音源距離推定方式の比較検討とコスト関数の一般化*

◎太刀岡勇気, 成田知宏, 石井純 (三菱電機・情報総研)

1 研究の背景と目的

高騒音下で遠隔マイクによる音声認識を行うには、音源位置を推定し目的音を強調する必要がある。著者らは既報 [1] において、事前分布を利用した CSP 法を提案し、騒音下においても方向推定は実用的であることを示した。方向に加え距離が推定できれば、話者が特定のゾーン内から発話した場合のみ音声認識を受け付ける等の対策ができ、誤受理削減に有効である。ところが距離推定は、2 次元平面上で角度と距離を同時推定する問題となり、角度を推定するよりも格段に難しい。本報では、距離推定用に提案されている手法 (2D-CSP 法、マルチチャネル CSP 法、r-means 法) の比較検討を行う。またこれらがあるコスト関数を最小化する枠組みで一般化できることを示す。

2 既存の距離推定法

2.1 音波伝播の仮定

点音源からの音波は、音源からの距離が等しい点が等位相となる球面波として伝搬する。マイクアレイの中心からの距離 ρ が $\frac{2D^2}{\lambda}$ より小さい近傍場では球面波と考えられる [2]。ここで、D はマイクアレイの最大の幅であり、λ は音波の波長である。1 kHz の場合、$D = 0.3$ [m] で $\rho = 0.52$ [m]、$D = 0.6$ [m] で $\rho = 2.1$ [m] となる。球面波の場合、音源座標を $\mathbf{s} = (x^s, y^s)$、i 番目 $(1 \leq i \leq N)$ のマイク座標を $\mathbf{r}_i = (x_i^r, y_i^r)$ とすると、マイク i, j の到来時間差は $\tau_{ij}^{sp} = \frac{d_i - d_j}{c}$ で表される。c は音速、d_i は音源からマイク i までの距離 $|\mathbf{s} - \mathbf{r}_i|$ である。

一方、この条件を満たさない場合、音波の進行方向に直交する面で等位相となる平面波と考えられる。この場合、マイク i, j 間の到来時間差は入射角 θ の関数として $\tau_{ij}^{pl} = -\frac{\delta_x}{|\delta_x|}\frac{\sqrt{\delta_x^2+\delta_y^2}}{c}$ で表される。x^{og}, y^{og} はマイクアレイの中心の座標である。ここで $\theta = \tan^{-1}\left(\frac{y^s - y^{og}}{x^s - x^{og}}\right)$, $\delta_x = -(x_i^r - x_j^r)\cos\theta$, $\delta_y = (y_i^r - y_j^r)\sin\theta$ である。

2.2 CSP 法 (平面波仮定)

Cross-Spectrum Phase(CSP) 法は、2 ch 信号のクロススペクトルから信号間の到来時間差 τ を求める方法である。まず、Eq. (1) より CSP 係数を算出する。到来時間差 τ_{ij}^{csp} は $\arg\max_{\tau}(CSP(\tau))$ によって計算される [3]。

$$CSP(\tau) = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\eta_i(t))\,\mathcal{F}(\eta_j(t))^*}{|\mathcal{F}(\eta_i(t))\,||\mathcal{F}(\eta_j(t))\,|}\right) \quad (1)$$

η_i, η_j はマイク $i, j (1 \leq i, j \leq L)$ の入力、$\mathcal{F}$ は短時間フーリエ変換、*は複素共役を表す。平面波仮定では、音源の方向 θ は求まった到来時間差 τ_{ij}^{csp} から $\theta = \sin^{-1}\left(\frac{\tau_{ij}^{csp} c}{|\mathbf{r}_i - \mathbf{r}_j|}\right)$ により求まる。算出した複数の θ の交点から位置を推定する手法も提案されている [4]。これはコスト関数

$$P(\theta) = \left(\tau_{ij}^{pl} - \tau_{ij}^{csp} + \epsilon\right)^2 \quad (2)$$

を最小化する問題ともいえる。τ_{ij}^{csp} は誤差 ϵ を持つ。

2.3 2D-CSP 法 (球面波仮定)

2.2 は平面波を仮定して、θ を求める 1 次元の音源定位問題である。球面波を仮定して、$\mathbf{s}$ を求める 2 次元の問題を解く手法が 2D-CSP 法である [5]。

ここで、2 つのマイク対 (マイク 1,2 とマイク 3,4) を考える。簡単のためマイク間隔は同じとする。平面波の場合 $|d_1 - d_2| = |d_3 - d_4|$ であるため、マイク対間で時間差はない。球面波の場合には $|d_1 - d_2| \neq |d_3 - d_4|$ であり、この差を利用して音源までの距離を推定できる。理論上のマイク i, j 間の到来時間差は、τ_{ij}^{sp} により表される。これに対し CSP 法により、マイク間の到来時間差 τ_{ij}^{csp} を求める。ここで音源がある範囲を含む音源の候補点 $\mathbf{s}$ について、M 個のマイク対に対して、それぞれ理論値からのずれを加算したコスト関数 $P(\mathbf{s})$ の値を計算する (Eq. (3))。

$$P(\mathbf{s}) = \sum_{m=1}^{M}\left(\tau_{\varphi(m)}^{sp} - \tau_{\varphi(m)}^{csp} + \epsilon\right)^2 \quad (3)$$

ここで $\varphi(m)$ は m 番目のマイク対である。

$\tau_{\varphi(m)}^{csp}$ に、理論値 $\tau_{\varphi(m)}^{sp}$ が近い値をとるとき P が小さくなるから、球面波を仮定でき、かつ誤差 ϵ が小さければ、$P(\mathbf{s})$ を最小化する $\mathbf{s}$ が音源の座標であると推定できる。1 つのマイク対だけでは、ある双曲線上に音源があるとわかるだけなので、この推定には 2 つ以上のマイク対 (3 つ以上のマイク) が必要である。

2.4 マルチチャネル CSP 法 (M-CSP 法)

CSP 法はマイク対から到来方向を求めるが、M-CSP 法では N 本のマイクの全ペアの相関行列 $\mathbf{R} =$

*Comparative study on source's distance estimation methods and generalization of cost functions, by TACHIOKA, Yuuki, NARITA, Tomohiro, ISHII, Jun (Mitsubishi Electric Corp.).

(r_{ij}) $(1 \le i, j \le N)$ を求め、所与のステアリングベクトルと比較することで音源位置を推定する [6]。これにより、各マイク対での相関を参照できるため、推定精度が向上するとされる。各成分は

$$r_{ij} = \frac{\mathcal{F}(\eta_i(t))\,\mathcal{F}(\eta_j(t))^*}{|\mathcal{F}(\eta_i(t))\,||\mathcal{F}(\eta_j(t))\,|}$$

で表される。

あらかじめ複数の音源座標 $\mathbf{s}$ に対するステアリングベクトル $\mathbf{a}_k(\mathbf{s}) = \left[e^{-j\omega_k d_1/c}, \ldots, e^{-j\omega_k d_N/c}\right]^T$ を求めておく。ここで k は短時間フーリエ変換の周波数 bin である (ω_k はその時の角周波数)。各 $\mathbf{s}$ に関して、$P_k(\mathbf{s}) = 1/\mathbf{a}_k^{\mathrm{H}}(\mathbf{s})\mathbf{R}_k\mathbf{a}_k(\mathbf{s})$ を計算する。H はエルミート転置である。

$\mathbf{s}$ が真の音源位置に近い場合に、$P_k(\mathbf{s})$ が小さくなるので、対象 bin $(k_L \le k \le k_H)$ にわたり平均化した

$$P(\mathbf{s}) = \frac{k_H - k_L}{\sum_{k=k_L}^{k_H} 1/P_k(\mathbf{s})} + \epsilon$$

が最小となる座標 $\mathbf{s}$ を音源位置の推定結果とする。

2.5 r-means 法

Eq. (3) の M を全ペア ($_N\mathrm{C}_2$ ペア) に拡張すると、

$$P(\mathbf{s}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\tau_{ij}^{sp} - \tau_{ij}^{obs} + \epsilon\right)^2 \qquad (4)$$

のようになる。τ_{ij}^{obs} は観測された到来時間差である。P を最小とする音源位置は解析的には解けないので、文献 [7] では補助関数 $\tilde{P}$ の反復法による最小化を行う。

$$\tilde{P}(\mathbf{s}, \tilde{\mathbf{s}}) = 2N\sum_{i=1}^{N}(\mathbf{s} - (\mathbf{r}_i + (\bar{r} + \bar{\tau}_i^{obs})\mathbf{e}_i))^2 + \mathrm{Const.}$$

ここで $\bar{\tau}_i^{obs} = \frac{1}{N}\sum_{j=1}^{N}\tau_{ij}^{obs}$ は観測値より定まるので、最適化に関係しない。$\tilde{\mathbf{s}} = \{\bar{r}, \mathbf{e}_1, \ldots, \mathbf{e}_L\}$ は補助変数である。$\tilde{P}$ の最小化条件を考えて更新式は、

$$\bar{r} \leftarrow \frac{1}{N}\sum_{i=1}^{N} d_i,\ \mathbf{e}_i \leftarrow \frac{\mathbf{s} - \mathbf{r}_i}{d_i} \qquad (5)$$

$$\mathbf{s} \leftarrow \frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i + (\bar{r} + \bar{\tau}_i^{obs})\mathbf{e}_i) \qquad (6)$$

のようになる。Eq. (5) で音源の方向を探索し、Eq. (6) で音源の位置を更新する。この更新には多数の繰り返しが必要なので、加速法を用いる [7]。

Fig. 1 にアルゴリズムを模式的に示す。センサから音源位置に向かうベクトルが $\mathbf{e}_i$ であり、$\mathbf{e}_i$ を更新し音源位置を同定する。音源を囲むように配置した分散マイクアレイ (Fig. 1(a)) では、センサごとの $\mathbf{e}_i$ に角度差がついており、収束が期待される。これに対し、直線マイクアレイ (Fig. 1(b)) では、センサごとの $\mathbf{e}_i$ に角度差がつかず、収束性はよくないと予想される。

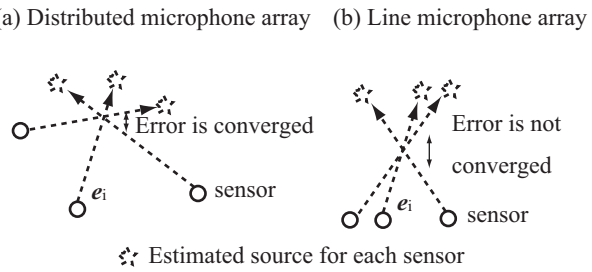

Fig. 1 Method of updating source position in r-means localization.

3 コスト関数の一般化とテンプレート法

M-CSP 法以外の上記手法は、M' 組のマイク対 $\varphi(m)$ における Eq. (7) のコスト関数を最小化する問題といえる。

$$P(\mathbf{s}) = \sum_{m=1}^{M'}\left|\tau_{\varphi(m)}^{ref} - \tau_{\varphi(m)}^{obs} + (\epsilon_L + \epsilon_E)\right|^{\kappa} \qquad (7)$$

ここで $\tau_{\varphi(m)}^{obs}, \tau_{\varphi(m)}^{ref}$ は、何らかの手法による観測および参照到来時間差で、各手法で何にあたるかを Table 1 にまとめた。κ は距離の次元である。ϵ_L はマイクや音源の配置誤差、ϵ_E は τ^{obs} の推定誤差である。

本報では既存法に加えて、配置誤差 ϵ_L を減らすため τ^{ref} を τ^{sp} の代わりにインパルス応答から求まる時間差 τ^{imp} を用いた手法を、テンプレート法 I として実験している。また推定誤差 ϵ_E も考慮するために、τ^{ref} として、ある話者がそれぞれの地点で発話した際に得られる遅れ時間 $\tau_{ij}^{csp(ref)}$ を用いた手法を、テンプレート法 II としている。

Table 1 φ, $\tau_{\varphi(m)}^{ref}$, $\tau_{\varphi(m)}^{obs}$, and κ for CSP, 2D-CSP, and r-means method.

	φ	$\tau_{\varphi(m)}^{ref}$	$\tau_{\varphi(m)}^{obs}$	κ	Eq.
CSP	1 pair	τ_{ij}^{pl}	τ_{ij}^{csp}	2	(2)
2D-CSP	M pair	τ_{ij}^{sp}	τ_{ij}^{csp}	2	(3)
r-means	all pair	τ_{ij}^{sp}	any	2	(4)
templateI	any	τ^{imp}	any	any	(7)
templateII	any	$\tau_{ij}^{csp(ref)}$	any	any	(7)

4 距離推定精度の検証

4.1 実験条件

音源をマイクアレイに対して {30,60,90,120,150} °、{50,100,150,200,300}cm の 25 地点に設置して、Fig. 2 のように配置したマイクによりインパルス応答を測定した。角度 D[°]、距離 R[cm] の点を D{D}R{R} と呼ぶ。機器操作用語の音声に、インパルス応答を畳み込み評価データを作成した。実験室のオールパスの残響時間 T_{20} は 0.58 秒である。本報ではクリーン

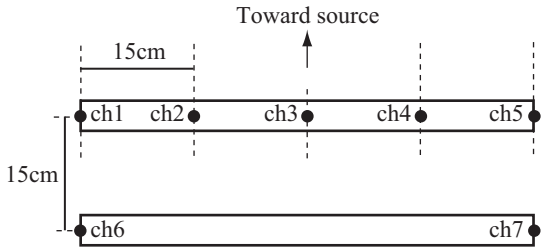

Fig. 2　Microphone array settings.

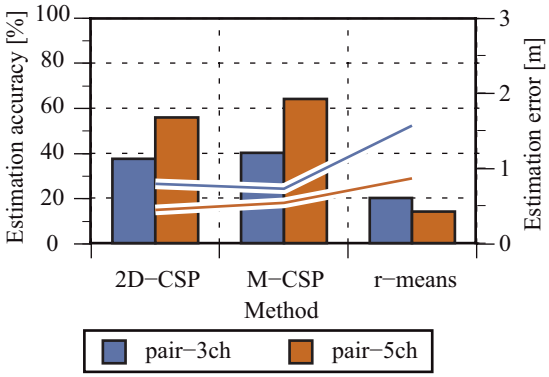

Fig. 3　Average estimation accuracy[%] (bar) (tolerance is ±25%) and estimation error[m] (line).

とエアコンの騒音 (12 dB) の場合を示す。サンプリング周波数は 16 kHz、短時間フーリエ変換の窓長は 60 ms、フレームシフトは 30 ms とし、150 Hz から 8 kHz の帯域を利用した。距離の候補点は上記 25 地点とした。ch1-2-5(pair-3ch) の 3 マイクと、ch1-3-5-6-7(pair-5ch) の 5 マイクの結果を比較する。比較手法に共通で $\varphi = \text{allpair}$, $\tau^{obs}_{\varphi(m)} = \tau^{csp}_{ij}$, $\kappa = 1$ である。

4.2　結果と考察 (クリーン環境)

4.2.1　既存手法の比較

距離推定の性能を、推定精度 (25%許容誤差)(棒グラフ)[%] と平均絶対値誤差 (折れ線グラフ)[m] の 2 つの尺度で評価した。前者が遠方で、後者は近傍で有利な指標であるため両方評価した。各地点での平均を Fig. 3 に示す。M-CSP 法が最も性能が高かった。M-CSP 法は 2D-CSP 法よりは、推定精度が平均的には向上するものの、推定できない点は多く存在する。r-means 法は、ほとんど推定できておらず、補助変数 $\mathbf{e}_i$ の初期値に定位結果が依存した。これの推定が原理的に難しいことは、例えば pair-3ch で $\mathbf{e}_i$ の y 成分の初期値を 0 とした時、マイクアレイが x 軸上にあるため、更新を経ても 0 のままであることからわかる。

4.2.2　方向、距離の違いによる評価値の比較

Eq. (7) において、τ^{obs}_{ij} を τ^{csp}_{ij}、τ^{ref}_{ij} を τ^{sp}_{ij}(図中 sp)、τ^{pl}_{ij}(図中 pl) とした場合の P の値を、Fig. 4 に示す。pair-5ch である。角度ごとの P の差は大きいが、距離ごとの P の差は小さい。また 50 cm の場合を除き、τ^{sp}_{ij}、τ^{pl}_{ij} の差異は小さい。

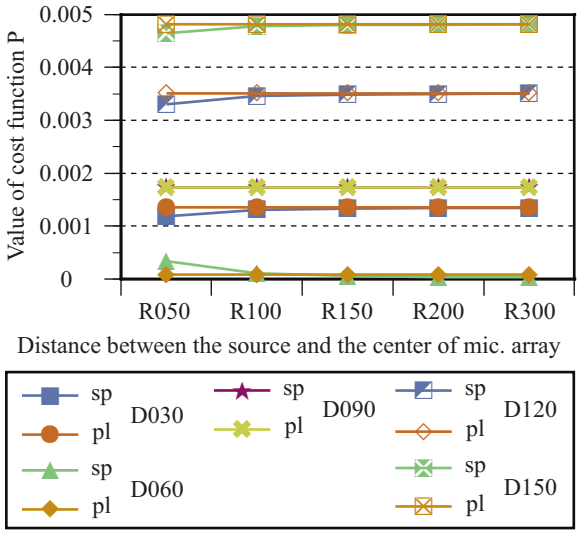

Fig. 4　Value of cost function P calculated by Eq. (7). Source is located at “D060R300”. ($\tau^{obs}_{ij} = \tau^{csp}_{ij}$ for all. $\tau^{ref}_{ij} = \tau^{sp}_{ij}$ for sp and $\tau^{ref}_{ij} = \tau^{pl}_{ij}$ for pl.)

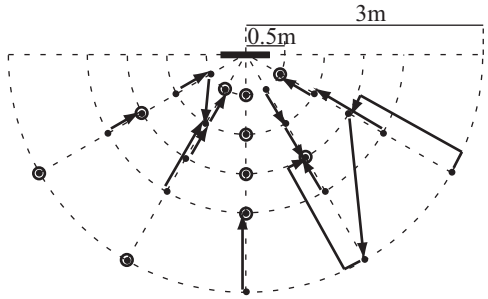

Fig. 5　Relationship between source and localized point under spherical wave assumption. (source)→(localized point)

4.2.3　インパルス応答から求まる時間差との比較

以上の既存手法で、推定精度の悪い点が多かった。$\epsilon_L \approx 0$ であればインパルス応答から得られる時間差は理論値に近づき、定位誤りは起きないはずなので、配置誤差 ϵ_L が大きいと考えられる。そこでインパルス応答から得られる時間差を用いて、理論値と比較した。すなわち Eq. (7) で、$\tau^{ref}_{\varphi(m)} = \tau^{sp}_{ij}$, $\tau^{obs}_{\varphi(m)} = \tau^{imp}_{ij}$ とした。τ^{imp}_{ij} は、i, j のインパルス応答の相互相関関数が最大となる遅れ時間である。結果を Fig. 5 に示す。図中矢印は、音源位置から定位位置に向いている。角度の誤りは少ないが、距離に関しては誤りが多い。遠方程定位誤りが増加している訳ではないので、遠方で平面波に近くなるわけでもない。ϵ_L の誤差の影響で、理論値との比較では距離の推定が難しい。

4.2.4　学習データによるテンプレート法

テンプレート法 I による結果を Fig. 6 に示す。平均的には推定精度が向上したが、これまでの手法と同じように推定精度が極端に低い点が存在する。これはまだ推定誤差 ϵ_E が含まれるためと考えられる。

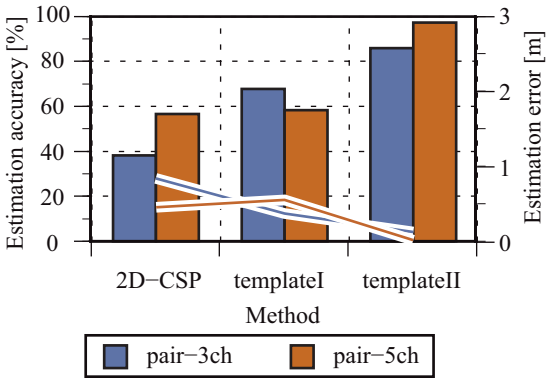

Fig. 6　Average estimation accuracy[%] and estimation error[m].

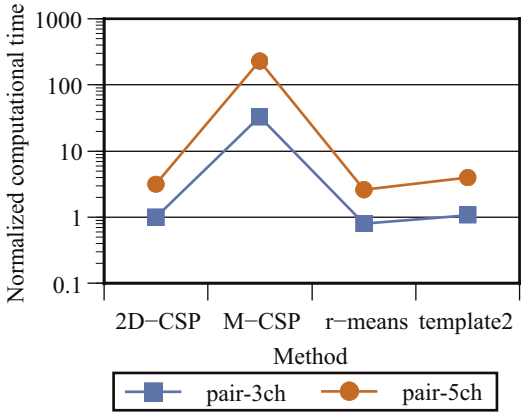

Fig. 7　Normalized computational time.

テンプレート法 II では、評価話者と異なる女性話者 1 名の各地点での 10 発話の音声区間部分の遅れ時間の平均 $\tau_{ij}^{csp(ref)}$ を参照時間差とした。結果を同じく Fig. 6 に示す。pair-3ch ではいくつかの点で推定精度が低いが、pair-5ch ではほとんどの点で 90%以上の推定精度となっている。これは学習により ϵ_L に加え ϵ_E も補正されたためである。

4.2.5　計算量の比較

上記手法の計算量比較を Fig. 7 に示す。2D-CSP 法の pair-3ch の場合の計算時間で規格化してある。M-CSP 法は、非常に計算量が大きい。r-means 法は、最も計算量が少ない。テンプレート法 II(テンプレート法 I も同程度) は 2D-CSP と同程度の計算量である。pair-5ch は pair-3ch と比して計算時間は 2 倍から 3 倍程度であり、おおむねペア数 (3→10) に比例する。

4.3　結果と考察 (騒音環境)

エアコンの騒音 (SNR=12 dB) 環境の結果を示す。テンプレート法 I,II を、2D-CSP 法と比較した。平均の推定精度を、Fig. 8 に示す。2D-CSP 法はクリーンな場合と比べて性能低下は少ない。テンプレート法 II の結果をコンターにしたものを、Fig. 9 に示す。多くの点で実用的に問題ない程度の誤差である。

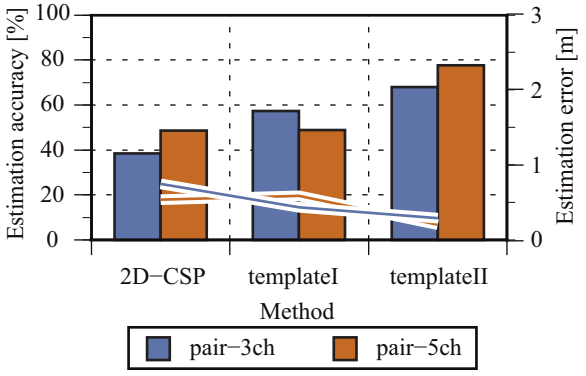

Fig. 8　Average estimation accuracy[%] and estimation error[m] . (Air Conditioner noise, SNR = 12 dB)

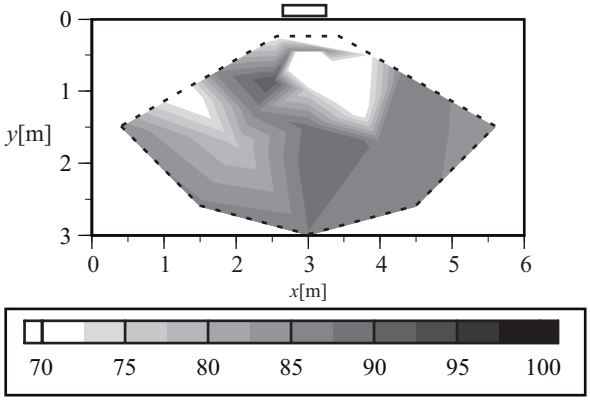

Fig. 9　Contour of estimation accuracy[%] (pair-5ch) (tolerance is ±25%).

5　まとめと今後の課題

音源距離推定の既存手法を、あるコスト関数を最小化する問題に一般化し、その推定精度を比較した。今回の実験では、従来法は十分な性能が出なかったが、これは測定の誤差が原因と考えられる。実測定では誤差は不可避のため、誤差を考慮したテンプレート法が有効であった。事前のテンプレート作成にはコストが掛かるので、今後は、テンプレート不要 (もしくは少ない測定点) で誤差調整できる手法を開発する。

参考文献

[1] Y. Tachioka *et al.*, *AST*, **33**, 68–71 (2012).
[2] R. Kennedy *et al.*, *IEEE Trans. on SP*, **46**, 2147–2156 (1998).
[3] C.H. Knapp *et al.*, *IEEE Trans. on ASSP*, **24**, 320–327 (1976).
[4] 西浦他, 信学論, **J83-D-II(7)**, 1610-1619 (2000).
[5] D.V. Rabinkin *et al.*, *Proc. of SPIE*, 88–99 (1996).
[6] 林田他, 信学技報, **EA2010-9**, 49–54 (2010).
[7] N. Ono *et al.*, *Proc. of ICASSP*, 2718–2721 (2010).