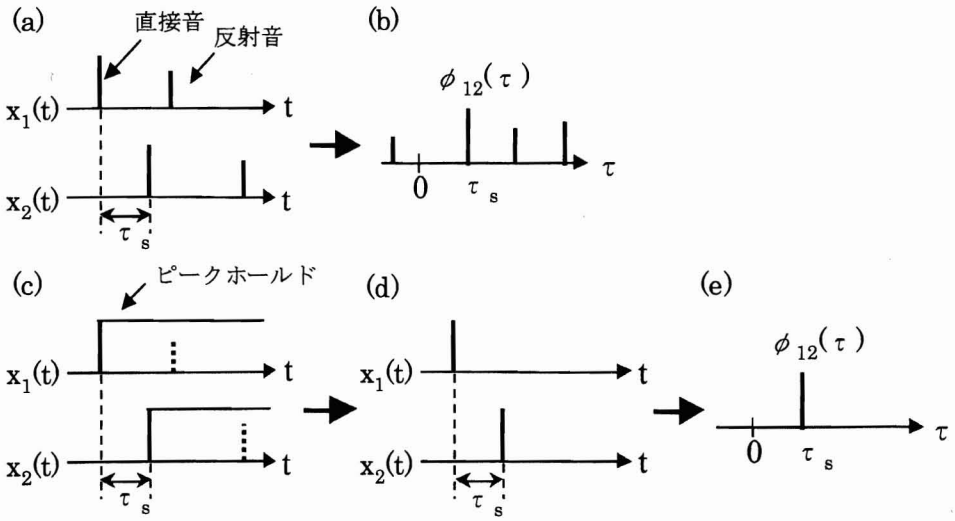
室内遠距離で受音する場合，近距離に比べ，直接音対反射音（DR）比が低下する。また，図–2 のようにアレーが壁際に置かれた場合には，近接した壁からの初期反射音の影響が大きくなる。そして，それらの初期反射音が特定の時刻に集中すると，その影響は更に大きなものとなる。よって，これらのような環境下では，良好な推定結果を得ることが難しい。

次に，ピークホールド処理を用いることで，この問題を改善できることを図–3 によって説明する。図–3(a) はパルス信号の直接音に単一反射音が付加された場合の受音信号 $x_1(t)$, $x_2(t)$ を表しており，第 1 のマイクロホンの受音信号 $x_1(t)$ には直接音及び反射音が含まれている。また，第 2 のマイクロホンの受音信号 $x_2(t)$ には $x_1(t)$ に比べて直接音が時間 τ_s 遅れて受音されており，反射音は $x_1(t)$ の場合とは異なった時間間隔で含まれている。

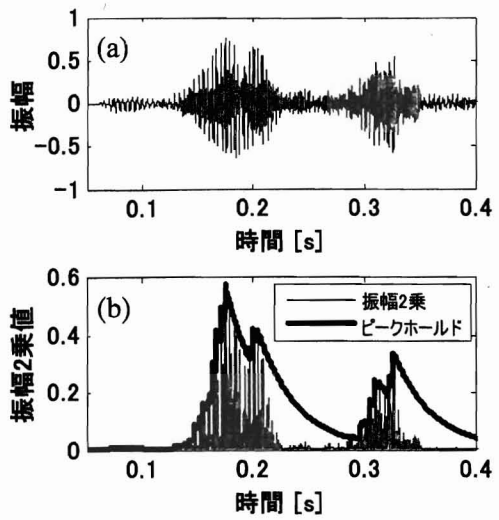
図–3(b) はそれらから直接計算した相互相関関数 $\varphi_{12}(\tau)$ を示している。これより，相互相関関数には，直接音の時間差 τ_s 以外に直接音と反射音，反射音同士の時間差に起因する複数のピークが発生し，誤推定の原因となりうることが分かる。そこで，この影響を軽減するために信号の最大値を一定時間保持するピークホールド処理を用いた手法を検討する。

図–3(c) は受音信号にピークホールドをかけた信号を示し，(d) はそれらの信号に対して時間差分をとった信号を表している。そして，(e) は (d) の二つの信号から算出される相互相関関数を表している。このように，ピークホールドは直接音の大きさを保持して，後続する低振幅の反射音をマスクするので，直接音の時間差 τ_s が明確となる。

ただし実際の適用において，一定値での保持を行うと，時間的に継続して発声される直接音をもマスクしてしまい，後続する直接音成分の利用ができなくなったり，話者の移動などへの追随が困難となったりする。



図–3 ピークホールド処理の効果を示すモデル図



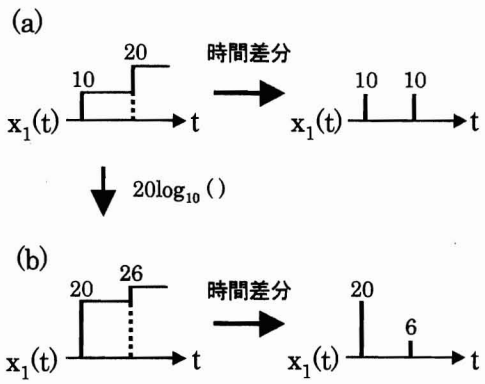
図–4 ピークホールド信号の例
(a) 受音信号（音声），(b) ピークホールド信号

そこで，ピークホールド値に室内残響相当の減衰を持たせることにした。図–4 に，実際の音声を用いた例を示す。(a) は受音信号波形，(b) はその振幅 2 乗とそれにピークホールドをかけた信号を示す。

4.2 対数操作

複数の初期反射音が近接した時刻に到来した場合，それらの振幅が加算され，直接音の振幅よりも大きくなることがある。

図–5 に，直接音（パルス音）の 2 倍の反射音が付加された例を示す。(a) はピークホールド信号とその時間差分，(b) はピークホールド後に対数をとった信号とその時間差分を示す。このように，ピークホールドのみでは，振幅の大きい反射音をマスクしきれない。そこで，振幅を対数化することで，信号の立ち上がり部分を更に強調し，その影響を軽減することができる。



図–5 対数操作の効果を示すモデル図
(a) 対数操作なし，(b) 対数操作あり

5. サブバンドピークホールド（SBPH）処理

5.1 音声スペクトルの特徴

ここまでは，パルス信号を例にとって説明してきたが，本章では音声信号への応用を考える。まず，パルス信号に類似し，音源方向推定が比較的容易な拍手音と，困難な音声との時間–周波数スペクトル上の違いについて説明していく。

図–6(a) に拍手音のスペクトログラム，図–6(b) に音声のスペクトログラムを示す。一般的に，音の方向情報は信号の立ち上がり部分の直接音において明確となる。拍手音では，全帯域で信号の立ち上がり時刻が同一となっているので，この時刻の時間波形にピークホールドをかけることが有効となる。

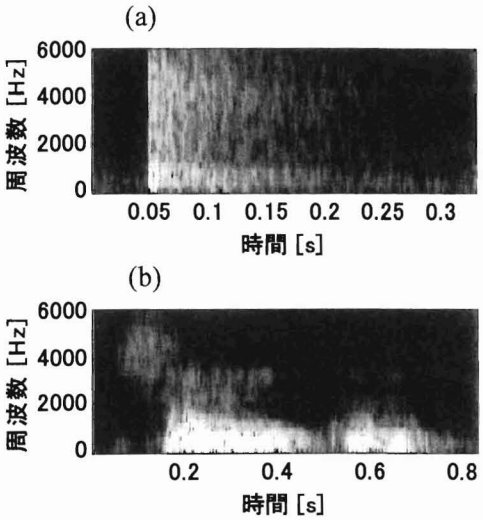
一方，音声の場合には，立ち上がり時刻が帯域ごとに異なり，また帯域によっては成分を持たないことがある。よって，時間波形上では各帯域の直接音成分が時間的に分散してしまい，不明確となるためにピークホールド処理の効果が十分発揮できない。

5.2 サブバンドピークホールド処理

前述した音声のように，立ち上がり時刻が帯域ごとに異なる信号に対しても対応可能とするためには，受音信号を帯域分割し，帯域ごとにピークホールドをかけることで直接音のみの観測機会を増加させることが有効となる [15]。この処理をサブバンドピークホールド（SBPH: Sub-Band Peak Hold）と呼び，その処理系のブロック図を図–7 に示す。

図において，音源方向推定の前処理に，音声区間推定処理（VAD: Voice Activity Detection）を設置している。これにより，方向性を持つような非音声の環境雑音（例えばドアの開閉音など）の影響を低減させている。また，4.2 節で述べた対数操作を無音声時に行うと，雑音の影響で誤差要因となるので，その影響の軽減効果も有している。使用した音声区間推定法は付録に示す。

処理の流れは，まず音声区間推定により受音信号から抽出した音声を短時間フーリエ変換（STFT）し，振幅成分の時系列 $|X_i(\omega,t)|$ $(i=1,2)$ を出力する。そして，周波数成分ごとにピークホールド処理（PH），対数操作（log），時間差分（Diff）の処理を行う。これらの処理は，具体的には，次のような手順で処理を行う。



図–6 拍手音と音声のスペクトログラム
(a) 拍手音，(b) 音声

1) ピークホールド処理（PH）

各周波数成分のピークホールド値を $Ph(\omega)$ と表し，以下の処理を各離散時間 t（時間間隔は第 6.1 節で述べるシフト長）において行う。

$|X_i(\omega,t)| \geqq Ph(\omega)$（ホールド値より大きい）場合

$|X_i(\omega,t)|_{\mathrm{h}} = |X_i(\omega,t)|$（振幅はそのまま）

$Ph(\omega) = |X_i(\omega,t)|$（ホールド値の更新）

$|X_i(\omega,t)| < Ph(\omega)$（ホールド値より小さい）場合

$|X_i(\omega,t)|_{\mathrm{h}} = Ph(\omega)$（ホールド値を振幅とする）

$Ph(\omega) = \alpha \cdot Ph(\omega)$（ホールド値を減衰させる）

ただし，

$|X_i(\omega,t)|_{\mathrm{h}}$：ピークホールド処理の出力

α：ピークホールド値の減衰係数（室内の残響時間 T_{60} を利用）

$$\alpha = 10^{-3/(T_{60} * F_{\mathrm{s}})}$$

F_{s}：標本化周波数

なお，ピークホールド値の初期値は雑音区間の平均値の数倍程度を与える。また，処理は図–4 に示すように振幅 2 乗値に行っても，以下で対数をとるので同等な処理（全体が 2 倍されるだけ）になる。

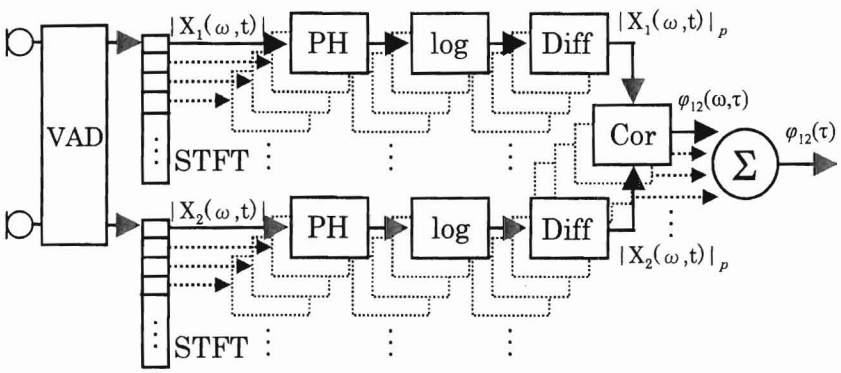
2) 対数操作（log）及び時間差分（Diff）

$$|X_i(\omega,t)|_{\mathrm{p}} = \log|X_i(\omega,t)|_{\mathrm{h}} - \log|X_i(\omega,t-1)|_{\mathrm{h}}$$

ただし，

$|X_i(\omega,t)|_{\mathrm{p}}$：対数操作及び時間差分の出力

その後，二つの受音信号の対応する周波数成分時系列ごとの正規化相関関数（Cor）$\varphi_{12}(\omega,\tau)$ を以下の式



図–7 サブバンドピークホールド（SBPH）処理のブロック図

(6) から求めて，これをすべての周波数に対して加算し（式 (7)），その最大値から到来方向を推定する。

$$\varphi_{12}(\omega,\tau)=\frac{\int |X_1(\omega,t)|_{\mathrm{p}}\cdot|X_2(\omega,t-\tau)|_{\mathrm{p}}dt}{\sqrt{\int |X_1(\omega,t)|_{\mathrm{p}}^{2}dt\cdot\int |X_2(\omega,t)|_{\mathrm{p}}^{2}dt}} \tag{6}$$

$$\varphi_{12}(\tau)=\sum_{\omega}\varphi_{12}(\omega,\tau) \tag{7}$$

なお，以下，本論文では対数操作，時間差分を含めた図–7 の処理系全体をサブバンドピークホールド処理と呼ぶことにする。

6. パラメータ設定

サブバンドピークホールド処理を効果的に行うためには短時間フーリエ変換（STFT）のパラメータである分析フレーム長（窓長），フレームシフト長，窓関数を適切に定める必要がある。

6.1 シフト長

フレームシフト長は，サブバンド信号及びその相関関数の，時間軸上の標本間隔となる。そして，この標本間隔は遅延時間 τ_{s} の推定精度に対応するので，式 (2) より，音源方向 θ_{s} の推定精度に対応する。このシフト長は，周波数分析などでは，（フレーム長/2），などとフレーム長に比例させる場合が多いが，本方式ではフレーム長とは独立に，必要な θ_{s} の推定精度に基づいて決定するものとする。

例えば，θ_{s} が 60 deg. 方向で 2.5 deg. の精度（角度分解能）を持たせる場合を考える。マイクロホン間距離 $d=0.3$ m のとき，音源方向が 60 deg. 及び 62.5 deg. のときの遅延時間は，式 (1) より，それぞれ 1.57 ms 及び 1.53 ms で，これらを区別するには，約 0.04 ms の標本間隔（時間分解能）が必要である。これは，標本化周波数 32 kHz では，約 1 サンプルに相当する。

なお，推定精度は正面からの角度が大きくなるほど劣化するので，この場合には，θ_{s} が 60 deg. 以下の性能を保証するシフト長となる。

6.2 フレーム長

一般的に，フレーム長を長くして周波数分解能を高めることで，より多くの帯域で直接音成分が観測可能となり，反射音耐性が向上する。しかし，フレーム長を長くし過ぎると，直接音と反射音が一つのフレーム内に混在することになり，立ち上がり時間が不鮮明になって推定精度が低下する（正解方向に対してズレ）。

このことよりフレーム長は，「反射音を含まない程度に長く」定めることが適切である。例えば，直接音と初期反射音との経路差が 35 cm 程度であるとすると，直接音と初期反射音との時間差は約 1 ms である。フレーム端が減衰するような分析窓を利用する場合には実効長は半分になるので，フレーム長はその 2 倍の 2 ms 程度が適当と考えられた。これは，標本化周波数 32 kHz の場合には，64 サンプルとなる。

ただし，最適なフレーム長は，室内音響条件や方向推定の許容誤差などによって多少変化するので，本論文では実験的に最適なフレーム長を定めるものとした。

6.3 窓関数

短時間フーリエ変換を行う際，フレーム長が短い場合には，周波数特性において窓関数のサイドローブの影響が大きくなる。本手法では，メインローブの幅は大きいが，サイドローブの影響を大幅に軽減できるブラックマン窓を利用した。

7. 評価実験

本手法（SBPH 法）の有効性を確認するために 2 種の評価実験を行った。最初に計算機シミュレーションを用いて，多数の室内残響条件，アレー配置，音声条件における有効性を確認した。その後，実際の室内音場において，有効性の再確認を行った。

7.1 鏡像法による計算機シミュレーション

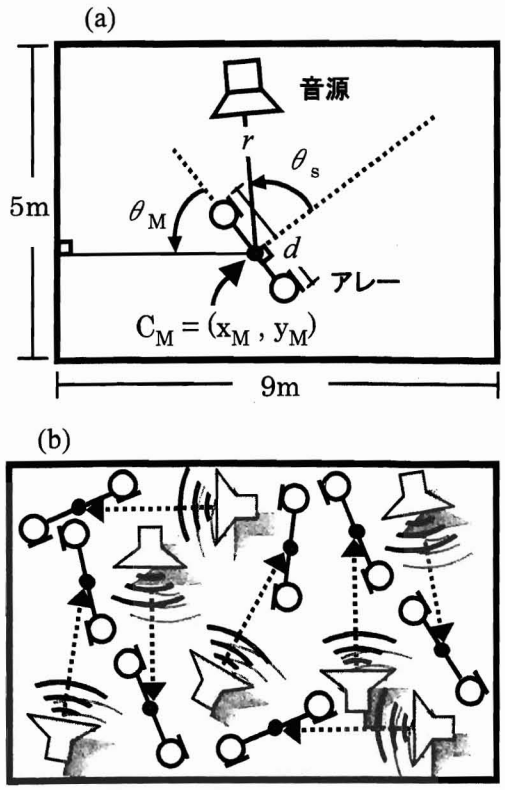
本手法の反射音耐性を評価するために，鏡像法に基づく室内インパルス応答の計算機シミュレーションを行った。反射音レベルは残響時間をパラメータとして 0〜0.4 s の範囲で 0.05 s ずつ変化させた。このシミュレーション条件を**表–1** に示す。

受音する反射音はアレーの位置や向きなどによって大きく変化するため，**図–8(a)** に示した音源距離 r を 3 m，アレーに対する音源方向 θ_{s} を 60 deg. に固定しながら，アレーの中心位置 $C_{\mathrm{M}}=(x_{\mathrm{M}},y_{\mathrm{M}})$，及び，壁面に対するアレーの角度 θ_{M} をランダムに 100 パターン変化させた（図–8(b)）。また，アレーの高さは，反射音の影響が大きい条件を想定し，床と天井の反射音が同時に到来する天井高の半分の高さに固定した。

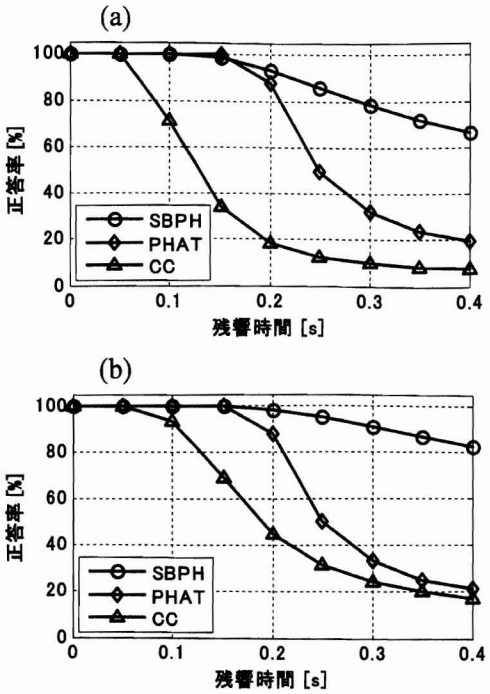
環境雑音は，標準的な室内騒音スペクトルを持つ Hoth 雑音 [16] が周囲から一様に到来している状態をシミュレートした。用いた音声は，電総研・単語音声データ

表–1 シミュレーション条件

標本化周波数	16,000 Hz
部屋の寸法	9.0 [W]×5.0 [D]×2.4 [H] [m]
アレーの高さ	1.2 m
マイク間距離 d	0.3 m
音源距離 r	3.0 m
音源方向 θ_{s}	60 deg.
SN 比	30 dB



図–8 鏡像法によるシミュレーション
(a) アレーの配置図，(b) アレーの配置の変化のイメージ図



図–9 シミュレーション結果
(a) 許容誤差 5 deg.，(b) 許容誤差 10 deg.

ベース（ETL-WD-I & II）から，男声 95 パターン，女声 63 パターンをランダムに選出し，アレーの配置条件と合わせて合計 15,800 条件において音源方向推定を行った。比較対象の従来法には，相互相関（CC）法と反射音耐性を持つとされている PHAT 法を用いた。

なお，シミュレーションにおける SBPH 処理のパラメータとして，フレーム長は予備実験より得られた最適値である 1 ms（標本化周波数 16 kHz で，16 サンプル）を用いた。

シミュレーション結果を図–9 に示す。(a) は許容誤差を 5 deg. としたとき，(b) は許容誤差を 10 deg. としたときの各残響時間における正答率を示している。許容誤差 5 deg. とした図–9(a) において，CC 法の正答率は，反射音レベルが増加して残響時間が 0.05 s より大きくなると急激に低下していることが分かる。一方，PHAT 法では CC 法より高い正答率が得られているが残響時間が 0.2 s を超えると急激に正答率が低下している。これに対し，SBPH 法は緩やかに低下し，残響時間 0.4 s においても，他の手法に比べて 3 倍以上の正答率（約 65%）が得られている。また，許容誤差を拡大して 10 deg. とすると，図–9(b) に示すように，SBPH 法及び CC 法ではともに正答率が向上している。これに対して PHAT 法では，許容誤差の拡大が正答率にあまり反映されない。このことは，第 7.2.2 項で示すように，PHAT 法の誤推定方向は正解方向の付近ではなく，大きく外れた方向であったことを意味している。

以上の結果より，提案手法は従来の音源方向推定法と比べて反射音の影響を大幅に軽減できると予想された。

なお，このシミュレーションにおける音源は無指向性音源を仮定しているので，指向性を有するスピーカなどの実音源と比べると反射音の影響を受け易く，実際の部屋の（残響時間の）場合より正答率が低く評価されていることを補足しておく。

7.2 実環境下における音源方向推定実験

次に，実際の室内における音源方向推定実験により，反射音耐性の評価を行った。ここでは，まず実験 1 として，部屋の中心位置付近で実験を行い，その後，実験 2 として，反射音の影響が大きいと予想される部屋の角で実験を行った。

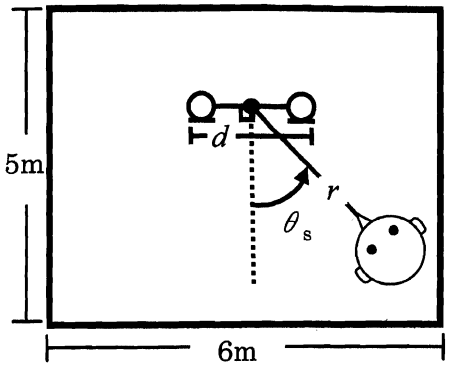
7.2.1 実験 1（部屋の中心位置付近）

音源方向推定に悪影響を及ぼす初期反射音が比較的少ない条件として，部屋の中心位置付近にアレーを配置して実験を行った。表–2 に実験条件を，図–10 に実験配置図を示す。

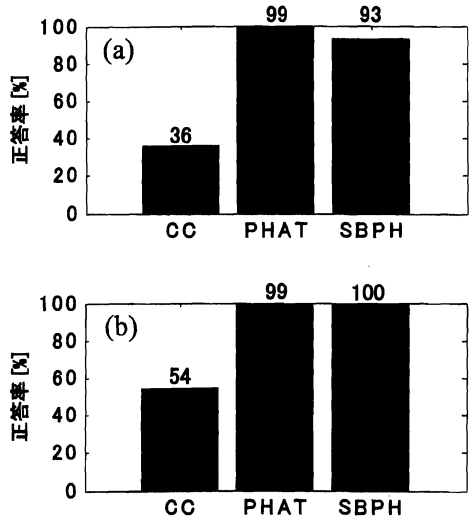
音源としてスピーカを用いると，人間が発声するより指向性が鋭いので反射音の影響が軽減される（実験条件下で DR 比が約 5 dB 上昇した）。よって，音源にはスピーカを用いず，人間（男性）が直接発声するよ

表–2 実環境下での実験条件

標本化周波数	32,000 Hz
部屋の寸法	5.0 [W]×6.0 [D]×2.5 [H] [m]
アレーの高さ	1.2 m
マイク間距離 d	0.3 m
音源距離 r	3.0 m
音源方向 θ_s	0, 30, 45 deg.
残響時間	0.38 s
SN 比	25～30 dB



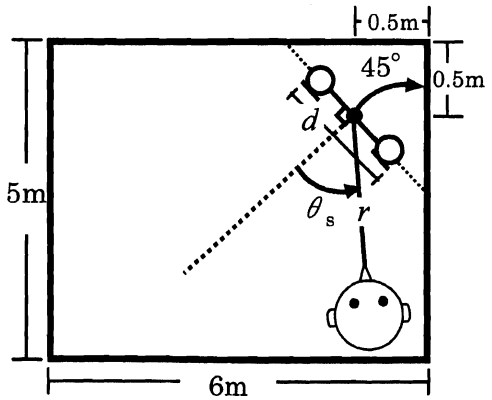
図–10 実験 1 の配置図



図–11 実験 1 の正答率
(a) 許容誤差 5 deg., (b) 許容誤差 10 deg.

うにした。音源距離 r は 3 m と固定した。実験で用いた発話単語は，シミュレーションで利用した単語から 30 語をランダムに選出した。各単語・角度（3 方向）ごとに 1 回ずつ発話したため，合計 90 条件で測定を行った。比較対象の従来法には，CC 法と PHAT 法を用いた。

実験 1 の結果の正答率を図–11 に示す。(a) が許容誤差を 5 deg. としたとき，(b) が許容誤差を 10 deg. としたときの正答率である。許容誤差 5 deg. のとき，CC 法の正答率が 36%であるのに対し，PHAT 法は



図–12 実験 2 の配置図

99%と高い正答率を示している。この結果は，従来知られている PHAT 法の反射音耐性の高さを確認するものである。また，SBPH 法は PHAT 法より少し低い 93%の正答率であった。

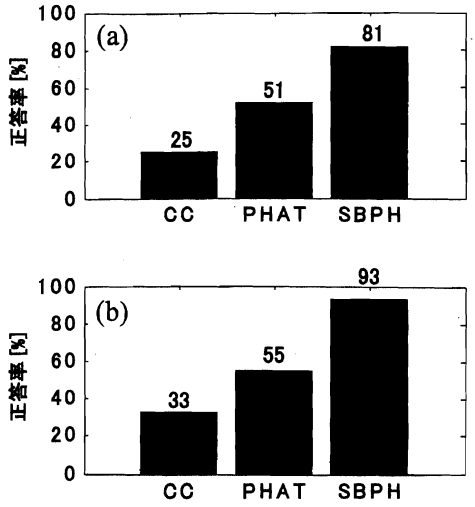
次に，許容誤差を 10 deg. として評価すると，SBPH 法の正答率が 100%まで向上することが分かった。これらのことより，初期反射音の影響が少ない環境では，1) PHAT 法は高い精度で音源方向推定することができる，2) SBPH 法は許容誤差を 10 deg. と多少大きめにとると PHAT 法と同様に高い正答率が得られる，ということが分かった。

7.2.2 実験 2（部屋の角）

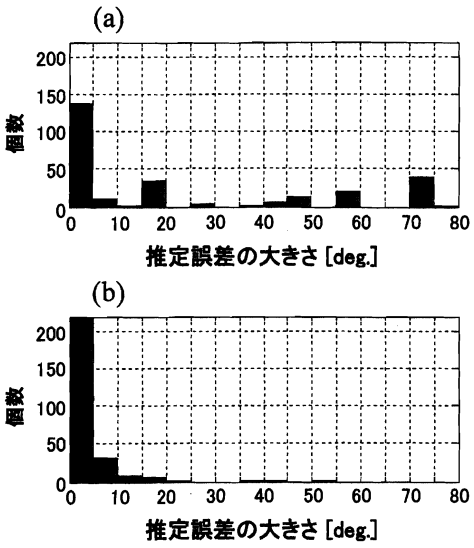
次に，初期反射音の影響が大きい条件として，部屋の角に近接するようにアレーを配置して実験を行った。標本化周波数などの条件は，実験 1 と同様の表–2 に示したもので，実験配置図は図–12 に示すものとした。具体的には，アレー中心が各壁面から 50 cm 離れ，またアレーの軸は図の縦方向の壁面に対して 45 deg. の角度で配置した。

音源，発声単語，音源距離，比較対象は，実験 1 と同じ条件とした。また，実験 2 では，各単語・角度（3 方向）ごとに 3 回ずつ発声したため，合計 270 条件で測定を行った。

実験結果の正答率を図–13 に示す。(a) が許容誤差を 5 deg. としたとき，(b) が許容誤差を 10 deg. としたときの正答率である。これより，許容誤差 5 deg. のとき，CC 法の正答率が 25%，PHAT 法が 51%であるのに対し，SBPH 法では 81%と高い正答率が得られている。更に，許容誤差 10 deg. とすると PHAT 法の正答率は 55%（誤答率約 45%）であるのに対し，SBPH 法の正答率が 93%（誤答率 7%）と大幅に向上していることが分かる。このことより，SBPH 法が PHAT 法に対する優位性は，許容誤差を大きくするほうが大きくなることが分かり，許容誤差を 10 deg. とした場合



図–13 実験 2 の正答率
(a) 許容誤差 5 deg., (b) 許容誤差 10 deg.

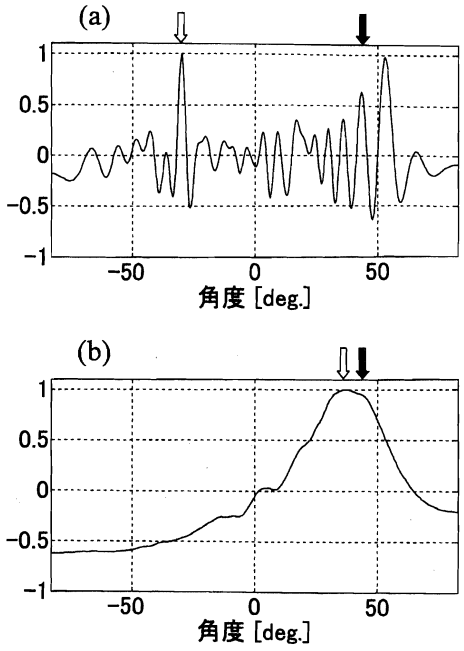


図–14 推定誤差のヒストグラム
(a) PHAT 法, (b) SBPH 法

には, PHAT 法による誤答率を約 1/6 に低減することが確認できた。

次に, このことを推定誤差の大きさのヒストグラムにおいて見てみた (図–14)。(a) が PHAT 法, (b) が SBPH 法による推定誤差の大きさである。図–14(a) より, PHAT 法では推定誤差の大きな誤りが幾つも見られる。これらは, 直接音とは大きく異なった方向から到来する反射音の方向を, 誤って音源方向として推定した結果と考えられる。これに対して, SBPH では, 誤差の大きさがほぼ 10 deg. の範囲に収まっていることが分かる。

更に, PHAT 法及び SBPH 法が典型的な誤り方を示した場合の相互相関関数を比較した (図–15 : 平均値を 0 として最大値で正規化表示)。(a) が PHAT 法,



図–15 典型的な誤推定の場合の相関関数の比較 (黒矢印 : 正解方向, 白矢印 : 推定方向)
(a) PHAT 法, (b) SBPH 法

(b) が SBPH 法の相互相関関数である。横軸は時間を推定角度に変換して表示した。それぞれの図において, 白矢印は相関関数が最大値をとる推定方向, 黒矢印は正解方向を示している。図 (a) より, PHAT 法は正解の方向に鋭いピークを示すが, それと同時に, 反射音に起因した複数のピークが発生することが分かる。その結果, 反射音の影響が大きい環境下では, 反射音によるピークの方が大きくなり, 大幅な誤推定を引き起こす。これに対して, 図 (b) より, SBPH 法ではピークが正解の方向より多少左右にずれてしまうが, 図 (a) に見られるような反射音によるピークは発生しておらず, その結果, 推定誤差を小さく抑えることができる。

これらの結果より, 本手法 (SBPH 法) は, 反射音の影響で音源方向推定に若干のズレが発生する場合もあるが, 反射音の影響による大きな誤りは低減できることを示した。

8. むすび

本論文では, 室内反射音の影響を軽減する新しい音源方向推定法として, サブバンドピークホールド処理を用いた手法 (SBPH 法) を提案し, その有効性を確認した。

SBPH 法は, 受音信号を短時間フーリエ変換によりサブバンド信号とし, その振幅時間系列にピークホールド処理を行った後に, 相関を計算する方法である。今回は二つのマイクの時間差に基づいて音源方向推定を

行う手法に適用し，最も一般的な相互相関（CC）法及び従来，反射音耐性が良好であるとされているPHAT法（CSP法）との比較評価を行った。

性能評価は，鏡像法による残響シミュレーションと実環境実験において行った。まず，残響シミュレーションの結果，SBPH法は，特に高残響下において，CC法やPHAT法よりも高い正答率を得ることができた。

次に，実環境実験を行い，以下の結果を得た。

1) CC法は反射音の影響による推定誤差が大きい。

2) PHAT法は，初期反射音の影響が少ない部屋の中心付近では高精度に音源方向推定を行うのに対して，初期反射音の影響が大きい壁際では，推定正答率が大幅に（約半分に）低下した。

3) SBPH法は，推定許容誤差を10 deg.とした場合には，初期反射音の影響が大きい壁際であっても93%の正答率が得られた。誤答率はPHAT法の45%に対してSBPH法は7%と，約1/6に低下させることができた。

4) 実験結果の誤り分析を行った結果，PHAT法は様々な方向から到来する初期反射音によって，方向を大幅に誤推定するのに対し，SBPH法は反射音の影響が小さく，多くの誤差が音源方向から10 deg.以内に含まれているという特徴を有していることが分かった。

以上の結果より，今回提案したサブバンドピークホールド（SBPH）法は，室内反射音の影響の大きい環境下での音源方向推定に有効な手法であることを確認した。

本論文で提案したサブバンドピークホールド処理は，相互相関関数への反射音の悪影響を軽減するものである。よって，相互相関関数を利用するMUSIC法などのサブスペース法にも組み合わせることができ，性能を改善できると考えるが，その検討は今後の課題とする。

謝　辞

本研究を行うにあたり，特にお世話になった方々を列記して，心より深く感謝いたします。

東京電機大学大学院 音響信号処理研究室卒業生の上杉信敏氏（現所属はヤマハ㈱）には，本研究を進める上でのご指導並びに貴重なご意見などをいただいた。木皿大介氏（現所属は三菱電機㈱）には，先行研究にあたる貴重な成果及び論文を残していただいた。また，本研究室在籍中の佐藤耕平氏には，評価実験を行うにあたり，協力いただいた。最後に，査読者の方には，大変有益なご指摘をいただき，論文の質の向上をはかることができた。

文　献

[1] 山本　潔, 浅野　太, 原　　功, 緒方　淳, 麻生英樹, 山田武志, 北脇信彦, “ヒューマノイドロボットHRP2における音響情報と画像情報を統合したリムタイム音声インタフェース,” 音響学会誌, 62, 161–172 (2006).

[2] 小林和則, 古家賢一, 羽田陽一, 片岡章俊, “複数の小型マイクロホンアレーと超音波距離計測を用いた高精度話者位置推定,” 信学技報, EA2007-88, pp. 13–18 (2007).

[3] 上杉信敏, 金田　豊, “音源方向推定に及ぼす室内反射音影響の分析的検討,” 信学技報, EA2006-105 (2007).

[4] A. Stephenne and B. Champagne, “Cepstral prefiltering for time delay estimation in reverberant environments,” *Proc. ICASSP 95*, Vol. 5, pp. 3055–3058 (1995).

[5] 大井堂史昌, 陶山健仁, “実環境におけるMUSIC法による到来方向推定の改善,” 信学技報, EA-2006-10 (2006).

[6] 日岡裕輔, 浜田　望, “反射音の存在する環境における音声の到来方向推定,” 信学技報, EA-2002-111 (2003).

[7] 木皿大介, 金田　豊, “音声に対するピークホールド音源方向検出法の検討,” 音講論集, pp. 631–632 (2006.3).

[8] 金田　豊, “室内残響下における広帯域音源の方向推定,” 音講論集, pp. 547–548 (1991.10).

[9] C.H. Knapp and G.C. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. Acoust. Speech Signal Process.*, **ASSP-24**, pp. 320–327 (1976).

[10] 大賀寿郎, 山崎芳男, 金田　豊, 音響システムとディジタル信号処理（㈳電子情報通信学会, 東京, 1995).

[11] 林　範章, 岩橋清勝, 山田一郎, “信号の白色化による航空機騒音識別手法の改良とハードウェアによる実現,” 信学技報, EA89-38, pp. 9–16 (1989).

[12] M. Omologo and P. Svaizer, “Acoustic event localization using a crosspower-spectrum phase based technique,” *IEEE ICASSP 94*, II-273-276 (1994).

[13] Y. Denda, T. Nishiura and Y. Yamashita, “Robust talker direction estimation based on weighted CSP analysis and maximum likelihood estimation,” *IEICE Trans. Inf. Syst.*, **E89-D**, pp. 1050–1057 (2006).

[14] M.S. Brandsteim, “Time-delay estimation of reverberated speech exploiting harmonic structure,” *J. Acoust. Soc. Am.*, **105**, 2914–2929 (1999).

[15] 鈴木　敬, 金田　豊, “サブバンドピークホールド処理を用いた音源方向推定の検討,” 音講論集, pp. 751–752 (2007.9).

[16] D.F. Hoth, “Room noise spectra at subscribers' telephone locations,” *J. Acoust. Soc. Am.*, 12, 499–504 (1941).

[17] J. Sohn, N.S. Kim and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Process. Lett.*, 6, 1 (1999).

[18] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust. Speech Signal Process.*, **ASSP**-32, 1109–1121 (1984).

[19] 古井貞熙, 新音響・音声工学（近代科学社, 東京, 2006).

付　録

今回使用した音声区間推定法（VAD）について説明する。音声区間は，主に二つの手法を組み合わせて推定した。一つは，確率モデルに基づく手法で，定常性雑音に対して高いVAD性能を示す。もう一つは，音声の周期性の有無に着目した手法で，非定常な突発性雑音を音声と誤判定することを防止する。

最終的な判定は，一つの音声区間には必ず有声音が存在すると仮定し，確率モデルに基づく判定結果のうち，有声音が存在しない区間を非音声区間と判定する

ことで，突発性雑音を排除する。

A.1 MMSE の確率モデルに基づく手法 [17]

定常雑音に頑健な推定法として，MMSE（Minimum Mean Square Error）[18] の確率モデルに基づく手法が提案されている。この手法は，受音信号が音声状態（H_1）と非音声状態（H_0）を遷移する信号であると仮定している。よって，周波数帯域ごとにそれぞれの状態に属する確率の比（尤度比 Δ_ω）を式 (A.1) から求め，その相乗平均 $\log \Delta$ 式 (A.2) が閾値より大きい区間を音声区間候補とするこれらの確率モデルは MMSE で定義されているものを用いる。

$$\Delta_\omega = p(X_\omega|H_1)/p(X_\omega|H_0) = 1/(1+\xi_\omega)\exp\{\gamma_\omega\xi_\omega/(1+\xi_\omega)\} \quad \text{(A.1)}$$

$$\log \Delta = \sum_\omega \log \Delta_\omega \quad \text{(A.2)}$$

ここで，X_ω は入力信号の短時間スペクトル，ξ_ω は事前 SN 比，γ_ω は事後 SN 比を示す。また，ξ_ω は直接決定法 [17] により推定した。

A.2 LPC 分析に基づく手法 [19]

音声の特徴でもある周期性の有無に着目した推定法を併用することで，突発性雑音に対しても頑健とする。具体的には，LPC（Linear Predictive Coding）分析の残差信号を算出し，その自己相関関数から周期性の有無を検出する。この時，周期性を持つ有声音の場合は基本周期の整数倍で大きな相関を示す。よって人間の基本周期（ピッチ周期）にあたる 3〜10 ms 間に大きな相関を示した場合には有声音が存在すると推定できる。

鈴木　敬

1984 年生。2007 年東京電機大学・工・情報通信工学科卒。2009 年同大大学院情報通信工学専攻修士課程修了。在学中は，音場計測分野における音響信号処理の研究に従事。日本音響学会会員。現在，東京電力㈱に勤務。

金田　豊

1951 年生。1975 年名大・工・電気卒。1977 年同大大学院修士課程了。同年日本電信電話公社（現 NTT）入社。NTT 研究所において，マイクロホンアレー信号処理，音響エコーキャンセラ，音響計測などの音響信号処理の研究に従事。2000 年より東京電機大学情報通信工学科教授。現在に至る。工博。1989 年日本音響学会佐藤論文賞，1989 年 IEEE ASSP Senior Award，2009 年日本音響学会佐藤論文賞など受賞。日本音響学会，電子情報通信学会，米国音響学会，IEEE 各会員。