方位クラスタリングと非負値行列因子分解を用いた音像深度自動推定*

☆宮内 智, 北村大地, 猿渡 洋, 中村 哲 (奈良先端大)

1 はじめに

近年，臨場感のある音の再現を可能とする立体音響システムの構築を目指し，音場再現に関する研究が盛んに行われている．音場再現手法の一つとして，wave field synthesis (WFS) [1] がある．WFS は再現対象音源を任意の位置に定位させた場合の波面を物理的に再現し，音源の方位や深度（奥行き感）を受聴者に呈示する技術である．ここで，WFS の再現によって生じる波動の伝播は，音源の定位及び音源のスペクトル情報により記述され，これらの情報は事前に得る必要がある．

市場の大半を占める既存コンテンツであるステレオやマルチチャネルサラウンド用の混合音源を WFS で再現する場合，混合前の各音源に相当する一次音源の定位やスペクトル情報はミックスダウンの際に縮退してしまっているため，そのままでは再現を行うことができない．従って，既存コンテンツから各一次音源を分離した上で音源の定位情報を復元し，WFS で再現可能な音源を自動生成する手法の確立が望まれる．ここで問題となるのが音像深度の復元である．人間が音像の深度をどの様に知覚するかというメカニズムは未解明な部分が多く，音像深度推定に関しても手法が確立されていない．この問題に対し，我々は，direction of arrival (DOA)，すなわち音源の到来方位を特徴量として用いる音像深度推定手法を提案する．

本稿では，まず，WFS の原理及び混合音源からの一次音源分離手法である方位クラスタリングに基づく音源分離について述べる．次に，重み付き DOA ヒストグラムを用いた音像深度推定及び非負値行列因子分解 (nonnegative matrix factorization: NMF) [2] を用いた次元圧縮による改良手法を提案する．最後に，実音場で収録した深度の異なる音源に対して客観評価実験を行い，提案法の有効性を示す．

2 WFS の原理

WFS は，Kirchhoff-Helmholtz 積分方程式より，所望の一次音場を二次音場 (受聴空間) 上の二次音源により再現する．WFS による音源生成時の幾何学を Fig. 1 に示す．Figure 1 において x 軸に平行な破線は参照受聴線を意味し，この線上では合成された音場の振幅と位相が一次音源と理論的に一致する．

n 番目の二次音源に与えられる時間-周波数領域における駆動関数は式 (1) 及び式 (2) で表される [3]．ここで，j は虚数単位，$S_{Sn}(\omega,\tau)$ は n 番二次音源に与えられる各時間-周波数グリッドにおける駆動関数，$S_P(\omega,\tau)$ は一次音源のスペクトル情報，k は波数，τ は時間-周波数領域における時間フレーム，ω は角周波数，r_{PSn} は一次音源と n 番二次音源の距離，θ_{PSn} は境界線に対する法線と線分 r_{PSn} の成す角，sign(·)

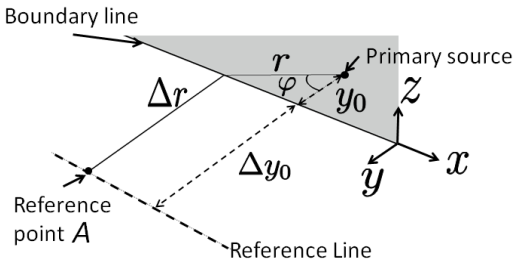

Fig. 1 Geometry of WFS.

$$S_{Sn}(\omega,\tau) = S_P(\omega,\tau) A \cos(\theta_{PSn}(\omega,\tau)) L \tag{1}$$

$$A = \sqrt{\frac{\mathrm{sign}(\zeta(\omega,\tau))\, k}{2\pi j}} \sqrt{\frac{\zeta(\omega,\tau)}{\zeta(\omega,\tau)-1}}$$

$$L = \frac{\exp\left(\mathrm{sign}(\zeta(\omega,\tau))\, jkr_{PSn}(\omega,\tau)\right)}{\sqrt{r_{PSn}(\omega,\tau)}} \Delta x$$

$$\zeta(\omega,\tau) = \frac{y_R}{y_P(\omega,\tau)} \tag{2}$$

は符号関数，Δx は二次音源間隔を示す．y_P は一次音源と二次音源の間の y 軸上における距離を示し，y_R は二次音源と参照受聴線の間の y 軸上における距離を示す．理論上，WFS には無限個の二次音源が必要となるが，実際にはこれは不可能であるため，有限個の二次音源で近似再現しており，これにより打ち切り誤差 [3] が発生し，WFS のスイートスポットに制限が生じる．

3 方位クラスタリングに基づく音源分離

混合音源中に含まれる一次音源の推定手法として，k-means 法を用いた方位クラスタリングに基づく音源分離 [4] を適用する．その概要を Fig. 2 に示す．今，既存のステレオ混合信号の時間周波数成分を $\boldsymbol{X}(\omega,\tau) = \left[X^{(L)}(\omega,\tau), X^{(R)}(\omega,\tau)\right]^{\mathrm{T}}$ とし，左チャネルの振幅 $|X^{(L)}(\omega,\tau)|$ と 右チャネルの振幅 $|X^{(R)}(\omega,\tau)|$ を軸とした二次元座標系で表現する．次に，全ての時間周波数成分を単位長さに正規化した後，一次音源数 V を指定し k-means クラスタリングを行う．そして，得られた各クラスタに含まれる成分のみ抽出したものを，各々の方位における一次音源スペクトル情報 $S_P(\omega,\tau)$ として WFS での再現に用いる．

4 提案手法

4.1 DOA に基づく音像深度推定

DOA は 電波や音波などの到来方位を意味するものであり，DOA を用いた音源分離手法が過去に提案されている [4]．前述の方位クラスタリングもその一つ

* "Automatic depth estimation of sound images using directional clustering and nonnegative matrix factorization," by Tomo Miyauchi, Daichi Kitamura, Hiroshi Saruwatari, Satoshi Nakamura (Nara Institute of Science and Technology).

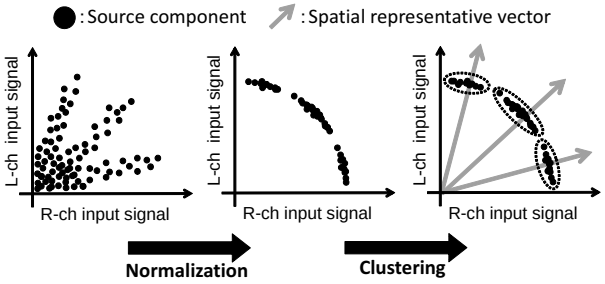

Fig. 2　Configuration of directional clustering.

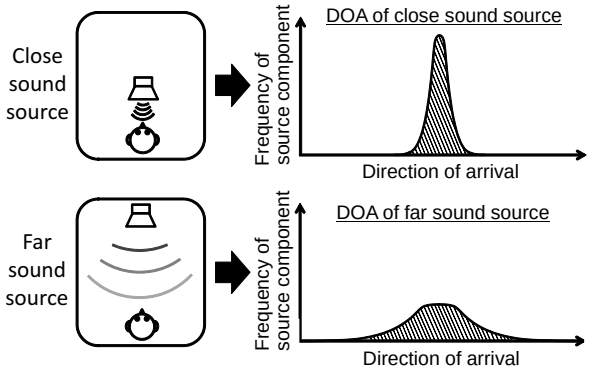

Fig. 3　Example of DOA histogram.

である．本研究では，この DOA を特徴量として解析することにより音像の深度推定を行うことを考える．人間の聴覚における音像の方位知覚において，両耳が受け取る音圧レベルの差は非常に重要な指標である．このことから，ステレオ信号 $\boldsymbol{X}(\omega,\tau)$ の左右チャネルにおける振幅比を用い，各時間周波数ビン毎の信号到来方位を $\theta = 2\arctan\left(|X^{(R)}(\omega,\tau)| \,/\, |X^{(L)}(\omega,\tau)|\right)$ として表現する．ここで，Fig. 2 に示す手法でクラスタリングを行った後，方位角に対する量子化を施し，各方位角ビンに存在する音源成分の個数をヒストグラムとして描く．通常，音の波面は音場を伝搬する中で方位方向に拡散が起こることを考慮すると，音源と受音点の距離に応じて DOA ヒストグラムの形状が異なると考えられる．即ち，Fig. 3 に示す様に，音源が近い場合，DOA ヒストグラムはある方位に急峻なピークを持ち，音源が遠くなる程に DOA ヒストグラムの裾野は拡がっていく．この定義を基に，音源から得られた DOA ヒストグラムの分散値が小さければ音源が近く，分散値が大きければ音源が遠くに存在するという定量的な評価尺度により音像深度推定を行う．

4.2　提案システムの処理フロー

WFS での再現のための，一次音源の分離及び音像深度推定に対する提案手法の信号フローを Fig. 4 に示す．前節で述べたように，ステレオ信号の DOA ヒストグラムを考え，その分散値を算出することにより音像深度の評価に行う．しかし，この手法をそのまま適用した場合，本来の信号成分がノイズに埋もれてしまったヒストグラムが形成され，真の分散の推定が困難になる．この問題は，信号を短時間フーリエ変換した際に生じる人工的なノイズが，クラスタリングの過程における正規化により本来の信号成分と同

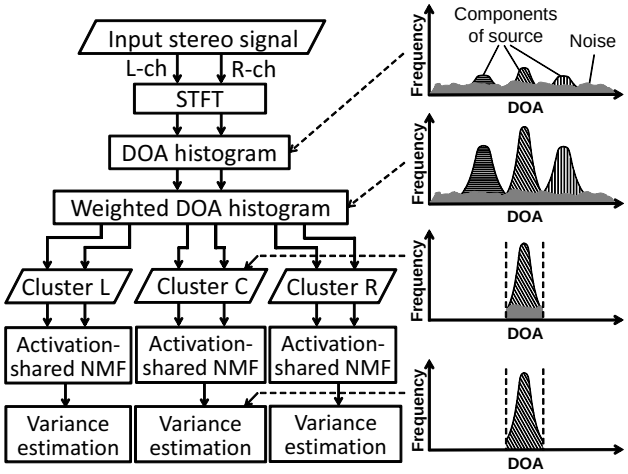

Fig. 4　Signal flow of proposed method.

等に扱われてしまうこと，そして元々の音源の録音環境における背景雑音がステレオ信号に収録されてしまっていることに起因する．以降では，この問題の改善手法として，重み付き DOA ヒストグラム及びアクティベーション共有型マルチチャネル NMF による次元圧縮法を提案する．

4.3　重み付き DOA ヒストグラム

人工的なノイズの影響を低減するために，重み付き DOA ヒストグラムを提案する．この手法では，ステレオ信号情報 $\boldsymbol{X}(\omega,\tau)$ のある時間周波数ビンにおける左右チャネルの振幅比から得られた方位 θ に対し，Fig. 2 に示す座標系における各音源成分のベクトル長に対応する $w = \left(|X^{(L)}(\omega,\tau)|^2 + |X^{(R)}(\omega,\tau)|^2\right)^{1/2}$ の重み付けを行いヒストグラムを作成する．これにより，本来の信号成分と人工的に生じたノイズとの差異化を図ることが可能となる．しかし，残存する人工的なノイズに加え，元信号に含まれる背景雑音を除去しなければ，一次音源の真の分散を推定することは出来ない．そこで，次節にて，これらのノイズの抑圧手法として提案するアクティベーション共有型マルチチャネル NMF について述べる．

4.4　アクティベーション共有型マルチチャネル NMF

NMF は信号のスパース表現手法であり，主成分分析や独立成分分析などの多変量解析手法と同様，個々の基底が観測データを構成する基本要素成分となるように学習することが目的となる．スパース表現の特質は，スパース性の制約により観測データの中に混在する特徴的なパターンが個々の基底となって表現される点にある．この性質を利用し，DOA ヒストグラムの分散を推定するに当たり問題となる人工的なノイズや背景雑音を抑圧することを考える．しかし，通常の NMF をステレオチャネルに別々にかけた場合，音源の空間情報である左右の振幅比が崩れてしまい正しい DOA ヒストグラムが算出できない．そこで，我々は左右チャネルのアクティベーションを共有した，音源の空間情報を保存しながら分解表現する NMF を提案する．次節において，その基底およびアクティベーションの更新式の導出について述べる．

4.4.1 アクティベーションの共有

M チャネルのマルチチャネル信号を観測したとき，m 番目のチャネルのスペクトログラムを $\boldsymbol{Y}^{(m)}(\in \mathbb{R}_{\geq 0}^{\Omega\times T})$ とし，以下のような NMF 分解モデルを考える．

$$\boldsymbol{Y}^{(m)} \simeq \boldsymbol{F}^{(m)}\boldsymbol{G} \quad (m = 1,\ 2,\ \cdots,\ M) \tag{3}$$

ここで，$\boldsymbol{F}^{(m)}(\in \mathbb{R}_{\geq 0}^{\Omega\times K})$，各チャネルに対応する基底行列であり，$\boldsymbol{G}(\in \mathbb{R}_{\geq 0}^{K\times T})$ は全てのチャネル間で共有するアクティベーション行列である．また，Ω は観測スペクトログラムの周波数ビン数，K は分解する基底数，T は観測スペクトログラムの時間フレーム数を示す．全てのチャネルの信号はアクティベーションを共有するため，マルチチャネル信号の空間情報 (振幅比) は保存しつつ基底分解を行うことができる．

4.4.2 β-divergence を用いたコスト関数

β-divergence は Eguchi らによって考案された距離尺度であり，あるデータ y に対する変数 x の一般的なダイバージェンスとして次の様に定義されている [5].

$$\mathcal{D}_\beta(y\|x) = \begin{cases} \dfrac{y^\beta}{\beta(\beta-1)} + \dfrac{x^\beta}{\beta} - \dfrac{yx^{\beta-1}}{\beta-1} & (\beta \in \mathbb{R}_{\setminus\{0,1\}}) \\ y(\log y - \log x) + x - y & (\beta \to 1) \\ \dfrac{y}{x} - \log\dfrac{y}{x} - 1 & (\beta \to 0) \end{cases} \tag{4}$$

コスト関数は β-divergence を用いて以下で表される．

$$\mathcal{J}(\Theta) = \textstyle\sum_m \mathcal{D}_\beta(\boldsymbol{Y}^{(m)}\|\boldsymbol{F}^{(m)}\boldsymbol{G}) \tag{5}$$

ここで，$\Theta = \left\{\boldsymbol{F}^{(m)}, \boldsymbol{G}\right\}$ は変数の集合である．このコスト関数を最小化する変数 Θ を，[6] と同様に補助関数法により求める．(4) 式を用いて (5) 式を書き改めると，以下のようになる．

$$\mathcal{J}(\Theta) = \textstyle\sum_{m,\omega,t}\left[\dfrac{\left(\sum_k f_{\omega,k}^{(m)} g_{k,t}\right)^\beta}{\beta} - \dfrac{y_{\omega,t}^{(m)}\left(\sum_k f_{\omega,k}^{(m)} g_{k,t}\right)^{\beta-1}}{\beta-1}\right] \tag{6}$$

ここで，$y_{\omega,t}^{(m)}$，$f_{\omega,k}^{(m)}$，及び $g_{k,t}$ はそれぞれ $\boldsymbol{Y}^{(m)}$，$\boldsymbol{F}^{(m)}$，及び $\boldsymbol{G}$ の要素値である．また，式中に現れる定数項は省略して表記している．

4.4.3 補助関数の設計

式 (6) の上限を与える補助関数を設計し，式 (6) を間接的に最小化する補助関数法を用いて，各変数の更新式を導出する．まず，式 (6) 中の第一項に対する補助関数を設計する．この項は $\beta \geq 1$ において凸関数であり，$\beta < 1$ において凹関数となる．従って，凸関数に対しては Jensen の不等式，凹関数に対しては接線不等式を用いて補助関数を設計する．$\beta \geq 1$ のとき，式 (6) 中の第一項の上限は Jensen の不等式及び補助変数 $\alpha_{\omega,t,k}^{(m)} \geq 0$，を用いると次のようになる．但し，補助変数は $\sum_k \alpha_{\omega,t,k}^{(m)} = 1$，を満たす．

$$\begin{aligned}\frac{1}{\beta}\left(\textstyle\sum_k f_{\omega,k}^{(m)} g_{k,t}\right)^\beta &\leq \frac{1}{\beta}\textstyle\sum_k \alpha_{\omega,t,k}^{(m)}\left(\dfrac{f_{\omega,k}^{(m)} g_{k,t}}{\alpha_{\omega,t,k}^{(m)}}\right)^\beta \\ &\equiv \mathcal{Q}_{\omega,t}^{(\beta)(m)}\left(\theta,\hat{\theta}\right)\end{aligned} \tag{7}$$

ここで，$\hat{\Theta}$ は補助変数の集合であり，式 (7) の等号成立条件は以下で与えられる．

$$\alpha_{\omega,t,k}^{(m)} = \frac{f_{\omega,k}^{(m)} g_{k,t}}{\sum_{k'} f_{\omega,k'}^{(m)} g_{k',t}} \tag{8}$$

また $\beta < 1$ のとき，式 (6) 中の第一項の上限は接線不等式及び補助変数 $\gamma_{\omega,t}^{(m)} \geq 0$ を用いると以下となる．

$$\begin{aligned}\frac{1}{\beta}\left(\textstyle\sum_k f_{\omega,k}^{(m)} g_{k,t}\right)^\beta &\leq \gamma_{\omega,t}^{(m)\beta-1}\left(\textstyle\sum_k f_{\omega,k}^{(m)} g_{k,t} - \gamma_{\omega,t}^{(m)\beta-1}\right) + \frac{\gamma_{\omega,t}^{(m)}}{\beta} \\ &\equiv \mathcal{R}_{\omega,t}^{(\beta)(m)}\left(\Theta,\hat{\Theta}\right)\end{aligned} \tag{9}$$

ここで，式 (9) の等号成立条件は以下で与えられる．

$$\gamma_{\omega,t}^{(m)} = \textstyle\sum_{k'} f_{\omega,k'}^{(m)} g_{k',t} \tag{10}$$

次に，式 (6) 中の第二項に対する補助関数を設計する．この項は $\beta \geq 2$ において凹関数であり，$\beta < 2$ において凸関数となる．式 (7) から式 (9) までと同様にして，補助関数は以下に定義できる．

$$\begin{aligned}&-\frac{1}{\beta-1} y_{\omega,t}^{(m)}\left(\textstyle\sum_k f_{\omega,k}^{(m)} g_{k,t}\right)^{\beta-1} \\ &\quad\leq \begin{cases} -y_{\omega,t}^{(m)}\,\mathcal{Q}_{\omega,t}^{(\beta-1)(m)}(\theta,\hat{\theta}) & (\beta < 2) \\ -y_{\omega,t}^{(m)}\,\mathcal{R}_{\omega,t}^{(\beta-1)(m)}(\theta,\hat{\theta}) & (\beta \geq 2) \end{cases}\end{aligned} \tag{11}$$

以上より，コスト関数式 (6) に対する補助関数 $\mathcal{J}^+$ は以下のように設計できるため，コスト関数式 (6) の代わりに補助関数式 (12) を最小化する θ を求める．

$$\mathcal{J}(\theta) \leq \mathcal{J}^+(\Theta,\hat{\Theta}) = \textstyle\sum_{\omega,t} \mathcal{S}_{\omega,t}^{(\beta)(m)}(\Theta,\hat{\Theta}) \tag{12}$$

ここで，$\mathcal{S}_{\omega,t}^{(\beta)(\mathrm{m})}$ は次式で与えられる．

$$\mathcal{S}_{\omega,t}^{(\beta)(\mathrm{m})} = \begin{cases} \mathcal{R}_{\omega,t}^{(\beta)(\mathrm{m})} - \mathcal{Q}_{\omega,t}^{(\beta-1)(\mathrm{m})} & (\beta < 1) \\ \mathcal{Q}_{\omega,t}^{(\beta)(\mathrm{m})} - \mathcal{Q}_{\omega,t}^{(\beta-1)(\mathrm{m})} & (1 \leq \beta < 2) \\ \mathcal{Q}_{\omega,t}^{(\beta)(\mathrm{m})} - \mathcal{R}_{\omega,t}^{(\beta-1)(\mathrm{m})} & (2 \leq \beta) \end{cases} \tag{13}$$

4.4.4 更新式の導出

補助関数式 (12) を最小化する補助変数 $\hat{\theta}$ は等号成立条件として式 (8) 及び式 (10) で与えられる．従って，先の補助関数を目的変数のそれぞれについて偏微分し，各補助変数の等号成立条件を代入することで，目的変数を最適化する反復型更新式が導出される．以下に導出した結果を示す．

$$f_{\omega,k}^{(m)} \leftarrow f_{\omega,k}^{(m)}\left(\frac{\sum_t y_{\omega,t}^{(m)} g_{k,t}\left(\sum_{k'} f_{\omega,k'}^{(m)} g_{k',t}\right)^{\beta-2}}{\sum_t g_{k,t}\left(\sum_{k'} f_{\omega,k'}^{(m)} g_{k',t}\right)^{\beta-1}}\right)^{\varphi(\beta)}, \tag{14}$$

$$g_{k,t} \leftarrow g_{k,t}\left(\frac{\sum_{m,\omega} y_{\omega,t}^{(m)} f_{\omega,k}^{(m)}\left(\sum_{m,k'} f_{\omega,k'}^{(m)} g_{k',t}\right)^{\beta-2}}{\sum_{m,\omega} f_{\omega,k}^{(m)}\left(\sum_{m,k'} f_{\omega,k'}^{(m)} g_{k',t}\right)^{\beta-1}}\right)^{\varphi(\beta)} \tag{15}$$

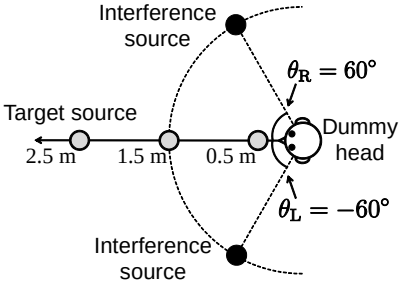

Fig. 5　Sound source geometry used in experiment.

ここで，$\varphi_{(\beta)}$ は次式で与えらえる．

$$\varphi_{(\beta)} = \begin{cases} 1/(2-\beta) & (\beta < 1) \\ 1 & (1 \le \beta \le 2) \\ 1/(\beta - 1) & (2 < \beta) \end{cases}. \qquad (16)$$

5　評価実験

5.1　実験条件

提案法の評価を目的として行った客観評価実験の概要を Fig. 5 に示す．評価には受聴者から見て左，正面，右の 3 方向に音源を配置し混合したテスト音源を用いる．この際，左右に配置する干渉音の距離は受聴者から見て 1.5 m に固定する一方で，目的音の距離は受聴者の正面 0.5 m, 1.5 m, 2.5 m の 3 位置に定位させ，各位置における分散の評価を行う．試験音には，Vocal, Guitar, Piano の 3 種類を用意し，図中の各音源位置でバイノーラル録音した室内インパルス応答を畳み込んで音源を作成した．また，背景雑音に対する提案法の頑健性を評価するために，ピンクノイズを混合したテスト音源に対しても評価を行った．この際，信号対雑音比は 10, 20, 30 dB に設定した．3 種類の試験音の方位の組み合わせ 6 通り，目的音の距離の違う 3 通り，ノイズ付加率の違いによる 4 通りを各々設定した計 72 個のテストセットを用い評価を行った．また，NMF のパラメータは β を 1, 基底数 K を 30 とした．

5.2　実験結果

Figure 6 に，ノイズの有無による 4 つの条件において，DOA ヒストグラムに重み付けを施した提案法 1 と，重み付けに加え NMF を施した提案法 2 の実験結果の平均値を示す．各条件において，最近傍距離である 0.5 m 地点で算出された DOA ヒストグラムの分散値を 1 に正規化し，それに対比する 1.5 m，2.5 m の分散値がどの様に推定されているかを評価した．図から，提案法及び提案法 2 ともに音源深度が大きくなるに従い分散値も大きくなっており，距離間で有意な推定が行われていることが分かる．また，全条件において NMF を用いた提案法 2 の方が若干優位であり，ノイズの強い音源における頑健性が示唆された．以上より，一次音源情報の音像深度推定問題に対する提案法の有効性が示された．

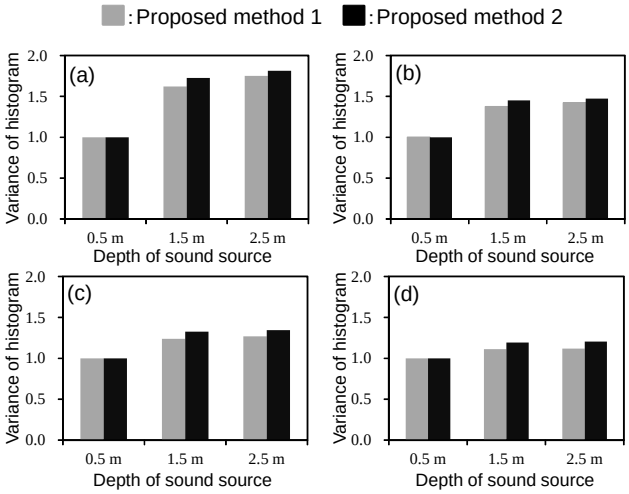

Fig. 6　Average scores for test source: (a) signal, (b) signal with pink noise (SNR=30 dB), (c) signal with pink noise (SNR=20 dB), (d) signal with pink noise (SNR=10 dB).

6　まとめ

本稿では混合音源中の音像深度推定問題に対し，音源の DOA ヒストグラムの分散値を評価することを考え，重み付け DOA ヒストグラム及びアクティベーション共有型マルチチャネル NMF を提案した．客観評価実験を行った結果，音源の距離の違いによる DOA ヒストグラム分散値の違いを有意に評価出来ており，音像深度推定問題における提案法の有効性が示された．

参考文献

[1] A. J. Berkhout, “A holographic approach to acoustic control,” *J. Audio Eng. Soc.*, vol.36, no.12, pp.977–995, 1988.

[2] D. D. Lee, H. S. Seung, “Algorithms for non-negative matrix factorization," *Neural Info. Process. Syst.*, vol.13, pp.556–562, 2001.

[3] E.Verheijen, “Sound reproduction by wave field synthesis,” *PhD. Thesis*, TU Delft, 1998.

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

[5] S. Eguchi, Y. Kano, “Robustifying maximum likefood estimation," *Technical Report of Institute of Statistical Mathematics*, 2001.

[6] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, S. Sagayama, “Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence," *Proc. MLSP*, 2010.