高残響下で混合された音声の音源分離に関する研究*

☆礒佳樹（筑波大），荒木章子（NTT 研究所），牧野昭二（筑波大），
中谷智広，澤田宏（NTT 研究所），山田武志（筑波大），中村篤（NTT 研究所）

1　はじめに

ブラインド音源分離（BSS）とは，それぞれのセンサが観測した混合信号の情報のみを用いて，各音源の信号を推定，分離する手法である。今回扱う音声信号における BSS の技術は, ハンズフリーテレビ会議システムなど, 多くの応用が期待されている。

これまで，残響が少ない混合信号に対しては，独立成分分析による方法[1]や音声のスパース性を用いた方法[2]など，性能の良い方法が多数考案されている。しかし，残響が多く含まれる混合信号に対しては，無響を仮定している上記の手法をそのまま残響がある場合にも適用するなど，まだ発展途上の段階にある。本研究では残響を考慮した N.Q.K. Duong らによるモデルパラメータ推定による音源分離法[3]（以下「従来法 1」）について議論する。

従来法 1 において Duong らは，残響などの空間特性を含んだ音源信号である Source image を複素ガウス分布による確率変数であると仮定し，その共分散である各パラメータを推定，求められたパラメータから作成されるウィーナーフィルタによって各音源の分離を行うことを提案した。モデルパラメータを推定するために EM アルゴリズムが使われているが，EM アルゴリズムは初期値依存性が高いということが従来法 1 の問題点として挙げられる。従来法では観測信号の成分同士の距離が近いものをクラスタリングした結果を初期値としているが，十分な性能が得られていなかった。そこで本研究ではこのクラスタリングについて，残響を特に考慮していない澤田らの時間周波数バイナリマスクによる音源分離法 [4] （以下「従来法 2」）で得られた結果を初期値として用いることを検討する。

2　問題設定

2.1 混合系

実環境において，J 人の音声信号 $s_j(j=1,\ldots,J)$ が I 個のセンサで観測されたとすると，観測信号は畳み込みによって次のようにモデル化できる。

$$x_i(t)=\sum_{j=1}^{J}\sum_{\tau}h_{ij}(t)s_j(t-\tau)\quad(i=1,\ldots,I)\quad(1)$$

x_i はセンサ i による観測信号，h_{ij} は音源 j からセンサ i へのインパルス応答である。この各チャンネルを列ベクトルして表すと次のようになる。

$$\mathbf{x}(t)=\sum_{j=1}^{J}\sum_{\tau}\mathbf{h}_j(\tau)s_j(t-\tau),\qquad(2)$$
$$\mathbf{x}(t)=[x_1(t),x_2(t),\ldots,x_I(t)]^T,$$
$$\mathbf{h}_j=[h_{1j},h_{2j},\ldots,h_{Ij}]^T$$

BSS の目的は，観測信号 $\mathbf{x}$ の情報のみを用いて分離信号 $\mathbf{y}_j$ を得ることである。本研究では劣決定問題（I < J）について議論し，I と J は既知であると仮定する。

本研究では信号の時間周波数領域表現を用いる方法を採用する。時間周波数領域での音声信号は時間領域よりスパースであることや[5]，時間領域での畳み込みは各周波数で積に変形できることを利用するためである。時間周波数領域の観測信号は次のようにモデル化される。

$$\mathbf{x}(n,f)=\sum_{j=1}^{J}\mathbf{h}_j(f)s_j(n,f)\quad(j=1,\ldots,J)\quad(3)$$
$$=\sum_{j=1}^{J}\mathbf{c}_j(n,f)$$
$$\mathbf{c}_j(n,f)\approx\mathbf{h}_j(f)s_j(n,f)\qquad(4)$$

* Research on source separation of mixed speech in a high reverberation, by Keiju ISO (University of Tsukuba) and Shoko ARAKI (NTT), Shoji MAKINO (University of Tsukuba), Tomohiro NAKATANI, Hiroshi SAWADA (NTT), Takeshi YAMADA (University of Tsukuba) and Atsushi NAKAMURA (NTT).

ここで$\mathbf{h}_j(f)$は音源jから各センサへの伝達関数，$s_j(n,f)$及び$\mathbf{x}(n,f)$はそれぞれ短時間フーリエ変換された原信号と観測信号を表す。nは時間フレーム番号，fは周波数である。$\mathbf{c}_j(n,f)$は，残響などの空間特性を含んだ音源信号である Source image である。

2.2 分離処理

従来法 1 は，モデルパラメータ推定を用いて，Source image $\mathbf{c}_j(n,f)$を推定する[3]。この手法では，原信号$s_j(n,f)$に伝達関数$\mathbf{h}_j(f)$がかけられた Source image $\mathbf{c}_j(n,f)$が，共分散行列$\mathbf{R}_{\mathbf{c}_j}$を持つ複素ガウス分布による確率変数であると仮定する。すなわち，以下のように表すことができる。

$$\mathrm{p}\left(\mathbf{c}_j;\mathbf{R}_{\mathbf{c}_j}\right)=N\left(0,\mathbf{R}_{\mathbf{c}_j}\right) \tag{5}$$

また，(4)より共分散行列$\mathbf{R}_{\mathbf{c}_j}$は以下のように表すことができる。

$$\begin{aligned}\mathbf{R}_{\mathbf{c}_j}(n,f)&=\mathbf{c}_j(n,f)\mathbf{c}_j^H(n,f)\\&=\left|s_j(n,f)\right|^2\mathbf{h}_j(f)\mathbf{h}_j^H(f)\\&=v_j(n,f)\mathbf{R}_j(f)\end{aligned} \tag{6}$$

(6)より共分散行列$\mathbf{R}_{\mathbf{c}_j}$は時間依存の信号の分散$v_j$と音源の空間特性を表す時不変の共分散行列$\mathbf{R}_j$に分けることができる。このとき，観測信号が音源信号の直接音のみで構成される場合は共分散行列$\mathbf{R}_j$のランクが 1 になるが，従来法 1 においては残響を考慮し，フルランクであると仮定している[3]。

モデルパラメータ$\theta=\{v_j,\mathbf{R}_j\}$を以下の尤度関数を最大化することによって推定する。

$$\mathrm{P}\left(\{\mathbf{x}\},\{\mathbf{c}_j\},\theta\right)=\prod_n\prod_f\mathrm{P}\left(\mathbf{x}\mid\{\mathbf{c}_j\};\theta\right)\prod_j\mathrm{P}\left(\mathbf{c}_j\mid\theta\right) \tag{7}$$

この尤度関数から，EM アルゴリズムを用いてモデルパラメータv_jと$\mathbf{R}_j$と，Source image $\mathbf{c}_j$を交互に更新する。

EM アルゴリズムの各ステップの計算は次のようになる。ただし，モデルパラメータの初期値は以下のようにする。

$$\mathbf{R}_j^{\mathrm{init}}(f)=\frac{1}{\left|C_j\right|}\sum_{\mathbf{x}(n,f)\in C_j}\mathbf{x}(n,f)\mathbf{x}(n,f)^H \tag{8}$$

$$v_j^{\mathrm{init}}(n,f)=1 \tag{9}$$

ここで$\left|C_j\right|$はクラスタの要素数である。従来法 1 では，各フレーム番号の各周波数成分について，階層的クラスタリングを用いてC_jを求めている[3]。

E-step: 分離のためのウィーナーフィルタ$\mathbf{W}_j$及び分離音$\hat{\mathbf{c}}_j$を更新

$$\mathbf{W}_j(n,f)=\mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f) \tag{10}$$

$$\hat{\mathbf{c}}_j(n,f)=\mathbf{W}_j(n,f)\mathbf{x}(n,f) \tag{11}$$

$$\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f)=\hat{\mathbf{c}}_j(n,f)\hat{\mathbf{c}}_j^H(n,f)+\left(\mathbf{I}-\mathbf{W}_j(n,f)\right)\mathbf{R}_{\mathbf{c}_j}(n,f) \tag{12}$$

ここで，

$$\mathbf{R}_{\mathbf{c}_j}(n,f)=v_j(n,f)\mathbf{R}_j(f) \tag{13}$$

$$\mathbf{R}_{\mathbf{x}}(n,f)=\sum_{j=1}^{J}v_j(n,f)\mathbf{R}_j(f) \tag{14}$$

M-step: モデルパラメータv_jと$\mathbf{R}_j$を更新

$$v_j(n,f)=\frac{1}{I}\mathrm{trace}\left(\mathbf{R}_j^{-1}(f)\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f)\right) \tag{15}$$

$$\mathbf{R}_j(f)=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{v_j(n,f)}\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f) \tag{16}$$

以上のアルゴリズムの反復により分離音を(11)式にて求める。得られた分離音は逆フーリエ変換によって時間信号に戻す。

3 提案手法

従来法 1 が EM アルゴリズムの初期設定に階層的クラスタリングを用いているのに対し，提案手法では澤田らのノルム正規化した$\mathbf{x}(n,f)$の音源方向推定によるクラスタリングを用いた音源分離法[4]を適用する。この方法は残響の少ないときに高い分離性能を得られることが知られている*。この結果を EM アルゴリズムの初期値に用いて，フルランクモデルによるウィーナーフィルタ推定を行う。

アルゴリズムの流れを図でまとめると以下のようになる。尚，図中の Permutation については，文献[6]の方法を用いた。

* http://sisec.wiki.irisa.fr/tiki-index.php

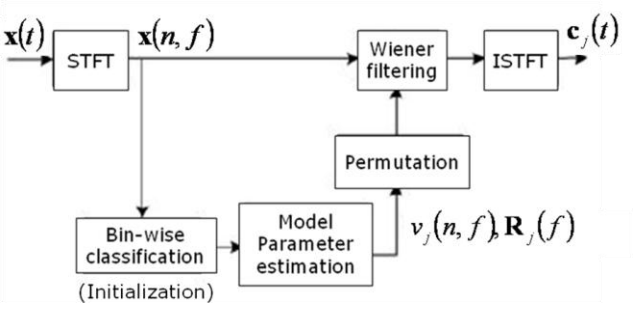

図 1：アルゴリズムの流れ

4　実験と評価

4.1 実験：二つの従来法との比較

提案法の有効性を検証するために，提案法の性能に対して，従来法 1（Duong らのモデルパラメータ推定による方法），従来法 2（澤田らの時間周波数バイナリマスクによる方法）の二つの性能をそれぞれ比較した。

実験では，3 話者音声の残響を含んだ 2 チャンネル混合信号に対して分離を行い，得られた分離信号を客観評価によって測定した。なお，客観評価値には文献[7]で提案された 4 つの歪み尺度を使用する。

- SDR (Signal to Distortion Ratio)
  ：総合的な歪み
- ISR (source Image to Spatial distortion Ratio)
  ：線形歪み
- SIR (Source to Interference Ratio)
  ：他話者音声の消し残りによる歪み
- SAR (Sources to Artifacts Ratio)
  ：非線形歪み

単位は dB であり，この歪み尺度の数値が高いほど性能が良い。

4.2 実験条件

A. 実験 1：従来法 1 との比較

SiSEC*2010 の音声データを使い，公表されている Duong らの方法による評価値との比較を行った。音声データの残響時間は 250ms，マイク間距離 5cm である。また，サンプリング周波数 16 kHz であり，FFT のサイズは 2048 とし，オーバーラップを 256 に設定した。提案法の反復回数は 20 回である。

B. 実験 2：従来法 2 との比較

3 人の話者による 8 秒，サンプリング周波数 8 kHz の混合された音声信号について分離を行った。FFT のサイズは 1024 とした。提案法の反復回数を 1, 5, 10, 20, 30, 40, 50 に設定して各反復回数の分離結果について評価し，また 6 種類の残響の長さ（130～450 ms）についてもそれぞれ検証を行った。尚，評価値は 6 通りの話者の組み合わせの平均値で算出している。

4.3 実験結果

実験 1 及び 2 の結果をそれぞれ表 1 及び表 2，図 2 に示す。

表 1：実験 1 の結果

Algorithm		SDR	ISR	SIR	SAR
従来法 1	音源 1	0.1	6.0	-0.5	7.8
	音源 2	5.2	7.6	9.1	11.6
	音源 3	3.0	4.8	5.7	8.0
従来法 2	音源 1	7.0	10.9	13.2	8.9
	音源 2	5.4	12.5	8.7	7.9
	音源 3	8.5	12.7	15.0	10.8
Proposed	音源 1	7.8	11.1	11.8	11.6
	音源 2	6.4	10.4	9.3	10.1
	音源 3	9.2	12.9	13.2	13.7

表 2：実験 2 の結果

	SDR	ISR	SIR	SAR
従来法 2	6.8	12.3	11.9	8.9
Proposed 1	5.4	8.7	6.0	13.4
5	5.9	9.4	8.0	11.44
10	7.1	10.9	10.1	11.36
20	7.72	12.1	11.1	11.65
30	7.68	12.5	11.2	11.76
40	7.8	12.82	11.4	11.83
50	7.9	12.84	11.5	11.80

* SiSEC: Signal Separation Evaluation Campaign (http://sisec.wiki.irisa.fr/tiki-index.php)

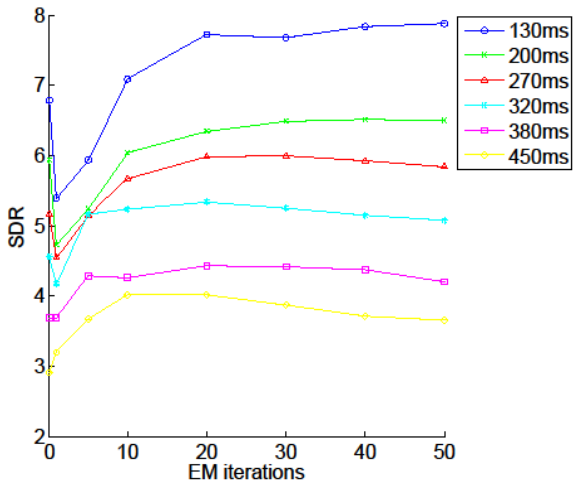

図 2：実験 2 における提案法の SDR

表 1 では，参考のため同条件での従来法 2 の結果も載せている。また表 2 における残響時間は 130 ms である。図 2 は，実験 2 における提案手法の各反復回数に対しての SDR の値の変化のグラフで，音声データの残響時間別に載せている。ここで反復回数 0 は従来法 2 による初期値の性能を指す。表 1，表 2（図 2）において，総合的歪みを表す SDR の値が共に上昇している。これにより，提案法の有効性が示された。

4.4 考察

ここでは二つの従来法に比べて提案法による性能がどのように改善したか考察する。

表 1 より，従来法 1 に対しては SDR の大幅な改善が見られ，今回提案した初期値設定が非常に有効であることが分かる。また表 2 より従来法 2 に対しては SIR の値が若干下がってはいるものの，SAR の値の大幅な上昇により結果として総合的歪みである SDR の改善が見られる。各歪み尺度とその意味合いの関係に当てはめると，従来法 2 に対しては分離性能を保ちながら音質が向上し，その結果総合的な性能が改善されていると考えられる。それぞれの手法による分離音を聞いたところ，この考察と一致した結果が得られていた。

また図 2 より，残響を考慮していない従来法 2 と比較して，残響時間が長い場合の SDR の改善の度合いが小さいことから，さらに性能が改善できるような適切なモデル化ができるのではないかと考えられる。

5 結論

本研究では，Duong らの残響を考慮したモデルパラメータ推定による音源分離法に，EM アルゴリズムの初期値設定として，残響を特に考慮していない澤田らの時間周波数バイナリマスクによる分離法を組み合わせた音源分離法の提案を行った。実験では Duong らの分離法と澤田らの分離法の両方に対して客観評価による比較を行い，提案法の有効性を示した。今後はさらに性能の改善を図るための検討を行う。

参考文献

[1] S. Makino, S. Araki, R. Mukai, and H. Sawada, “Audio source separation based on independent component analysis,” in *ISCAS 2004*, pp. V-668 – V-671, May 2004.

[2] S. Araki, H. Sawada, R. Mukai, and S. Makino, “Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors,” in *Signal Processing, vol. 87*, pp. 1833-1847, 2007.

[3] N. Q. K. Duong, E. Vincent and R. Gribonval, Under-determined reverberant audio source separation using a full-rank spatial covariance model, *IEEE Trans. on ASLP*, vol. 18, no. 7, pp. 1830-1840, Sep. 2010.

[4] H. Sawada, S. Araki, S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. on ASLP*, vol.19, no.3, pp.516-527, Mar. 2011.

[5] P. Bofill and M. Zibulevsky, “Blind separation of more sources than mixtures using sparsity of their short-time-Fourier transform,” in *ICA 2000*, pp.87-92, Jun. 2000.

[6] H. Sawada, S. Araki, S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in Frequency-domain BSS," in *ISCAS 2007*, pp. 3247-3250, May 2007.

[7] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, “First stereo audio source separation evaluation campaign: data, algorithms and results,” in *ICA 2007*, pp. 552-559, Sep. 2007.