音声スパース性と信頼度付時間差検出に基づく残響環境下での 2chBSS *

◎和泉洋介，小野順貴，嵯峨山茂樹 (東大情報理工)

1 はじめに

音声のスパース性を利用したブラインド音源分離 (BSS)[1, 2, 3] は，観測信号数より音源信号数が多くても適用可能でありマイクロフォン数が少ない場合に特に有効な分離手法であるが，実環境での適用のためには，残響環境下での性能向上が課題の 1 つとなっていた。本稿では先の報告 [4] にひきつづき 2ch の BSS に議論を絞り，残響環境下における具体的な音源定位と分離手法を提案し，実験によりその性能を検証した結果について報告する。

2 スパース性に基づく 2chBSS の問題設定

以下では，2 個のマイクロフォンにより観測された信号の時間周波数表現を $\mathbf{M} = (M_{Li}(\omega), M_{Ri}(\omega))$ と表す。ただし，i はフレーム番号，ω は角周波数，L, R の添え字はそれぞれ，左右のマイクロフォンで取得された信号であることを表す。以下では簡単のため，フレーム番号 i と角周波数 ω は省略して表記する。

音声のスパース性から，各時間周波数成分において観測信号に寄与する音源は 1 個だけであると仮定すると観測モデルは，

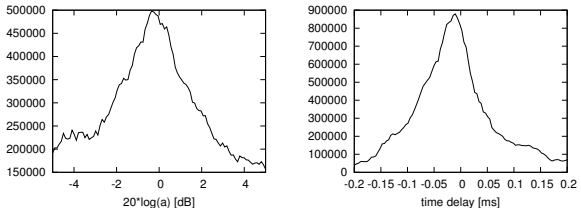
$$\boldsymbol{M} = S_n \boldsymbol{b}_n + \boldsymbol{N} \tag{1}$$

のように表される。S_n は推定したい音源信号，N は残響などに起因する観測誤差，$\boldsymbol{b}_n$ は音源 n の位置に対応するベクトルを表す。ここで音源信号の振幅の任意性を除くため，$\boldsymbol{b}_n$ は $|\mathbf{b}_n| = 1$ のように規格化されているものとする。

スパース性に基づく音源分離でよく行なわれる手法は，1) 強度比・時間差のクラスタリングに基づく音源定位 (式 (1) における $\boldsymbol{b}_n (n = 1, \cdots, N$ の決定)，2) 時間周波数マスキングによる音源分離 (寄与する音源 n の，時間周波数成分毎の推定) の 2 段階からなる。しかしながら残響環境下においては，強度比や時間差は大きな分散を生じ，1) の定位自体が困難なものとなっていた。

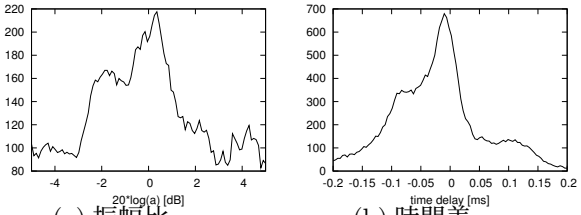
これに対し我々は，1) 残響を拡散音場とみなすことによる誤差モデルの導入，2) 連続したフレームの観測区間毎の信頼度付での強度比・時間差検出，により，信頼度の高い推定結果が真値のまわりにより密に分布し，音源定位を改善できる見込みを得た [4]。本稿ではこれに基づき，1) 観測区間毎の信頼度付での強度比・時間差検出の統合による具体的な音源定位法，2) 音源定位後の音源分離法，の 2 点について論じる。

3 観測区間毎の信頼度付検出の統合による音源定位



(a) 振幅比 a (b) 時間差 τ

Fig. 1 観測信号のパワーで重みづけした特徴量のヒストグラム。残響 ($T_R = 376$ms) の影響で 3 個の音源があるのかも判断しづらい。



(a) 振幅比 a (b) 時間差 τ

Fig. 2 信頼度で重みづけした特徴量のヒストグラム。(b) を見ると中央の顕著な音源に隣り合う小さな 2 つの音源がある。

複数の測定結果を統合する自然な方法の 1 つはヒストグラムであり，従来用いられてきた手法は，観測フレームのパワーを重みとして強度比 a，時間差 τ の 2 次元の特徴空間でヒストグラムを生成するものであった [1]。これに対し我々は [4] で提案した信頼度付の検出をいかし，信頼度を重みとしたヒストグラムを作成した。その比較の例を Fig.1，Fig.2 に示す。

従来法である Fig.1 に比べ，Fig.2 では時間差 τ における断面では音源方向に依存してピークが若干みられるものの，特に強度比ではピークを検出することは難しい。これは信頼度の値の範囲が大きいこと，ヒストグラムにおいては原点近傍 ($a = 0$，$\tau = 0$) のピークが特に大きくなるバイアスのような効果がみられること，などが原因であり，この統合方法はまだ改善の余地があると考えられるが，ここでは，時間差 τ の方がピークを形成しやすい性質に着目し，

1. 音源数 N は既知として与え，
2. τ のみのヒストグラムに対し，EM アルゴリズムを用いて混合正規分布をフィッティングすることで，小さなピークの検出を可能にし，各音源に対する時間差 $\tau_n (n = 1, \cdots, N)$ を求め，

---

* 2ch BSS based on speech sparseness and detection of time delay with reliability value in a reverberant environment. by IZUMI, Yosuke, ONO, Nobutaka and SAGAYAMA, Shigeki (Graduate School of Information Science and Technology, the Univ. of Tokyo)

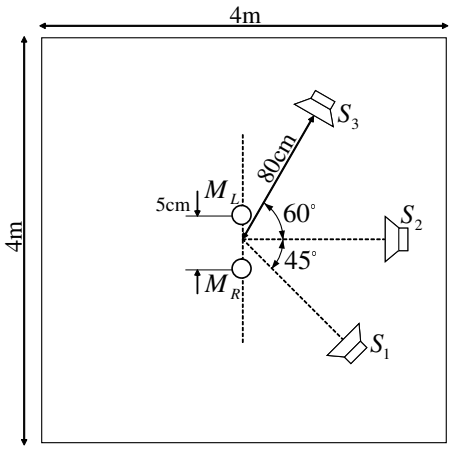


Fig. 3　シミュレーションにおけるマイクロフォンと音源の位置関係

3. 各 τ_n の近傍の値をとる点のみから振幅比-信頼度のヒストグラムを作成し，そのピークを音源位置に対応する振幅比 a_n として求める

という手法をとった。紙面の都合上，詳細は発表にて示す。

4　拡散音場モデルに基づくバイナリマスキングとビームフォーミング

音源の分離方法として，最尤法に基づくバイナリマスキングを行なうことを考える。拡散音場モデルに基づき誤差ベクトルの共分散行列を V とすると，式 (1) のようにある観測ベクトル $\boldsymbol{M}$ が音源 n に帰属すると仮定に基づく対数尤度 L_n と推定値 S_n は，

$$L_n = C - \frac{|\hat{\boldsymbol{b}}_n^h V^{-1}\boldsymbol{M}|^2}{\hat{\boldsymbol{b}}_n^h V^{-1}\hat{\boldsymbol{b}}_n}, \quad S_n = \frac{\hat{\boldsymbol{b}}_n^h V^{-1}\boldsymbol{M}}{\hat{\boldsymbol{b}}_n^h V^{-1}\hat{\boldsymbol{b}}_n} \tag{2}$$

のように与えられる。ここで C は n に依らない定数，^{h} はエルミート転置であり，また $\hat{\boldsymbol{b}}_n$ は前節で推定した特徴量 (a_n, τ_n) を用いて，以下のように表される。

$$\hat{\boldsymbol{b}}_n = \frac{1}{1+a_n^2}\begin{pmatrix} 1 \\ a_n e^{j\omega\tau_n} \end{pmatrix} \tag{3}$$

上記の $\hat{S}_n$ は誤差の共分散行列を考慮した 2ch のビームフォーミングにほぼ等しく，単なるバイナリマスクよりも雑音の低減効果が期待できる。以上より，バイナリマスキングとビームフォーミングを組み合わせた S_n の推定値として，

$$\hat{S}_n = \begin{cases} S_n & (L_n > L_{n'} \text{ for } {}^\forall n' \neq n) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

を得る。

5　残響シミュレーションによる分離実験

Fig.3 のように 3 個の音源と 2 個のマイクロフォンを配置し，残響時間 $T_R = 376$ms の部屋のシミュレーション分離実験を鏡像法 [5] により行った。分離性能の評価には，分離前後の元音声に対する S/N 比の改善値を用いた。音声データは研究用連続音声データベース (©板橋秀一 [日本音響学会／編]1991Vol. 1-3) を使用し，音源信号の組合せを変えて 84 通りの実験を行った。サンプリング周期 16kHz，フレーム長は 2^{10}，シフトは 32，窓関数を Hamming 窓として，時間周波数表現し，中心フレームの前後 7 フレーム計 15 フレームから信頼度を検出した。

Table 1　音源位置推定結果

$(20\log_{10}(a), \tau)$	s_1	s_2	s_3
真の位置	(2.03, -0.10)	(0.00, 0.00)	(-2.49, 0.13)
推定値	(1.76, -0.11)	(-0.03, 0.01)	(-1.11, 0.08)
分散	(2.87 0.00)	(0.01, 0.00)	(1.20, 0.00)

Table 2　音源分離性能 (S/N 比改善値) の平均値 [dB]

	s_1	s_2	s_3
0dB マスク	6.9	7.3	8.1
提案法	9.2	3.8	10.5

音源位置推定結果を Table.1，分離結果を Table.2 に示す。比較対象とした 0dB マスクとは，音源信号が既知であるときに各時間周波数成分を支配的な音源 n に帰属させる，残響がない場合には理想的なバイナリマスクである [1]。提案法は残響環境下でも音源定位ができており，さらに s_1, s_3 については 0dB マスクよりも分離性能が高いことが確認できる。

6　結論

拡散音場モデルに基づく観測区間毎の信頼度付強度比・時間差検出結果の統合と，バイナリマスキングとビームフォーミングを組み合わせた分離による残響環境下での 2chBSS 手法を提案した。2 個のマイクロフォンに入る残響環境下でのシミュレーション実験により，優れた分離性能を示すことを確認した。信頼度付強度比・時間差検出結果の統合方法にはまだ改善の余地があり，今後検討していきたいと考えている。

謝辞　本研究の一部は科学研究費補助金・若手研究 (B)(課題番号 18760303) の補助を受けて行なわれたので，ここに謝意を表する。

参考文献

[1] O. Yilmaz *et al.*, IEEE Trans. on Signal Processing, Vol. 52, No. 7, pp 1830-1847, 2004.
[2] S. Winter *et al.*, Proc. SAPA2004, S1.3, 2004.
[3] A. Blin *et al.*, Proc. ICASSP, vol. IV, pp. 85-88, 2004 .
[4] 小野他, 音講論 (秋), 1-1-10 in CD-ROM, 9 月, 2006.
[5] J. B. Allen et al., JASA, vol. 65, no. 4, pp. 943–950, Apr. 1979.