社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

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

鈴木 敬† 金田 豊‡

東京電機大学大学院工学研究科 〒101-8457 東京都千代田区神田錦町 2-2

E-mail: †07gmc08@ms.dendai.ac.jp, ‡kaneda@c.dendai.ac.jp

あらまし 本報告では，音源方向推定技術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処理を用いた手法を検討する。この方法は先行する直接音の振幅を保持し，後続の反射音をマスクする処理（ピークホールド）をサブバンド信号に対して適用することで，反射音を軽減する。この処理を適用する音源方向推定法としては，2 マイク信号間の時間差検出に基づく相互相関法を利用した。評価実験として，実環境下における音源方向推定実験を行い，提案法の有効性を検討した。その結果，従来手法である CC 法，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

A study of DOA estimation based on Sub-band Peak Hold processing

Takashi SUZUKI† Yutaka KANEDA‡

Graduate School of Engineering, Tokyo Denki University

2-2 Kanda-Nishiki-cho, Chiyoda-ku, Tokyo, 101-8457 Japan

E-mail: †07gmc08@ms.dendai.ac.jp, ‡kaneda@c.dendai.ac.jp

Abstract In this paper, we discuss the DOA (Direction Of Arrival) estimation method using sub-band peak hold processing, which avoid the undesirable effect of reflected sound in a room. In this method, the reflected sound is reduced by peak hold processing, which maintains the amplitude of the direct sound and masks subsequent reflected sound at each sub-band. The sub-band peak hold processing is applied with the cross-correlation function method, which is a typical DOA estimation method. To evaluate the effectiveness of the proposed method, a DOA estimation experiment was conducted under actual environmental conditions. As a result, the robustness of this method for reflected sound was shown to be higher than those of the conventional CC (Cross Correlation) and PHAT (PHAse Transform) methods.

Keywords reflected sound, DOA estimation, cross-correlation function, microphone array, sub-band 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]。この対策として我々は，先行する直接音の振幅を保持して，後続する反射音をマスクするピークホールド処理[4]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[5]の利用を提案する。その結果，反射音の影響を大幅に軽減することが確認できたので報告する。

代表的な音源方向推定の方法には，受音信号の時間差推定に基づく相互相関法[6]，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[7]，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[7]の 3 つが主に挙げられる。本論文では，最も基本的な推定法である 2 つのマイクロホンを用いた相互相関法を用いて検討を進める。

2. 音源方向推定モデル

図 1 に 2ch 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，2 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 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。

$$\tau_s = \xi / c = d\sin\theta_s / c \qquad (1)$$

この式を, θ_s について式を解くと次式のように表せる。

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

このとき，d,c の値は既知であるので，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. 相互相関関数法

2 つのマイクロホンに基づ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。

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

3.2. PHAT（Phase Transform）法

式(3)はクロススペクトル $\Phi_{12}(\omega)$を用いて，次式のように表すことができる。

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

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）[6]。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やCSP：Cross power Phase 法とも呼ばれる）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$\varphi_{P12}(\tau)$の最大値を与えるτの値として時間差τ_sを推定する。

$$\varphi_{P12}(\tau) = \int \Psi_{PHAT}(\omega)\, \Phi_{12}(\omega)\, e^{j\omega t}\, d\omega$$
$$= \int 1 / |\Phi_{12}(\omega)| \cdot \Phi_{12}(\omega)\, e^{j\omega t}\, d\omega \qquad (5)$$

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[9] ため，従来法として比較対象に用いることにする。

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となる場合には，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。

図 3 に，パルス信号の直接音に単一反射音が付加された場合の相互相関関数のモデル図を示す。図 3(a)は受音信号 $x_1(t)$, $x_2(t)$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。第 2 のマイクロホンの受音信号 $x_2(t)$には $x_1(t)$に比べて時間 τ_s 遅れて受音された直接音が含まれている。また，反射音が $x_1(t)$の場合とは異なった時間間隔

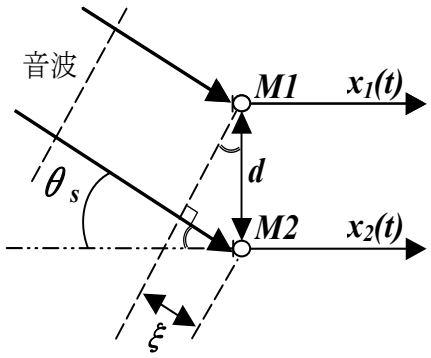


図 1　音源方向推定のためのモデル図

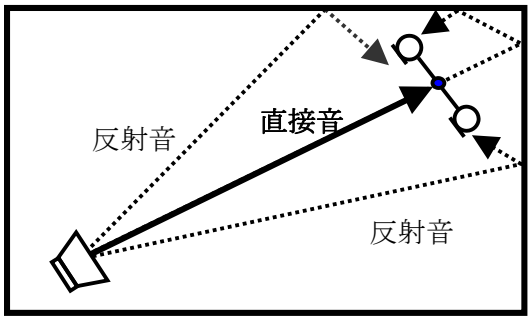


図 2　初期反射音の影響

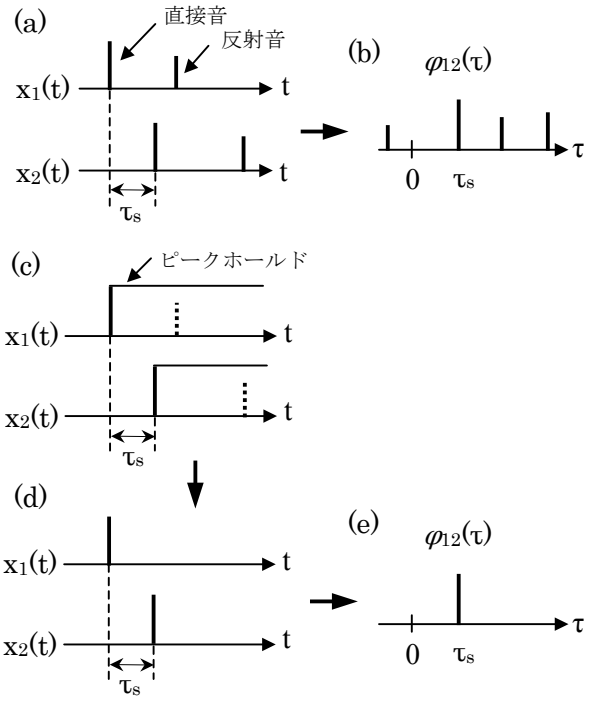


図 3　相互相関関数とピークホールドの効果を示すモデル図

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 2 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は，常に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。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。

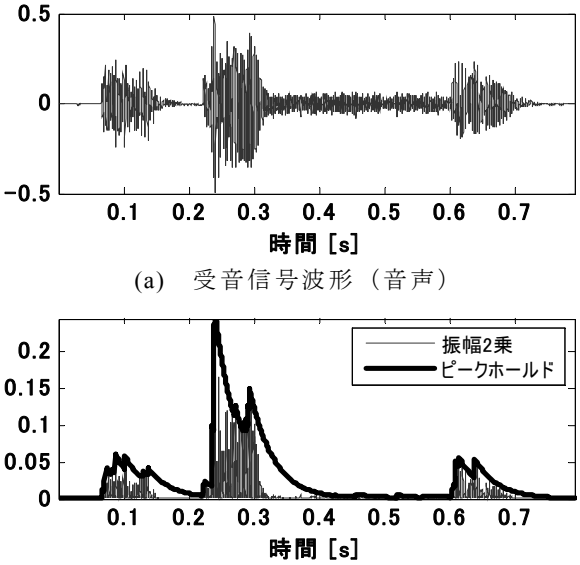
5. サブバンドピークホールド（SBPH）処理

5.1. 音声スペクトルの特徴

まず，比較的に音源方向推定が容易な拍手音と困難な音声との時間スペクトル上の違いについて説明する。

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毎に異なり，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



(a) 受音信号波形（音声）

(b) ピークホールド信号

図 4 ピークホールド信号の例

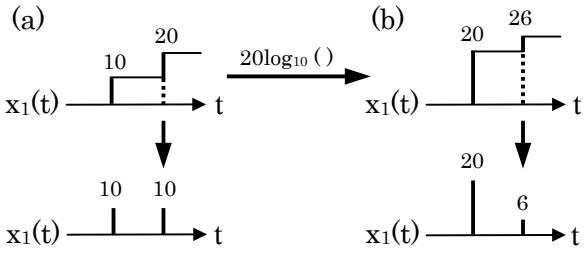
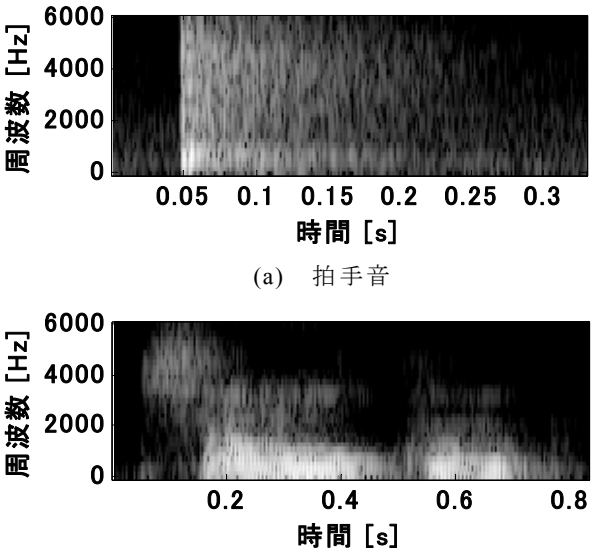


図 5 対数操作の効果を示すモデル図



(a) 拍手音

(b) 音声

図 6 拍手音と音声のスペクトログラム

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毎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，帯域毎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。この方法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，方向推定の前処理に，音声区間推定技術（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，周波数成分毎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。その後，2 つの受音信号の対応する周波数成分時系列毎の正規化相関関数（Cor）を式(6)から求めて，これを全ての周波数に対して加算（式(7)）し，その最大値から到来方向を推定する。ただし,$|X_i(\omega,t)|_p$は時間差分出力を示す。

$$\varphi_{12}(\omega,\tau) = \int |X_1(\omega,t)|_p \cdot |X_2(\omega,t-\tau)|_p \, dt \Big/ \left(\int |X_1(\omega,t)|_p \, dt \cdot \int |X_2(\omega,t-\tau)|_p \, dt \right) \quad (6)$$

$$\varphi_{12}(\tau) = \Sigma_\omega \, \varphi_{12}(\omega,\tau) \quad (7)$$

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のパラメータであるフレーム長，シフト長，分析窓関数を適切に定める必要がある。

6.1. STFT の分析フレーム長とシフト長

図 8(a)にシミュレーションで求めたインパルス応答を，(b)~(d)に分析フレーム長のサンプル数を 16, 32, 64 と変化させた場合のサブバンド信号（1000 Hz 付近）の振幅を示す。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る。しかし，(d)のように，長くし過ぎると直接音と反射音が融合して立ち上がり時間が不明確になる。このトレードオフ関係があるため，フレーム長は，用途によって使い分ける必要がある。本論文では,方向推定の許容誤差を 5deg.または 10deg.で評価しているため，ピーク値が少しずれても，予備実験において大きな誤差が比較的少なかったフレーム長の

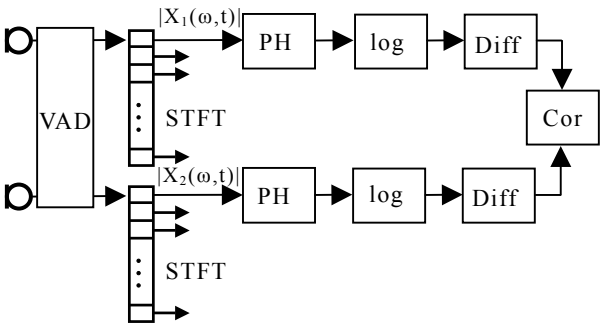


図 7　サブバンドピークホールド（SBPH）処理のブロック図

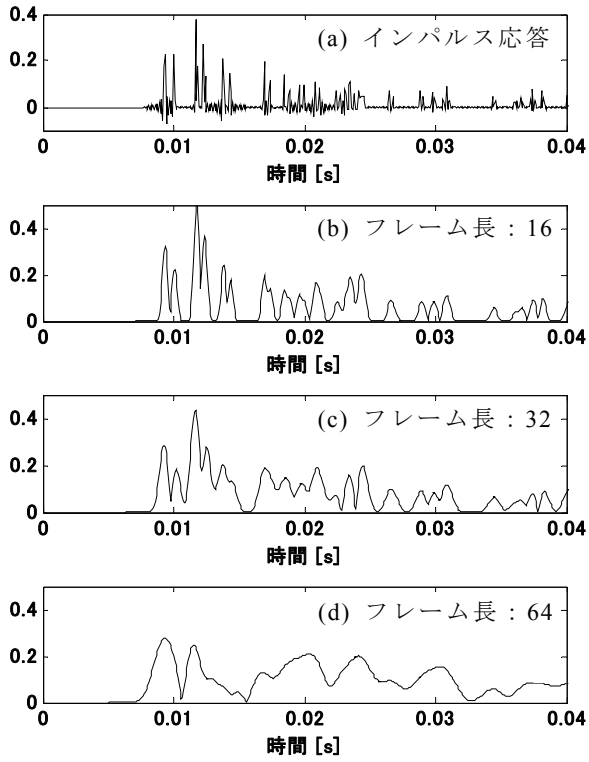


図 8　フレーム長と時間分解能

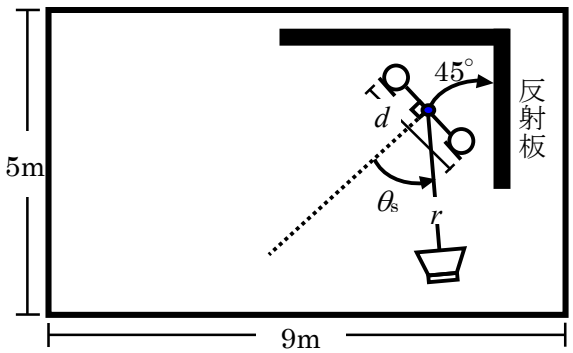


図 9　実験配置図

32 サンプルとした。また，今回は，演算量よりも推定精度の評価を狙いとしたため，シフト長は 1 サンプルとした。

6.2. 窓関数

短時間フーリエ変換する際，フレーム長が短いと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。そこで，本論文ではメインローブが幅を持っても,このサイドローブの影響を完全に抑えるために，メインローブの最大値に対してサイドローブの最大値が 100 dB 程度低下するようなチェビシェフ窓を設計し，使用した。

7. 評価実験（実環境下における方向推定実験）

提案法（SBPH 法）の有効性を確認するために，実際の室内音場において，反射音耐性の評価を行った。表 1 に実験条件を，図 9 に実験配置図を示す。この実験では，反射音の影響が大きい条件を想定し，アレーを囲むように 2 つの反射板を配置した。

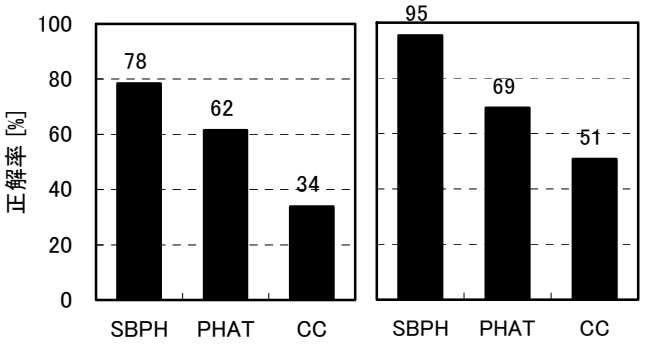
また，音源としてスピーカを用いると，人間が発声するより指向性が鋭いので反射音の影響が軽減される（実験条件下で DR 比が約 5dB 上昇した)。よって，音源にはスピーカを用いず，男性 2 人が直接発声するようにした。実験で用いた発声単語は「ばくおん」と「おんきょう」である。比較対象の従来法には，PHAT 法を用いた。

全 60 条件における実験結果の正解率を図 10 に示す。(a)が許容誤差を 5deg.とした時,(b)が許容誤差を 10deg.とした時の正解率である。これより，許容誤差 5deg.の時，PHAT 法の正解率が 62%であるのに対し，SBPH 法では 78%と高い正解率が得られている。更に，許容誤差 10deg.とすると PHAT 法の正解率は 69%（誤答率約 30%）であるのに対し, SBPH 法の正解率が 95%（誤答率 5%）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 10deg.とした場合には，PHAT 法による誤差率を 1/6 に低減することができた。

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詳しく見てみた（図 11)。(a)が PHAT 法，(b)が SBPH 法による推定誤差の大きさである。図 11(a)より，PHAT 法では推定誤差の大きな誤りがいくつも見られる。これらは，直接音とは大きく異なった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。一方, SBPH では，誤差の大きさがほぼ 10deg.の範囲に収まっている。このことは，反射音の影響で音源方向推定に若干のズレは発生するが，全く異なった方向からの反射音の影響は抑圧できていることを示している。この結果より，提案法（SBPH 法）が反射

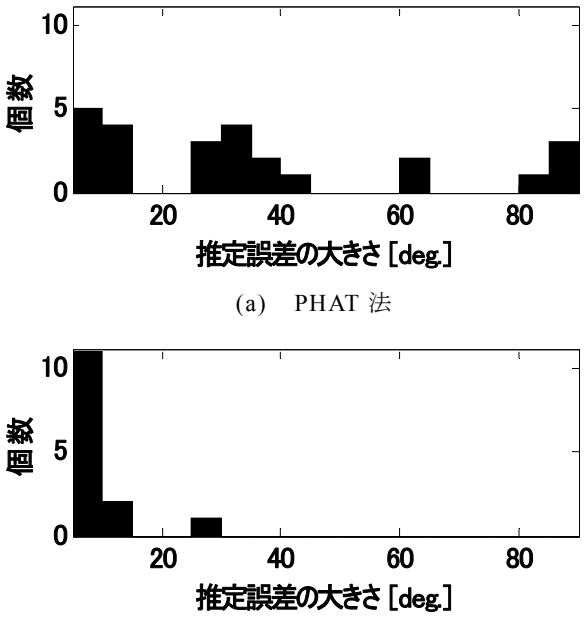
表 1　実験条件

標本化周波数	12000 Hz
部屋の寸法	9.0[W]×5.0[D]×2.4[H] [m]
残響時間	0.4 s
騒音レベル	45 dB
マイク間距離 d	0.6 m
音源距離 r	1, 2, 3 m
音源方向 θ_s	-60, -30, 0, 30, 60 deg.



(a) 許容誤差 5deg.　　(b) 許容誤差 10deg.

図 10　正解率



(a)　PHAT 法

(b) SBPH 法

図 11　推定誤差のヒストグラム

音の影響軽減に有効であることが示せたものと考える。

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 2 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法（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,実環境下実験において行い,その結果,SBPH 法は CC 法や PHAT 法よりも高い正解率を得ることができた。更に,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反射音の方向を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 10deg.以内に含まれていることが分かり，その有効性の確認ができた。

今後の検討課題としては以下のことが考えられる。

1) ピークホールド処理は,立ち上がりの異なる単語や音韻によって反射音耐性が変わってくると予想される。本報告では,「ばくおん」と「おんきょう」の 2 単語でのみ評価したが，もっと多くの単語で評価する必要がある。

2) 今回の実験では,騒音レベル 45 dB と比較的静かな環境での測定であったため，耐雑音性の評価もしていく必要がある。

3) これら数多くの条件における評価実験となるため，鏡像法を用いた計算機シミュレーションを併用させ，評価していく。

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 2 つの手法を組み合わせて推定した。1 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 1 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，1 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。

(1) MMSE の確率モデルに基づく手法 [9]

定常雑音に頑健な推定法として，MMSE (Minimum Mean Square Error) [10] 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，周波数帯域毎にそれぞれの状態に属する確率の比（尤度比 $\mathbf{\Lambda}_\omega$）を式(A1)から求め，その相乗平均 $\log\mathbf{\Lambda}$（式 A2）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。

$$\mathbf{\Lambda}_\omega = p(\mathrm{X}_\omega|\, H_1) \,/\, p(\mathrm{X}_\omega|\, H_0)$$
$$= 1 \,/\, (1+\boldsymbol{\xi}_\omega)\, \exp\{\boldsymbol{\gamma}_\omega\boldsymbol{\xi}_\omega \,/\, (1+\boldsymbol{\xi}_\omega)\} \qquad \text{(A1)}$$
$$\log\mathbf{\Lambda} = \mathbf{\Sigma}_\omega \log\mathbf{\Lambda}_\omega \qquad \text{(A2)}$$

ここで，X_ω は入力信号の短時間スペクトル，$\boldsymbol{\xi}_\omega$ は事前 SN 比，$\boldsymbol{\gamma}_\omega$ は事後 SN 比を示す。ここで，$\boldsymbol{\xi}_\omega$ は直接決定法[10] により推定した。

(2) LPC 分析に基づく手法 [11]

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 3ms ~10ms 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。

文　献

[1] 山本潔 他, “音響・画像情報の統合によるヒューマノイドロボット HRP-2 の音声インタフェース,” 音講論集, pp.35-36, Sep. 2005.

[2] 小林和則 他, “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, Dec. 2006.

[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, Jan. 2007.

[4] 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, Mar. 2006.

[5] 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, Oct. 1991.

[6] C. H. Knapp, G. C. Carter, ”The Generalized Correlation Method for Estimation of Time Delay,” IEEE Trans. on Acoust., Speech and Signal Proc., ASSP-24, 4, pp.320-327, Aug. 1976.

[7] 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, (社)電子情報通信学会, 1995.

[8] M. S. Brandsteim, ”Time-delay estimation of reverberated Speech exploiting harmonic structure,” J. Acoust. Soc. Am., 105, 5, pp.2914-2929, 1999.

[9] J. Sohn *et al*., ”A Statistical Model-Based Voice Activity Detection,” IEEE Signal Proc. Letters, 6, 1, Jan. 1999.

[10] Y. Ephraim, D. Malah, “Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator,” IEEE Trans. on Acoust., Speech and Signal Proc., ASSP-32, 6, pp.1109-1121, Dec. 1984.

[11] 古井貞熙, 新音響・音声工学, (株)近代科学社, 2006.