

周波数依存到来時間差推定に基づく 劣決定ブラインド音源分離の高速化*

丸山卓郎^{*1,*2} 荒木章子^{*2} 中谷智広^{*2} 宮部滋樹^{*1}
山田武志^{*1} 牧野昭二^{*1} 中村 篤^{*2}

【要旨】 本論文では EM アルゴリズムを用いたスパース性に基づく劣決定ブラインド音源分離 (BSS) の計算を高速化する手法を提案する。Izumi *et al.* は、雑音・残響下でも頑健な劣決定 BSS を提案したが、到来時間差パラメータをイタレーションごとに離散全探索で求める更新則のために計算量に問題があった。そこで、到来時間差パラメータが周波数に依存すると捉えた、時間差パラメータが解析的に更新される計算量の少ない更新則を提案する。また、帯域重み付け平均による帯域非依存到来時間差推定によってパラメータ数を削減し、収束性を向上させる。実験により、提案手法が計算時間を 1/10 程度に削減することを確認した。

キーワード ブラインド音源分離, EM アルゴリズム, スパース性, 周波数依存到来時間差推定, 高速アルゴリズム

Underdetermined blind source separation, EM algorithm, Speech sparseness, Frequency-dependent TDOA estimation, Computationally-efficient algorithm

1. はじめに

ブラインド音源分離 (BSS: Blind Source Separation) 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する手法である。音声信号を対象とした 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おり、研究が活発に行われている [1–6]。BSS では観測信号だけから未知の混合行列と音源成分を推定する必要があるため、通常は音源に関して何等かの仮定を置き、これにより立てられる規準をもとに両未知変数を最適推定する問題として定式化される。例えば、観測信号数が音源数以上の場合には、独立成分分析 (ICA: Independent Component Analysis) が有効な手法として知られている [6]。しかし、観測信号の数が音源数よりも少ない劣決定の条件下では、たとえ混合過程が既知であったとしても音源成分の解は無数に存在し、独立性の規準ではこの中から一意に解が決められない。

劣決定条件での音源分離を実現する枠組として、音源信号のスパース性を利用する手法が一般的である [1]。音声はホルマント構造や調波構造、スピーチブレイクなどの要因により、時間周波数領域では一部の時刻の狭い帯域にエネルギーが集中してスパースに分布する。そのため、複数の音声混在している場合でも、短時間の狭帯域では一つの音源のみが支配的になることが多い [7]。

この音声のスパース性を利用した音源分離の代表的な先行研究として、時間周波数マスキングが挙げられる [1, 2, 5, 8]。これは目的の音源信号成分が支配的である時間周波数成分を、マスキング処理によって抜き出すことで分離する手法である。各音源信号に対するマスクの推定は、主に各時間周波数成分におけるマイクロホン間の到来時間差が信号の到来方向に対応するという性質から、方位のクラスタリング [1] や観測信号の空間的特徴量のクラスタリング [2, 5]、また L1 ノルム最小化 [8] 等により行われる。また、スパース性を仮定せず確率変数の重ね合わせをモデル化した手法も提案されており [9–11]、スパース性を仮定したものよりも高品質な音源分離が得られる。これらの手法は初期値依存性が高いため、初期値生成のための前段処理にスパース性を仮定した手法が有効であることが知られている [12]。しかし、スパース性に基づく音源分離手法の多くでは、残響や背景雑音が多く存在する環境下において各音源の時間差の推定値がばらついてお互いに重なり合うため、特徴空間でのクラスタリングが

* Reduction of computational cost in underdetermined blind source separation based on frequency-dependent time-difference-of-arrival estimation, by Takuro Maruyama, Shoko Araki, Tomohiro Nakatani, Shigeki Miyabe, Takeshi Yamada, Shoji Makino and Atsushi Nakamura.

^{*1} 筑波大学大学院システム情報工学研究科

^{*2} NTT コミュニケーション科学基礎研究所

(問合先: 宮部滋樹 e-mail: miyabe@tara.tsukuba.ac.jp)

(2013年3月26日受付, 2013年9月30日採録決定)

困難になるという問題点があった。

それに対して Izumi *et al.* は, EM アルゴリズムによる混合正規分布モデルの推定問題を適用することにより, 雑音環境下でも頑健に分離できる手法を提案した [13]。しかし, Izumi *et al.* のモデルは音源の時間差を示すパラメータを解析的に求めることができず, 離散全探索によって求めているため, 計算コストがかかってしまうという問題点があった。

本論文では, Izumi *et al.* の手法に対して時間差が周波数に依存するモデルを用いることにより, 時間差を周波数ごとに解析的に求める手法を提案する。これにより時間差を離散全探索で求める必要がなくなり, 計算コストを削減することが可能となる。更に本論文では, 収束性を向上させるための時間差パラメータの帯域重み付けについて議論する。

本論文の構成を以下に示す。第 2 章では, Izumi *et al.* が提案した手法の概要と, その問題点について述べる。第 3 章では, 我々が提案する手法について述べる。第 4 章では, 音源分離の比較実験と考察を行う。第 5 章では, 本論文の結論を述べる。

2. スパース性に基づく雑音環境下の 2ch BSS

本章では, 本論文で用いる Izumi *et al.* の手法 [13] の概要を述べる。この手法は, 幾つかの音源が混合した信号がマイクロホンによって観測されているときに, 観測信号のみから音源信号とその音源方向を最尤推定によって求める。

今, 各音源はスパースであり, 時間周波数領域上で重なり合わない, すなわち, 各時間周波数ビンにおいてアクティブな音源は高々一つであると仮定する。更に, それぞれの音源信号は平面波として到来することを仮定すると, 観測信号は

$$\begin{aligned} \mathbf{x}_{f,t} &= \begin{bmatrix} x_{f,t,L} \\ x_{f,t,R} \end{bmatrix} \\ &\approx \begin{bmatrix} 1 \\ \exp(j2\pi f\delta_k) \end{bmatrix} s_{f,t,k} + \begin{bmatrix} n_{f,t,L} \\ n_{f,t,R} \end{bmatrix} \end{aligned} \quad (1)$$

$$= \mathbf{b}_{f,k} s_{f,t,k} + \mathbf{n}_{f,t} \quad (2)$$

と表される。ここで f は周波数ビンの中心周波数, t はフレームインデックス, L 及び R はマイクロホンアレーの左右それぞれの観測チャネルを表すラベル, k は時間周波数 f, t でアクティブな音源のインデックス, $s_{f,t,k}$ はそのアクティブな音源のスペクトル, $\mathbf{b}_{f,k}$ は音源 k のステアリングベクトル, δ_k は R マイクロホンに対する L マイクロホンの到来時間の遅延の連続時

間による表記である。ここで $\mathbf{n}_{f,t}$ は各音源からの直接音成分以外のすべてを含む雑音項であり, 残響や背景雑音などを含んでいる。ここでは, $\mathbf{n}_{f,t}$ は時間によらず, 平均 0, 共分散行列 $\sigma_f^2 \mathbf{V}_f$ の正規分布に従う定常確率過程としてモデル化する。ただし σ_f^2 は雑音のパワー, $\mathbf{V}_f$ は雑音のマイクロホン間の相関係数からなる行列で, 拡散性雑音を仮定すると以下のような [15]。

$$\mathbf{V}_f = \begin{bmatrix} 1 & \text{sinc}(2\pi f D/c) \\ \text{sinc}(2\pi f D/c) & 1 \end{bmatrix} \quad (3)$$

ここで, c は音速, D はマイクロホン間距離である。パラメータ $s_{f,t,k}$, δ_k , k が与えられた場合に

$$\mathbf{n}_{f,t,k} = \mathbf{x}_{f,t} - s_{f,t,k} \mathbf{b}_{f,k} \quad (4)$$

のように与えられる雑音の推定 $\mathbf{n}_{f,t,k}$ が, 式 (3) の共分散行列 $\sigma_f^2 \mathbf{V}_f$ を持つ零平均多変量正規分布に従うとして, 時間差が δ_k となる方向から k 番目の音声信号 $s_{f,t,k}$ が到来して $\mathbf{x}_{f,t}$ が観測される尤度 $p(\mathbf{x}_{f,t}|k, \theta)$ は,

$$\begin{aligned} p(\mathbf{x}_{f,t}|k, \theta) &= \frac{1}{\pi^2 \sigma_f^4 |\mathbf{V}_f|} \exp \left(-\frac{\mathbf{n}_{f,t,k}^H \mathbf{V}_f^{-1} \mathbf{n}_{f,t,k}}{\sigma_f^2} \right) \end{aligned} \quad (5)$$

のように与えられる。ただし $\{\cdot\}^H$ は複素共役転置を表す。ここでパラメータ θ は, マイクロホン間の到来時間差 δ_k , 雑音のパワースペクトル σ_f^2 及び音源信号 $s_{f,t,k}$ であり, θ は下記の対数尤度

$$L = \sum_{f,t} \log \sum_k p(\mathbf{x}_{f,t}|k, \theta) p(k|\theta) \quad (6)$$

を最大化するパラメータとして最尤推定することができる。ここで, $p(k|\theta)$ は音源の存在確率を表す (ただし $\sum_k p(k|\theta) = 1$)。

この問題には, 時間周波数成分が帰属する音源のインデックス k を未知の隠れ変数とし, パラメータ $\theta = \{\sigma_f^2, \delta_k, s_{f,t,k}\}$ を推定する EM アルゴリズムを適用できる。この問題における EM アルゴリズムの Q 関数は

$$\begin{aligned} Q(\theta|\theta') &= \mathbb{E} [\log p(\mathbf{x}_{f,t}; \theta) | \theta'] \\ &= \sum_{f,t,k} m_{f,t,k} \log p(\mathbf{x}_{f,t}|k, \theta) p(k|\theta) \end{aligned} \quad (7)$$

と書かれる。ただし, $m_{f,t,k}$ は占有確率と呼ばれ, 更新前の推定パラメータ θ' が与えられたもとの $\mathbf{x}_{f,t}$

が k 番目の音源に帰属する事後確率 $p(k|\mathbf{x}_{f,t}, \theta')$ を表す。また, $m_{f,t,k}$ は同時に観測信号から各音源信号へと分離するソフトマスクとして用いることができる。分離信号 $y_{f,t,k}$ は各パラメータの推定値に基づき, 以下の式でソフトマスキングによる期待値として求められる。

$$y_{f,t,k} = m_{f,t,k} \frac{\mathbf{b}_{f,k}^H \mathbf{V}_f^{-1} \mathbf{x}_{f,t}}{\mathbf{b}_{f,k}^H \mathbf{V}_f^{-1} \mathbf{b}_{f,k}} \quad (8)$$

この問題では, E ステップで $m_{f,t,k}$ と Q 関数を更新し, M ステップで残りのパラメータを更新する。E ステップでは占有確率 $m_{f,t,k}$ が以下のように更新される。

$$\begin{aligned} m_{f,t,k} &\leftarrow p(k|\mathbf{x}_{f,t,k}, \theta') \\ &= \frac{p(k|\theta')p(\mathbf{x}_{f,t,k}|k, \theta')}{\sum_k p(k|\theta')p(\mathbf{x}_{f,t,k}|k, \theta')} \end{aligned} \quad (9)$$

M ステップのパラメータ更新のうち, σ_f^2 , $s_{f,t,k}$ は, Q 関数をそれぞれのパラメータで偏微分することにより, 以下の式で解析的に求めることによって更新される。

$$\sigma_f^2 \leftarrow \frac{1}{2T} \sum_{t,f} m_{f,t,k} \mathbf{n}_{f,t,k}^H \mathbf{V}_f^{-1} \mathbf{n}_{f,t,k} \quad (10)$$

$$s_{f,t,k} \leftarrow \frac{\mathbf{b}_{f,k}^H \mathbf{V}_f^{-1} \mathbf{x}_{f,t}}{\mathbf{b}_{f,k}^H \mathbf{V}_f^{-1} \mathbf{b}_{f,k}} \quad (11)$$

ここで T はフレーム数を表す。また, 音源の存在確率 $p(k|\theta)$ は以下の式で更新される。

$$p(k|\theta) \leftarrow \frac{1}{TF} \sum_t \sum_f m_{f,t,k} \quad (12)$$

ここで T と F はそれぞれ時間フレームと周波数ビンの総数である。

以上のパラメータの更新は解析的に行えるが, 時間差 δ_k については, [13] では解析的な更新式が与えられておらず, Q 関数を最大化する時間差 δ_k を全探索することで求めている。

$$\delta_k \leftarrow \arg \max_{\delta_k} Q(\theta|\theta') \quad (13)$$

これは, 例えばマイクロホンから見た音源位置を一定の角度刻みに設定し, それぞれの位置に対する時間差 δ_k について Q 関数を計算する必要があるため, 計算コストが大きいという問題点があった。

3. 提案手法

本章では, 我々が提案する手法について述べる。前章の最後に述べたように, 従来手法には時間差 δ_k が解析的に計算できないという問題点があった。これは, 時間差 δ_k については Q 関数の偏微分が非線形な関数

となってしまう解析的に解くことができないためである。そこで我々は, 時間差 δ_k が周波数に依存するというモデル化を行うことにより, 周波数ごとの時間差 $\delta_{f,k}$ を解析的な更新式により推定する手法を提案する。これにより, 従来手法における計算コストの削減が期待できる。

3.1 時間差パラメータ $\delta_{f,k}$ の推定

尤度関数 (5) の右辺を時間周波数成分ごとに整理する。式 (1) の関係より, 第 k 音源がアクティブであると仮定した場合の観測信号には以下の関係が成り立つ。

$$\begin{bmatrix} n_{f,t,k,L} \\ n_{f,t,k,R} \end{bmatrix} \approx \begin{bmatrix} x_{f,t,L} \\ x_{f,t,R} \end{bmatrix} - \begin{bmatrix} 1 \\ \beta_{f,k} \end{bmatrix} s_{f,t,k} \quad (14)$$

これを用いると, 式 (5) 中の二次形式 $\mathbf{n}_{f,t,k}^H \mathbf{V}_f^{-1} \mathbf{n}_{f,t,k}$ は以下のように変形できる。

$$\begin{aligned} &\mathbf{n}_{f,t,k}^H \mathbf{V}_f^{-1} \mathbf{n}_{f,t,k} \\ &= \begin{bmatrix} n_{f,t,k,L} \\ n_{f,t,k,R} \end{bmatrix}^H \left(\frac{1}{1-\phi_f^2} \begin{bmatrix} 1 & -\phi_f \\ -\phi_f & 1 \end{bmatrix} \right) \begin{bmatrix} n_{f,t,k,L} \\ n_{f,t,k,R} \end{bmatrix} \\ &= \frac{1}{1-\phi_f^2} \left\{ |n_{f,t,k,L}|^2 + |n_{f,t,k,R}|^2 \right. \\ &\quad \left. - 2\phi_f \operatorname{Re} [n_{f,t,k,L}^* n_{f,t,k,R}] \right\} \\ &= \frac{1}{1-\phi_f^2} \left\{ |n_{f,t,k,R} - \phi_f n_{f,t,k,L}|^2 \right. \\ &\quad \left. + (1-\phi_f^2) |n_{f,t,k,L}|^2 \right\} \\ &= \frac{|\xi_{f,t,k} - \beta_{f,k} s_{f,t,k}|^2}{1-\phi_f^2} + |x_{f,t,L} - s_{f,t,k}|^2 \end{aligned} \quad (15)$$

ここで $\operatorname{Re}[\cdot]$ は実部, $\{\cdot\}^*$ は複素共役を表し, 変数 $\phi_f, \beta_{f,k}, \xi_{f,t,k}$ は以下のように定義されるものとする。

$$\phi_f = \operatorname{sinc}(2\pi f D/c) \quad (16)$$

$$\beta_{f,k} = \exp(j2\pi f \delta_{f,k}) \quad (17)$$

$$\xi_{f,t,k} = [x_{f,t,R} - \phi_f(x_{f,t,L} - s_{f,t,k})] \quad (18)$$

更に, 式 (15) 中の $\beta_{f,k}$ を含む項は次のように書き換えることができる。

$$\begin{aligned} &|\xi_{f,t,k} - \beta_{f,k} s_{f,t,k}|^2 \\ &= |\xi_{f,t,k}|^2 + |s_{f,t,k}|^2 \\ &\quad - 2|\xi_{f,t,k}||s_{f,t,k}| \\ &\quad \cdot \cos(\psi_{s_{f,t,k}} - \psi_{\xi_{f,t,k}} - 2\pi f \delta_{f,k}) \end{aligned} \quad (19)$$

ここで, $\psi_{s_{f,t,k}}$ 及び $\psi_{\xi_{f,t,k}}$ は, $s_{f,t,k}$ 及び $\xi_{f,t,k}$ の

位相であり (すなわち $s_{f,t,k} = |s_{f,t,k}| \exp(j\psi_{s_{f,t,k}})$, $\xi_{f,t,k} = |\xi_{f,t,k}| \exp(j\psi_{\xi_{f,t,k}})$ である), 雑音のマイクロホン間相関 $\phi_f = 0$ のとき, $\psi_{s_{f,t,k}} - \psi_{\xi_{f,t,k}}$ は左右チャンネルの観測信号の位相差となる。

以上を用いると, 式 (5) の尤度は

$$p(\mathbf{x}_{f,t} | k, \theta) = \frac{1}{\pi^2 \sigma_f^4 |\mathbf{V}_f|} \exp(-C_{f,t,k}) \cdot \exp\left(2 \frac{|\xi_{f,t,k}| |s_{f,t,k}|}{\sigma_f^2 (1 - \phi_f^2)} \cdot \cos(\psi_{s_{f,t,k}} - \psi_{\xi_{f,t,k}} - 2\pi f \delta_{f,k})\right) \quad (20)$$

となる。ここで

$$C_{f,t,k} = \frac{|\xi_{f,t,k}|^2 + |s_{f,t,k}|^2}{\sigma_f^2 (1 - \phi_f^2)} + \frac{|x_{f,t,L} - s_{f,t,k}|^2}{\sigma_f^2} \quad (21)$$

は $\delta_{f,k}$ によらない定数である。式 (20) の最終項である

$$\exp\left(2 \frac{|\xi_{f,t,k}| |s_{f,t,k}|}{\sigma_f^2 (1 - \phi_f^2)} \cdot \cos(\psi_{s_{f,t,k}} - \psi_{\xi_{f,t,k}} - 2\pi f \delta_{f,k})\right) \quad (22)$$

は von Mises 分布 [16]

$$g(x | \kappa, \mu) = \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(x - \mu)) \quad (23)$$

の形をしている¹。すなわち式 (22) は, 位相差 $\psi_{s_{f,t,k}} - \psi_{\xi_{f,t,k}} \approx \arg(x_{f,t,L}) - \arg(x_{f,t,R})$ が von Mises 分布に従うことを意味している。平均値 μ は周波数に依存するマイクロホン間の位相差に相当し ($\mu = 2\pi f \delta_{f,k}$), 集中母数 κ は以下の式のように SN 比² に相当する。

$$\kappa = 2 \frac{|\xi_{f,t,k}| |s_{f,t,k}|}{\sigma_f^2 (1 - \phi_f^2)} \quad (24)$$

この von Mises 分布を用いた混合モデルに対して EM アルゴリズムを用いることで, 周波数依存時間差 $\delta_{f,k}$ を求めることができる。式 (20) を式 (7) の Q 関数に代入し $\delta_{f,k}$ によらない項を $C'_{f,k}$ とまとめて整理すると,

$$Q(\delta_{f,t} | \theta') = \frac{2}{\sigma_f^2 (1 - \phi_f^2)} \sum_t m_{f,t,k} |\xi_{f,t,k}| |s_{f,t,k}| \cdot \cos(\psi_{s_{f,t,k}} - \psi_{\xi_{f,t,k}} - 2\pi f \delta_{f,k}) + C'_{f,k} \quad (25)$$

¹ここで x の取り得る範囲は $-\pi < x \leq \pi$ である。また, μ は平均値 ($-\pi < \mu \leq \pi$), $\kappa > 0$ は集中母数, $I_0(x)$ は 0 次の第 1 種ベッセル関数である。

² $\phi_f = 0$ のとき $\frac{|\xi_{f,t,k}| |s_{f,t,k}|}{\sigma_f^2 (1 - \phi_f^2)} = \frac{|s_{f,t,k}|^2}{\sigma_f^2}$ となり SN 比に対応する。

となる。

この $Q(\delta_{f,t} | \theta')$ の $\delta(f, t)$ 偏微分を 0 と置くことにより, 周波数ごとの時間差パラメータ $\delta_{f,t}$ の解析的な更新則が得られる。

$$\begin{aligned} \frac{\partial Q(\delta_{f,k} | \theta')}{\partial \delta_{f,k}} &= 0 \\ &\Leftrightarrow \sum_t m_{f,t,k} |\xi_{f,t,k}| |s_{f,t,k}| \cdot \pi f \sin(\psi_{s_{f,t,k}} - \psi_{\xi_{f,t,k}} - 2\pi f \delta_{f,k}) \\ &= \sum_t \pi f m_{f,t,k} |\xi_{f,t,k}| |s_{f,t,k}| \cdot \left\{ \sin(\psi_{s_{f,t,k}} - \psi_{\xi_{f,t,k}}) \cos(2\pi f \delta_{f,k}) \right. \\ &\quad \left. - \cos(\psi_{s_{f,t,k}} - \psi_{\xi_{f,t,k}}) \sin(2\pi f \delta_{f,k}) \right\} \\ &= 0 \end{aligned} \quad (26)$$

これを解くことで

$$\begin{aligned} \frac{\sin(2\pi f \delta_{f,k})}{\cos(2\pi f \delta_{f,k})} &= \tan(2\pi f \delta_{f,k}) \\ &= \frac{\sum_t m_{f,t,k} |\xi_{f,t,k}| |s_{f,t,k}| \sin(\psi_{s_{f,t,k}} - \psi_{\xi_{f,t,k}})}{\sum_t m_{f,t,k} |\xi_{f,t,k}| |s_{f,t,k}| \cos(\psi_{s_{f,t,k}} - \psi_{\xi_{f,t,k}})} \end{aligned} \quad (27)$$

が得られる。従って,

$$\begin{aligned} 2\pi f \delta_{f,k} &= \arctan \frac{\sum_t m_{f,t,k} |\xi_{f,t,k}| |s_{f,t,k}| \sin(\psi_{s_{f,t,k}} - \psi_{\xi_{f,t,k}})}{\sum_t m_{f,t,k} |\xi_{f,t,k}| |s_{f,t,k}| \cos(\psi_{s_{f,t,k}} - \psi_{\xi_{f,t,k}})} \end{aligned} \quad (28)$$

となり, 時間差 $\delta_{f,k}$ を求めることができる。

また, ここで式 (28) の左辺 $2\pi f \delta_{f,k}$ が $-\pi$ から π の値を取り得るのに対し, 右辺の $\arctan$ が $-\pi/2$ から $\pi/2$ の値しか返すことができないため, 式 (28) の解が Q 関数の極大値ではなく, 極小値を与えてしまう可能性がある。よって, Q 関数の 2 階微分 $\frac{\partial^2 Q(\delta_{f,k} | \theta')}{\partial (\delta_{f,k})^2}$ の値を確かめることで解 $\delta_{f,k}$ が Q 関数の極大値を与えるように補正を行う。二階微分 $\frac{\partial^2 Q(\delta_{f,k} | \theta')}{\partial (\delta_{f,k})^2}$ は以下の式ようになる。

$$\begin{aligned} \frac{\partial^2 Q(\delta_{f,k} | \theta)}{\partial (\delta_{f,k})^2} &= \sum_t m_{f,t,k} \frac{|\xi_{f,t,k}| |s_{f,t,k}|}{\sigma_f^2 (1 - \phi_f^2)} (2\pi f)^2 \\ &\quad \cdot \left[-\sin(\psi_{\xi_{f,t,k}} - \psi_{s_{f,t,k}}) \sin(2\pi f \delta_{f,k}) \right] \end{aligned}$$

$$- \cos(\psi_{\xi_{f,t,k}} - \psi_{s_{f,t,k}}) \cos(2\pi f \delta_{f,t,k}) \Big] \quad (29)$$

$\delta_{f,k}$ と $\frac{\partial^2 Q(\delta_{f,k}|\theta)}{\partial(\delta_{f,k})^2}$ の値によって、以下のように場合分けをして補正を行う。

$$2\pi f \delta_{f,k} \leftarrow \begin{cases} 2\pi f \delta_{f,k} + \pi & \delta_{f,k} < 0, \frac{\partial^2 Q(\delta_{f,k}|\theta')}{\partial(\delta_{f,k})^2} \geq 0 \\ 2\pi f \delta_{f,k} - \pi & \delta_{f,k} > 0, \frac{\partial^2 Q(\delta_{f,k}|\theta')}{\partial(\delta_{f,k})^2} \geq 0 \\ 2\pi f \delta_{f,k} & \text{otherwise} \end{cases} \quad (30)$$

以上のようにして、周波数ごとに独立の時間差パラメータ $\delta_{f,k}$ を推定することができる。

ただしこの方法では周波数ごとに時間差パラメータ $\delta_{f,k}$ の推定を行うため、パーミュテーション問題を解く必要がある。今回は、EM アルゴリズムの収束後にそれぞれの周波数で時間差パラメータ $\delta_{f,k}$ を値の小さな順に並べることで、パーミュテーション問題を解決している。

3.2 解析的更新アルゴリズムの計算量削減の理論的見積

ここでは従来手法と提案手法の到来時間差パラメータの更新に要する演算量の比較を、積演算の回数から議論する。従来法の EM アルゴリズムにおける周波数非依存時間差パラメータ δ_k の更新は、以下のように、 Q 関数の δ_k に依存する項の式 (25) と類似の最大化問題に相当する。

$$\begin{aligned} \delta_k &\leftarrow \arg \max_{\delta_k} Q(\theta|\theta') \\ &= \arg \max_{\delta_k} \sum_{t,f} \frac{m_{f,t,k} |\xi_{f,t,k}| |s_{f,t,k}|}{\sigma_f^2 (1 - \phi_f^2)} \\ &\quad \cdot \cos(\psi_{s_{f,t,k}} - \psi_{\xi_{f,t,k}} - 2\pi f \delta_k) \end{aligned} \quad (31)$$

この最大化問題は解析的に解くことができないため、 δ_k の解の候補を N 個に離散化して探索する必要がある。この計算の例としては、まず $m_{f,t,k} |\xi_{f,t,k}| |s_{f,t,k}|$ をすべての時間周波数スロットについて求め、更に N 個の各解候補ごとに $\cos(\varphi_{s_{f,t,k}} - \varphi_{\xi_{f,t,k}} - 2\pi f \delta_k)$ を掛ける演算を行い、最後に $\sigma_f^2 (1 - \phi_f^2)$ の積を計算するとすると、合計 $(4 + N)TF$ 回程度の積演算が必要になる。従来手法の場合は N を小さくすれば演算量を削減することができるが、小さい値に設定すると、特に高域において位相の誤差が大きくなり、音源分離精度の劣化に直結してしまう。次章の実験では $N = 181$ として評価を行っている。

提案手法の場合は、 $m_{f,t,k} |\xi_{f,t,k}| |s_{f,t,k}|$ をすべての

時間周波数スロットについて求め、次に式 (28) において $\sin(\varphi_{s_{f,t,k}} - \varphi_{\xi_{f,t,k}})$ 及び $\cos(\varphi_{s_{f,t,k}} - \varphi_{\xi_{f,t,k}})$ を掛ける演算を行う必要がある。更に式 (29) の 2 階偏導関数を求める必要があるが、この演算では式 (28) の積和に $\sin(2\pi f \delta_{f,k})$ 及び $\sin(2\pi f \delta_{f,k})$ を掛けるだけである。従って、 $6TF$ 回程度の積演算となり、積演算を $N = 2$ の場合の従来法と同程度にまで抑えられることになる。

3.3 周波数平均による周波数非依存時間差推定

ここまでで求めた時間差パラメータの推定法には、各周波数ごとの時間差 $\delta_{f,k}$ を計算しそれをパラメータとして用いるため、パラメータの総数が増えて EM アルゴリズムの解探索が複雑になってしまうという問題点がある。また、時間差 $\delta_{f,k}$ の推定にその周波数ビンの情報しか用いることができないことから、時間差の推定が無残響のものからかけ離れてしまうことが確認されている [14]。EM アルゴリズムは反復によって各パラメータを最尤推定する手法であることから、解析的に求められる帯域依存時間差パラメータ $\delta_{f,k}$ を平均して帯域非依存時間差パラメータ $\bar{\delta}_k$ に置き換えることで、各イタレーションで更新するパラメータがすべての帯域の $\delta_{f,k}$ からすべての周波数で共通の $\bar{\delta}_k$ のみに削減され、解探索の単純化により収束性がより向上すると期待できる。以下では時間差パラメータ推定の収束性向上とそれに伴う計算速度の高速化を目的として、以下の 2 種類の手法の導入を行う [17]。

3.3.1 帯域制限付き平均による周波数非依存時間差パラメータ推定

周波数ごとに独立して推定された時間差 $\delta_{f,k}$ の平均値 $\bar{\delta}_k$ を時間差 $\delta_{f,k}$ に代入し、他のパラメータの計算に用いる。

$$\delta_{f,k} \leftarrow \bar{\delta}_k \quad (32)$$

ここで図-1 は提案手法で推定された時間差 $\delta_{f,k}$ の一例である。横軸が周波数、縦軸が時間差 $\delta_{f,k}$ の値であり、細線が提案手法によって求められた時間差 $\delta_{f,k}$ 、太線が無残響における正解の時間差 δ_k の値を示している。この図より、低域、高域において時間差 $\delta_{f,k}$ の推定精度が無残響の正解値とかけ離れていることが見て取れる。

よって、時間差 $\delta_{f,k}$ の平均値を求める際に、使用する周波数帯域に制限を設ける。すなわち、時間差 $\bar{\delta}_k$ を以下の式のようにする。

$$\bar{\delta}_k = \frac{1}{|f_L < f \leq f_H|} \sum_{f_L < f \leq f_H} \delta_{f,k} \quad (33)$$

ここで f_L, f_H は平均を取る周波数帯域の下限と上限を、

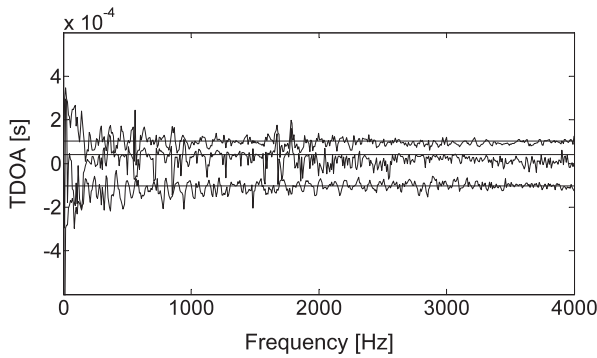


図-1 3音源の場合の $\delta_{f,k}$ の例

$\delta_{f,k}$ はソートしたものを繋げて曲線で表示し、太線は無残響の場合の到来時間差を示す。低域の推定が無残響のものからかけ離れていることが分かる。

$|f_L < f \leq f_H|$ は平均を取る帯域に含まれる周波数ビンの総数を表し、実験では $f_L = 200$ Hz, $f_H = 3,000$ Hz とした。これにより、複数の周波数の情報を用いて時間差パラメータの更新を行うことができる。また、すべての周波数ビンで同じ情報を扱うためパーミュテーション問題を解く必要がなくなる。

3.3.2 周波数パワー平均による帯域非依存時間差パラメータ推定

各時間周波数スロットにおいて、パワーの大きいところは情報の信頼性も高いと考え、各スロットにおいてパワーによる重み w_f を付ける。 w_f は以下の式とする。

$$w_f = \sum_t |x_{f,t}|^2 \quad (34)$$

前項における時間差 $\delta_{f,k}$ の平均値を求める際にこの重み w_f を使用する。すなわち、時間差 $\bar{\delta}_k$ は以下の式のようにする。

$$\bar{\delta}_k = \frac{\sum_{f_L < f \leq f_H} w_f \delta_{f,k}}{\sum_{f_L < f \leq f_H} w_f} \quad (35)$$

3.4 提案手法のまとめ

本章では解析的な更新則を用いて Izumi *et al.* の従来手法の計算量を削減する三つの手法を提案した。次章の比較実験では、第 3.1 節で周波数依存の時間差パラメータ推定を行い、パーミュテーション解決を行った手法を提案手法 A と呼ぶことにする。また、3.3.1 項で述べた周波数の平均化により周波数非依存の時間差パラメータを推定する手法を提案手法 B, 3.3.2 項で述べたパワー重み付けによる手法を提案手法 C と呼ぶこととする。

4. 比較実験

4.1 実験条件

2ch BSS に対し従来手法、提案手法を適用し、シミュレーション実験によりその分離性能、計算時間、到来角度推定誤差を確認した。

三つの音源及び二つのマイクロホンを配置し、実測した部屋のインパルス応答を音源信号に畳み込んで観測信号を作成した。比較対象は Izumi *et al.* の従来手法と提案手法 A-C の 4 手法である。雑音としては、平均 0, 共分散行列 $\sigma_f^2 V_f$ (V_f は式 (3)) のガウスノイズを SN 比約 20 dB 及び 10 dB にて重畳した。部屋の残響時間は 130 ms 及び 300 ms, サンプリング周波数は 8 kHz, 短時間フーリエ変換の窓長及びシフト長はそれぞれ 64 ms, 16 ms とした。音源には信号長 5 秒の英語話者音声を用い、男性 4 発話と女性 4 発話の合計 8 発話の音源から 3 発話を選ぶ全順列 336 通りにて実験を行った。音源方位は残響時間 130 ms の部屋では $30^\circ, 70^\circ, 150^\circ$, 残響時間 300 ms の部屋では $30^\circ, 90^\circ, 150^\circ$ とした。また、EM アルゴリズムの反復は、尤度の増加量がある閾値以下になったら終了とした。

従来手法の時間差パラメータ推定部については、[13] と同様に音源位置を 0° から 180° まで 1° きざみで変化させ、それに対応する $N = 181$ 種類の時間差 δ_k の値について全探索を行った。

評価として、分離性能の客観品質評価、計算時間、到来角度推定誤差を比較した。客観品質評価には SIR (信号対妨害音比, 妨害音には他話者音声と雑音を含む) と SDR (信号対ひずみ比) を用いる [18]。これらは値が大きいほど品質が高いことを表す。計算時間は、計算機による分離の際の収束までの EM アルゴリズムの計算時間で、パラメータが周波数間で共有されない提案手法 A についてはパーミュテーション解決を含めた計算時間とした。その計算時間は、AMD Opteron 4184 2.8 GHz のコア 1 個を占有する MATLAB R2010b 上での計算を評価した。到来角度 (DOA) 推定誤差は、時間差 δ_k より求めた DOA と、正解の DOA との平均二乗誤差 (RMSE) である。また、すべての評価においてエラーバーは 95%信頼区間を表す。

4.2 実験結果

図-2 に EM アルゴリズムの計算時間の実験結果を、図-3 に収束までの反復回数の実験結果を示す。従来手法と A-C すべての提案手法を比較すると、残響時間にかかわらずすべての条件で計算時間が 10 分の 1 程度になっていることが分かる。従って、従来手法の計算のボトルネックであった離散値全探索による更新を

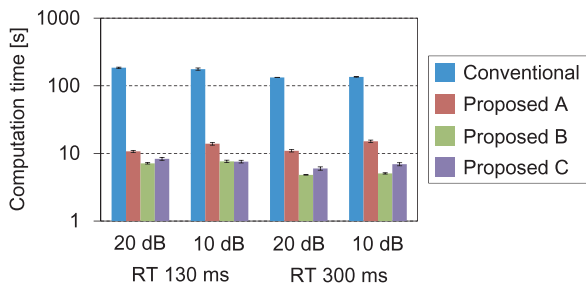


図-2 SNR 20 及び 10 dB, 残響時間 130 ms 及び 300 ms の計算時間の比較

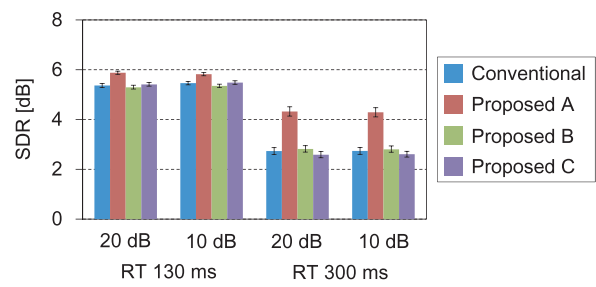


図-5 SNR 20 及び 10 dB, 残響時間 130 ms 及び 300 ms の SDR の比較

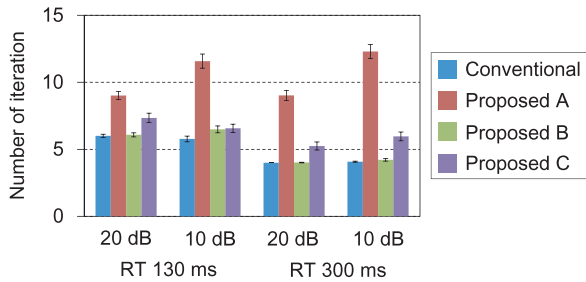


図-3 SNR 20 及び 10 dB, 残響時間 130 ms 及び 300 ms の EM アルゴリズムの反復回数の比較

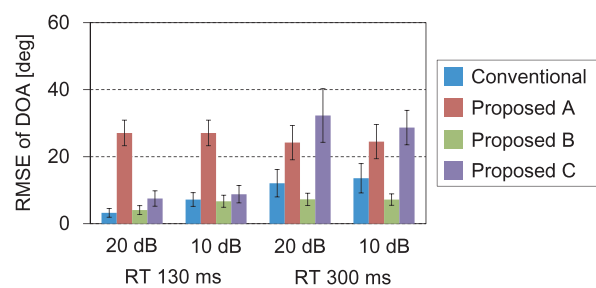


図-6 SNR 20 及び 10 dB, 残響時間 130 ms 及び 300 ms の DOA 推定誤差の比較

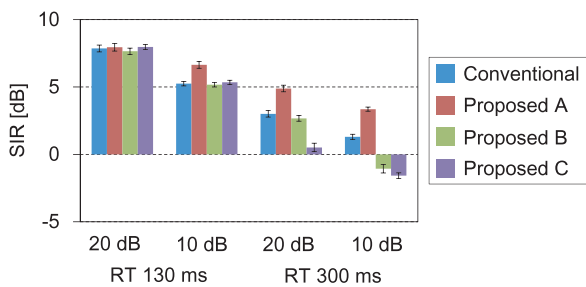


図-4 SNR 20 及び 10 dB, 残響時間 130 ms 及び 300 ms の SIR の比較

解析的な更新に置き換えることにより、計算時間が大幅に短縮できていることが分かる。また、図-3を見ると、提案手法 A は反復回数が最も大きくなっているのに対し、提案手法 B, C は反復回数が半分程度に小さくなっている。その結果、図-2 において提案手法 B, C の計算時間が提案手法 A の半分程度に削減されている。これにより、周波数依存時間差パラメータを周波数平均で置き換える提案手法 B, C は、パラメータを削減して収束性が向上させることが確認された。

図-4 に SIR, 図-5 に SDR, 図-6 に DOA 推定誤差の実験結果を示す。図-4 及び図-5 を見ると、低残響 (130 ms) の条件では、どの手法も同程度の推定精度であることが分かる。従って、低残響下では、どの提案手法も従来手法と同程度の精度の音源分離をはるかに小さい計算量で達成できるということが分かる。また、高残響下の提案手法 A の性能は、従来手法を含めたどの手法よりも分離精度が高いことが見て取れる。これは、周波数ごとに違う時間差を用いた自由度の高いモ

デル化により、図-6 の大きな誤差に現れているように DOA に従う到来時間差を推定するのではなく、残響の影響での周波数ごとに無残響のものからずれてしまう位相差をよく推定できるためであると考えられる。このことは、Izumi *et al.* の手法は観測雑音のモデル化により残響への頑健性を高めているとはいえ、高残響下の音源分離には周波数独立な時間差モデルは不十分で、帯域ごとに別々に無残響モデルからずれてしまう位相差に適応する必要があるということを示唆している。同様のことは提案手法 B と提案手法 C の比較からも言える。低残響の場合には、提案手法 C の DOA 推定誤差が大きいにもかかわらず、分離性能はやや高いことが見て取れる。これは、パワーで重みづけることにより、無残響の時間差推定の正確さを犠牲にしてもパワーの大きい帯域の分離に特化したパラメータ推定となっているためであると考えられる。しかし特に残響時間が長い場合には、伝達関数が時間差のみでは表せなくなり、提案手法 B, C の到来時間差の周波数平均が音源分離性能を損ねてしまうことが分かる。

4.3 考察

以上の考察をまとめる。まず、提案手法 A-C の周波数依存の解析的な到来時間差パラメータ更新により、EM アルゴリズムの計算量を大幅に削減できることが確認された。周波数依存の到来時間差モデルはパラメータ数が大きくなるため、やや反復回数が大きくなるが、それでも計算量削減の効果は大きい。また、提案手法 B 及び C の到来時間差の周波数平均により、パラメー

タ数を削減して収束性を回復できることが確認された。しかし高残響下では、伝達関数が単純な時間差モデルでは表せないために周波数非依存到来時間差モデルが音源分離性能を損ねてしまうため、到来時間差を周波数独立とみなした提案手法 A が最も頑健に音源分離を達成できることが分かった。

5. おわりに

本論文では、EM アルゴリズムを用いたスパース性に基づく雑音・残響環境下の 2ch BSS 問題において、Izumi *et al.* の提案した手法に対し、時間差が周波数に依存するモデルを用い時間差を周波数ごとに解析的に求める手法を提案した。従来手法では離散全探索で求めている時間差を示すパラメータを解析的に求めることで、分離性能はほぼそのままに、計算時間を大幅に短縮できることを確認した。

謝 辞

本研究を進めるにあたり、有益なご意見を頂きました、NTT コミュニケーション科学基礎研究所の吉岡拓也博士、木下慶介博士、筑波大学の葉袋良一助教、Tomas Rutkowski 講師、寺澤洋子助教、松原正樹博士に感謝いたします。

文 献

- [1] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, 52, 1830–1847 (2004).
- [2] S. Araki, H. Sawada, R. Mukai and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *IEEE Trans. Audio Speech Lang. Process.*, 77, 1833–1847 (2007).
- [3] M. Mandel, D. Ellis and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," *Proc. Neural Inf. Process. Syst.*, pp. 1–8 (2007).
- [4] C. Fevotte and S.J. Godsill, "A bayesian approach for blind separation of sparse sources," *IEEE Trans. Speech Audio Process.*, 14, 2174–2188 (2006).
- [5] H. Sawada, S. Araki and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio Speech Lang. Process.*, 19, 516–527 (2011).
- [6] A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis* (John Wiley & Sons, New York, 2001).
- [7] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," *Proc. Int. Conf. Acoust. Speech Signal Process.*, Vol. 1, pp. 529–532 (2002).
- [8] S. Winter, H. Sawada, S. Araki and S. Makino, "Overcomplete BSS for convolutive mixtures based on hierarchical clustering," *Proc. Statistical and Perceptual Audition*, 4 pp. (2004).
- [9] N.Q.K. Duong, E. Vincent and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio Speech Lang. Process.*, 18, 1830–1840 (2010).
- [10] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, 18, 550–563 (2010).
- [11] H. Sawada, H. Kameoka, S. Araki and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. Audio Speech Lang. Process.*, 21, 971–982 (2013).
- [12] K. Iso, S. Araki, S. Makino, T. Nakatani, H. Sawada, T. Yamada and A. Nakamura, "Blind source separation of mixed speech in a high reverberation environment," *Proc. Hands-free Speech Commun.*, pp. 36–39 (2011).
- [13] Y. Izumi, N. Ono and S. Sagayama, "Sparseness-based 2ch BSS using the EM Algorithm in Reverberant Environment," *Proc. Workshop Appl. Signal Process. Audio Acoust.*, pp. 147–150 (2007).
- [14] T. Maruyama, S. Araki, T. Nakatani, S. Miyabe, T. Yamada, S. Makino and A. Nakamura, "New analytical update rule for TDOA inference for underdetermined BSS in noisy environments," *Proc. Int. Conf. Acoust. Speech Signal Process.*, pp. 269–272 (2012).
- [15] R.K. Cook, R.V. Waterhouse, R.D. Berendt, S. Edelman and J.M.C. Thompson, "Measurement of correlation coefficients in reverberant sound fields," *J. Acoust. Soc. Am.*, 27, 1072–1077 (1955).
- [16] C.M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2008).
- [17] T. Maruyama, S. Araki, T. Nakatani, S. Miyabe, T. Yamada, S. Makino and A. Nakamura, "New analytical calculation and estimation of TDOA for underdetermined BSS in noisy environments," *Proc. Asia-Pacific Signal Inf. Process. Assoc. Ann. Summit Conf.*, 6 pp. (2012).
- [18] E. Vincent, H. Sawada, P. Bofill, S. Makino and J. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," *Proc. Int. Conf. Independent Component Analysis Signal Separation*, pp. 552–559 (2007).

丸山 卓郎



平 23 筑波大学情報学群情報科学類卒。
平 25 同大大学院システム情報工学研究科
コンピュータサイエンス専攻博士前期課程
修了。在学中、音声認識、ブライント音源
分離に関する研究に従事。

荒木 章子



平 12 東大大学院・工・計数修士課程了、
同年日本電信電話(株)入社。実環境にお
ける音響信号処理の研究に従事、現在、同
社コミュニケーション科学基礎研究所主任
研究員。博士(情報科学)。日本音響学会、
IEEE、電子情報通信学会会員。



中谷 智広

平 3 京大大学院工学研究科修士課程了。平 13 同大学院博士号取得。平 3 日本電信電話株式会社入社。以来、実環境における頑健な音声認識の実現に向けて音源分離・残響除去の研究に従事。現在、同社コミュニケーション科学基礎研究所主幹研究員。平 20 から名大客員准教授を兼任。博士 (情報学)。平 17 年度本会論文賞受賞。



牧野 昭二

昭 56 東北大学院修士課程了。同年日本電信電話公社入社。以来、NTT 研究所において、電気音響変換器、音響エコーキャンセラ、ブラインド音源分離などの音響信号処理の研究に従事。工博。現在、筑波大学生命領域学際研究センター教授。IEEE Distinguished Lecturer。IEEE Fellow。電子情報通信学会 Fellow。



宮部 滋樹

平 19 奈良先端大博士後期課程了。平 20 米ジョージア工科大客員研究員。平 21 東大特任研究員。平 21 同大助教。平 22 筑波大助教。音響信号処理の研究に従事。博士 (工学)。日本音響学会、IEEE、電子情報通信学会会員。



中村 篤

昭 62 九大大学院・工・情報修士課程了、同年日本電信電話 (株) 入社。実環境自然発話音声認識を中心とするメディア情報処理の研究に従事、現在、同社コミュニケーション科学基礎研究所主幹研究員、信号処理研究グループリリーダ。博士 (工学)。本会評議員。IEEE シニアメンバ、電子情報通信学会会員。



山田 武志

平 11 奈良先端大博士後期課程了。同年、筑波大学講師。現在、同准教授。音声認識、音環境理解、多チャネル信号処理、メディア品質評価、e ラーニングの研究に従事。博士 (工学)。IEEE、電子情報通信学会、情報処理学会、日本音響学会、日本言語テスト学会各会員。