1章 統計的音響信号処理の新展開

吉井和佳[†]，糸山克寿[†]

キーワード：統計的音響信号処理，板倉・斎藤ダイバージェンス，確率モデル，線形予測分析，非負値行列分解，ソース・フィルタ理論

1．まえがき

本稿では，音を聴き分けるという観点から，モノラルの混合音を分離する技術の最新動向を解説する．マルチチャネル信号処理においては，音源数がマイク数以下（優決定）であれば，マイク間の位相差や独立性などに着目することで，高精度な分離が可能である．一方，モノラル信号の分離は数学的に不良設定問題（劣決定）であり，音響信号（音声・音楽・環境音）に内在する「スパース性」や「低ランク性」といった何らかの性質を音の聴き分けの手がかりに用いる必要がある．具体的には，各音源信号のスペクトルは局所的な周波数領域にエネルギーが集中していることや，観測信号のスペクトルは高々有限個の音源スペクトルが重畳して構成されていることなどを音源分離の制約に用いることができる（3.3項）．

本稿では，板倉・斎藤（IS）ダイバージェンス最小化という一貫した立場から，音声信号に対する古典的解析法である線形予測分析[1)]（Linear Predictive Coding: LPC）をはじめ，モノラル音響信号の音源分離において優れた性能を示す非負値行列分解[2)]（Nonnegative MatrixFactorization: NMF）や半正定値テンソル分解[3)4)]（Positive Semidefinite Tensor Factorization: PSDTF）など最新技術を一挙に解説する（2～4節）．各手法におけるISダイバージェンスの最小化は確率モデルの最尤推定に対応しており，LPCとNMFを確率的に統合した複合自己回帰モデル[5)6)]（Composite Autoregressive Model: CAR）が自然に導ける（5節）．同様に，文献7）8）を参考に，LPCとPSDTFを確率的に組み合せることで，従来のモデルをすべて内包する統一的な確率モデルを構成できることを示す（6節）．

本稿で使用する数学記法について以下の通り定める．まず，$\hat{\boldsymbol{x}} \in \mathbb{R}^M$を時間領域でサンプリングされた離散信号，$\tilde{\boldsymbol{x}} \in \mathbb{C}^M$を複素スペクトル，$\boldsymbol{x} \in \mathbb{R}^M$を非負のパワースペクトルとする．ここで，$M$は離散フーリエ変換の窓幅を表す．同一記号を共有する変数は同一実体の異なる表現であるが，$\boldsymbol{x}$から$\hat{\boldsymbol{x}}$や$\tilde{\boldsymbol{x}}$に変換はできないことに注意する．また，$*$は共役，^{T}は転置，^{H}は共役転置を表すものとする．$\odot$はベクトル間の要素同士の積を表す．

2．線形予測分析

本節では，音声信号（単独発話）の音色分析によく利用される線形予測分析[1)]（LPC）について解説する．LPCを用いると，与えられた音声信号の周波数スペクトルの概形（スペクトル包絡）を求めることができる．音素を識別するうえでスペクトル包絡のピーク（フォルマント）の位置や形状は重要な手がかりを与えるため，LPCは歴史的に重要な音響的特徴量抽出法としての役割を果たしてきた．

2.1　ソース・フィルタ理論

音声信号の音響的な性質は，人間の発声機構に基づいて説明できる[9)]．声帯から生成される「音源信号」が，声道の形状に合わせて変化する「フィルタ」を通過することで，多様な音声が生成されると考える．音源信号としては，声帯の振動（周期信号）や雑音などがある．一方，調音フィルタは共振特性のみで記述できる（周波数応答は極しか持たない）と考えるのが一般的である．実際，調音器官を単純な音響管の接続と考えれば，鼻子音を除く音素には反共振は存在しない．

音素を識別するには，音声信号が通過した声道の形状を表す特徴量，すなわち調音フィルタを推定することが重要になる．しかし，音声信号だけから調音フィルタと音源信号を同時に推定する問題は不良設定問題であるため，何らかの制約が必要になる．音声信号のスペクトルにおいては，音源スペクトルは微細構造（パワーの急峻な増減）に，調音フィルタのスペクトルはなめらかな包絡構造に対応していると仮定し，それぞれの成分を分離することがよく行われる．

2.2　確率モデルの定式化

LPCの目的は，離散信号の将来の値をそれまでの標本群の線型和として予測することである．まず，与えられた局所的な音声信号$\hat{\boldsymbol{x}}$（信号全体では$\hat{\boldsymbol{x}}$が周期Mで無限に繰り返すと仮定）がP次の自己回帰過程

$$\hat{x}_m = -\sum_{p=1}^{P} a_p \hat{x}_{m-p} + \hat{s}_m \quad \left(\sum_{p=0}^{P} a_p \hat{x}_{m-p} = \hat{s}_m \right) \tag{1}$$

† 京都大学 大学院情報学研究科
"Recent Progress of Statistical Audio Signal Processing" by Kazuyoshi Yoshii and Katsutoshi Itoyama (Kyoto University, Kyoto)

に従うことを仮定する．ここで，$\boldsymbol{a} = [a_0, \ldots, a_P]^T$ は自己回帰フィルタの係数（a_0=1）であり，$\hat{\boldsymbol{s}} = \{\hat{s}_m\}_{m=1}^M$ は線形予測誤差である．ソース・フィルタ理論では，$\hat{\boldsymbol{x}}$が音声信号，$\hat{\boldsymbol{s}}$が声帯（ソース）から生成される音源信号に対応し，$\boldsymbol{a}$が声道（フィルタ）の特性を決定づける．

式（1）は，$\hat{\boldsymbol{s}}$を入力にとり，$\hat{\boldsymbol{x}}$を出力する線形系とみなすことができ，その振る舞いはパラメータ$\boldsymbol{a}$で決定される．式（1）は$\boldsymbol{a}$と$\hat{\boldsymbol{x}}$との畳み込みであるから

$$A(z)X(z) = S(z) \ \ i.e., \ X(z) = S(z)F(z) \tag{2}$$

が成立する．ここで，$X(z) \stackrel{\text{def}}{=} \sum_{m=-\infty}^{\infty} \hat{x}_m z^{-m}$ および $S(z) \stackrel{\text{def}}{=} \sum_{m=-\infty}^{\infty} \hat{s}_m z^{-m}$ は，それぞれ$\hat{\boldsymbol{x}}$および$\hat{\boldsymbol{s}}$のz変換である．$F(z) \stackrel{\text{def}}{=} \frac{1}{A(z)}$は全極型伝達関数であり，

$$F(z) = \frac{1}{A(z)} = \frac{1}{\sum_{p=0}^{P} a_p z^{-p}} \tag{3}$$

で定まる．これは，フィルタが共振特性のみで説明できることを意味し，ソース・フィルタ理論と相性がよい．いま，式（2）に$z = e^{i\omega_m}$（ただし$\omega_m = 2\pi \frac{m}{M}$）を代入することで，この線形系の伝達特性のフーリエ領域表現

$$\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{s}} \odot \tilde{\boldsymbol{f}} \tag{4}$$

を得る．ここで，観測信号$\hat{\boldsymbol{x}}$，音源信号$\hat{\boldsymbol{s}}$，フィルタの複素スペクトルをそれぞれ$\tilde{\boldsymbol{x}} = \{X(e^{i\omega_m})\}_{m=1}^M$，$\tilde{\boldsymbol{s}} = \{S(e^{i\omega_m})\}_{m=1}^M$，$\tilde{\boldsymbol{f}} = \{F(e^{i\omega_m})\}_{m=1}^M$とした．また，対応するパワースペクトルをそれぞれ$\boldsymbol{x} = \tilde{\boldsymbol{x}} \odot \tilde{\boldsymbol{x}}^*$，$\boldsymbol{s} = \tilde{\boldsymbol{s}} \odot \tilde{\boldsymbol{s}}^*$，$\boldsymbol{f} = \tilde{\boldsymbol{f}} \odot \tilde{\boldsymbol{f}}^*$と定義しておく．

LPCでは，音源信号$\hat{\boldsymbol{s}}$がガウス性白色雑音である，すなわち，複素スペクトル$\tilde{\boldsymbol{s}}$がすべての周波数mで独立同分布な複素ガウス分布に従うことを仮定する．

$$\tilde{\boldsymbol{s}} \sim \mathcal{N}_c(0, \sigma^2 \boldsymbol{I}) \tag{5}$$

ここで，σ^2は各周波数ビンにおける平均的なパワーを表す．式（4）および式（5）を用いると，

$$\tilde{\boldsymbol{x}} \sim \mathcal{N}_c(0, \mathrm{diag}\,(\sigma^2 \boldsymbol{f})) \tag{6}$$

を得る．すなわち，各要素のパワーx_mは指数分布

$$x_m \sim \mathrm{Exponential}\,(\sigma^2 f_m) \tag{7}$$

に従う．図1に，観測信号$\hat{\boldsymbol{x}}$のパワースペクトル$\boldsymbol{x}$から推定されたスペクトル包絡$\boldsymbol{f}$を示す．

2.3　乗法更新アルゴリズムに基づく最適化

観測スペクトル$\boldsymbol{x}$が与えられたとき，式（6）で与えられる尤度を最大化するスペクトル包絡$\boldsymbol{f}$（すなわち$\boldsymbol{a}$）およびパワーσ^2を求めたい．式（6）の対数をとって符号反転させると，ISダイバージェンス

$$\mathcal{D}_{\mathrm{IS}}(\boldsymbol{x} \,|\, \sigma^2 \boldsymbol{f}) = \sum_{m=1}^{M} \left(\frac{x_m}{\sigma^2 f_m} - \log \frac{x_m}{\sigma^2 f_m} - 1 \right) \tag{8}$$

と定数を除いて等しくなることから，式（6）の最大化は式

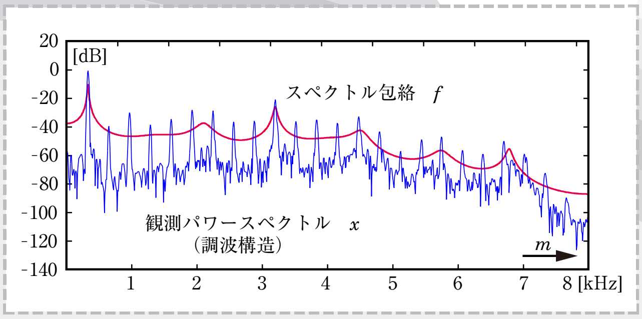

図1　調波構造をもつ観測スペクトルに対する線形予測分析

（8）の最小化と等価である．ここで，f_mは

$$f_m = \frac{1}{\left| \sum_{p=0}^{P} a_p e^{-i\omega_m p} \right|^2} = \frac{1}{\boldsymbol{a}^T \boldsymbol{U}_m \boldsymbol{a}} \tag{9}$$

であり，$\boldsymbol{U}_m$は，各要素が$[\boldsymbol{U}_m]_{pq} = \cos(\omega_m(p - q))$となる$(P+1) \times (P+1)$のテプリッツ行列である．

まず，式（8）をσ^2に関して偏微分してゼロとおくと，

$$\sigma^2 = \frac{1}{M} \sum_{m=1}^{M} \frac{x_m}{f_m} \tag{10}$$

を得る．一方，$\boldsymbol{a}$を求めるには，式（8）を各a_pに関して偏微分してゼロとおいたものを連立して得られるYule-Walker方程式を解けばよいが[1]，乗法更新アルゴリズムと呼ばる効率的な反復解法も提案されている[10]．結果のみ記すと，ベクトル$\boldsymbol{a}$に関する乗法更新則は，

$$a \leftarrow \left(\frac{1}{\sigma^2} \sum_{m=1}^{M} x_m U_m \right)^{-1} \left(\sum_{m=1}^{M} f_m U_m \right) a \tag{11}$$

となる．式（8）が収束するまで式（10）および式（11）を反復する．ただし，反復ごとにσ^2を調節して，$a_0 = 1$を満たすようスペクトル包絡$\boldsymbol{f}$を正規化しておく．

3．非負値行列分解

本節では，モノラル音響信号の音源分離によく利用される非負値行列分解（NMF）について解説する．最小化すべきコスト関数の違いによりさまざまな変種が存在するが，音源分離にはKullback-Leibler（KL）ダイバージェンスに基づくKL-NMF[11]やISダイバージェンスに基づくIS-NMF[2]がよく利用される．本稿では，最適化が難しいが，理論的には音源分離により適しているIS-NMFに着目する．

3.1　コスト関数最小化としての定式化

NMFでは，非負値行列$X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{M \times N}$に対し，$X \approx WH \stackrel{\text{def}}{=} Y$となる二つの非負値行列$W = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K] \in \mathbb{R}^{M \times K}$，$H = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K]^T \in \mathbb{R}^{K \times N}$への低ランク分解を行う．ただし，$\boldsymbol{w}_k$および$\boldsymbol{h}_k$はそれぞれ基底ベクトルおよび対応するアクティベーションベクトルであり，$K \ll \min(M, N)$とする．ここで，再構成行列を$Y = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N] \in \mathbb{R}^{M \times N}$とすると，

$$\boldsymbol{x}_n \approx \sum_{k=1}^{K} h_{kn}\boldsymbol{w}_k \overset{\text{def}}{=} \boldsymbol{y}_n \tag{12}$$

と書ける．観測ベクトル$\boldsymbol{x}_n$と再構成ベクトル$\boldsymbol{y}_n$との間の誤差$\mathcal{D}(\boldsymbol{x}_n \mid \boldsymbol{y}_n)$を評価する尺度として，本稿では以下で定義されるISダイバージェンスに着目する．

$$\mathcal{D}_{\mathrm{IS}}(\boldsymbol{x}_n \mid \boldsymbol{y}_n) = \sum_{m=1}^{M}\left(\frac{x_{nm}}{y_{nm}} - \log\frac{x_{nm}}{y_{nm}} - 1\right) \tag{13}$$

全体のコスト関数$\mathcal{D}_{\mathrm{IS}}(\boldsymbol{X} \mid \boldsymbol{Y}) = \sum_n \mathcal{D}_{\mathrm{IS}}(\boldsymbol{x}_n \mid \boldsymbol{y}_n)$を最小化する$\boldsymbol{W}$および$\boldsymbol{H}$を求めるため，乗法更新アルゴリズム[12]が提案されている．

3.2 乗法更新アルゴリズムに基づく最適化

本項では，補助関数法に基づく収束性が保証された乗法更新アルゴリズムを紹介する．導出は文献12)に詳しいので，本稿では結果のみを記すと，乗法更新則は

$$w_{km} \leftarrow w_{km}\left(\frac{\sum_n x_{mn}h_{kn} / y_{mn}^2}{\sum_n h_{kn} / y_{mn}}\right)^{\frac{1}{2}} \tag{14}$$

$$h_{kn} \leftarrow h_{kn}\left(\frac{\sum_m x_{mn}w_{km} / y_{mn}^2}{\sum_m w_{km} / y_{mn}}\right)^{\frac{1}{2}} \tag{15}$$

となる．ただし，$\sum_m w_{km} = 1$を満たすよう，反復ごとに$\boldsymbol{w}_k$および$\boldsymbol{h}_k$をスケーリングしておく．この更新則では，w_{km}およびh_{kn}の非負性は自然に保たれる．

3.3 音源分離への応用

観測されるモノラル音響信号（混合音）の複素スペクトログラムを$\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_N] \in \mathbb{C}^{M\times N}$，$k$番目の音源信号の複素スペクトログラムを$\tilde{\boldsymbol{X}}_k = [\tilde{\boldsymbol{x}}_{k1}, \ldots, \tilde{\boldsymbol{x}}_{kN}] \in \mathbb{C}^{M\times N}$とする．ここで，$M$は周波数ビン数，$N$はフレーム数である．観測した混合音が$K$個の音源信号の瞬時混合であると仮定すると，以下が成り立つ．

$$\tilde{\boldsymbol{X}} = \sum_{k=1}^{K}\tilde{\boldsymbol{X}}_k \quad \left(\tilde{\boldsymbol{x}}_n = \sum_{k=1}^{k}\tilde{\boldsymbol{x}}_{kn}\right) \tag{16}$$

観測変数$\tilde{\boldsymbol{X}}$を潜在変数$\tilde{\boldsymbol{X}}_k$に分解する不良設定問題を解くには，$\tilde{\boldsymbol{X}}_k$に関して何らかの制約が必要になる．そこで，複素スペクトログラム$\tilde{\boldsymbol{X}}_k$に対応するパワースペクトログラム$\boldsymbol{X}_k$ $[\boldsymbol{x}_{k1}, \ldots, \boldsymbol{x}_{kN}] \in \mathbb{R}^{M\times N}$は，ランク1の行列$\boldsymbol{Y}_k$で近似できると仮定する（図2）．

$$\boldsymbol{X}_k \approx \boldsymbol{w}_k\boldsymbol{h}_k^T \overset{\text{def}}{=} \boldsymbol{Y}_k \tag{17}$$

すなわち，$\boldsymbol{Y}_k = [\boldsymbol{y}_{k1}, \ldots, \boldsymbol{y}_{kN}] \in \mathbb{R}^{M\times N}$をどのフレーム$n$でスライスしても，パワースペクトル$\boldsymbol{y}_{kn}$は基底スペクトル$\boldsymbol{w}_k \in \mathbb{R}^M$を重み$h_{kn}$でスケーリングするだけで得られるものとする（$\boldsymbol{y}_{kn} = h_{kn}\boldsymbol{w}_k$）．この仮定は，同じ形状のパワースペクトルが音量を変えながら繰り返し現れるような打楽器音に対しては特に相性がよい．一方，調波構造をもつ楽器音のスペクトルは，倍音の相対強度は時間変化する．ただし，実際のスペクトル$\boldsymbol{X}_k$と簡単化したスペクトル$\boldsymbol{Y}_k$は厳密に一致している必要はないため，上記仮定は有効に働くと考えられる．

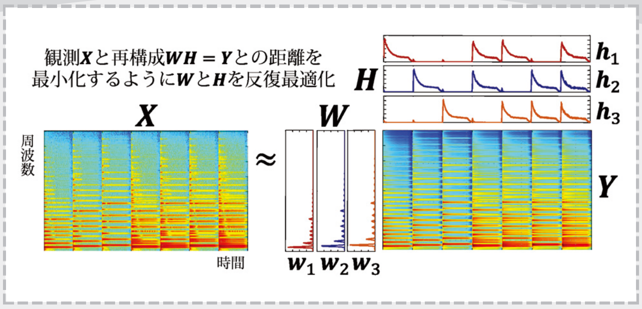

図2 パワースペクトログラムに対する非負値行列分解（NMF）

まず，潜在変数$\tilde{\boldsymbol{x}}_{kn}$が$\boldsymbol{y}_{kn}$で定まる対角共分散行列をもつ複素ガウス分布に従うことを仮定する

$$\tilde{\boldsymbol{x}}_{kn} \sim \mathcal{N}_c(0, \mathrm{diag}(\boldsymbol{y}_{kn})) \tag{18}$$

式（16）に着目すると，複素ガウス分布の再生性から

$$\tilde{\boldsymbol{x}}_n \sim \mathcal{N}_c(0, \mathrm{diag}(\boldsymbol{y}_n)) \tag{19}$$

を得る．ただし，$\boldsymbol{y}_n = \sum_k \boldsymbol{y}_{kn}$である．したがって，$x_{nm} = |\tilde{x}_{nm}|^2$は指数分布に従うことがわかる．

$$x_{nm} \sim \mathrm{Exponential}(y_{nm}) \tag{20}$$

ここで，式（19）の対数をとって符号反転させると，式（13）と定数項を除いて等しい．したがって，式（19）の最大化は式（13）の最小化と等価であり，IS-NMFを用いて$\boldsymbol{y}_n$や$\boldsymbol{y}_{kn} = h_{kn}\boldsymbol{w}_k$を求めることができる．

最終的に，式（18）および式（19）に着目すると，$\tilde{\boldsymbol{x}}_n$が与えられたときの$\tilde{\boldsymbol{x}}_{kn}$の事後分布は複素ガウス分布になることがわかり，その平均と分散は

$$\mathbb{E}[\tilde{\boldsymbol{x}}_{kn} \mid \tilde{\boldsymbol{x}}_n] = \mathrm{diag}(\boldsymbol{y}_{kn})\mathrm{diag}(\boldsymbol{y}_n)^{-1}\tilde{\boldsymbol{x}}_n \tag{21}$$

$$\mathbb{V}[\tilde{\boldsymbol{x}}_{kn} \mid \tilde{\boldsymbol{x}}_n] = \mathrm{diag}(\boldsymbol{y}_{kn}) - \mathrm{diag}(\boldsymbol{y}_{kn})\mathrm{diag}(\boldsymbol{y}_n)^{-1}\mathrm{diag}(\boldsymbol{y}_{kn}) \tag{22}$$

で与えられる．この処理はウィナーフィルタリングと呼ばれ，$\tilde{\boldsymbol{X}}_k$の位相は$\tilde{\boldsymbol{X}}$の位相と同一であると仮定されている．最後に，逆フーリエ変換を用いて，$\mathbb{E}[\tilde{\boldsymbol{X}}_k \mid \tilde{\boldsymbol{X}}]$から$k$番目の音源信号を復元することができる．

4．複合自己回帰モデル

本節では，LPCとNMFとを確率的に統合した複合自己回帰モデル[5][6]（CAR）について解説する．LPCには，音高をもつ音響信号を解析すると，観測スペクトル$\boldsymbol{x}$中の調波構造に影響を受け，推定されるスペクトル包絡$\boldsymbol{f}$は倍音周波数において不要に急峻なピークをもつ欠点があった．この理由は，音源スペクトルはすべての周波数帯域で平均的

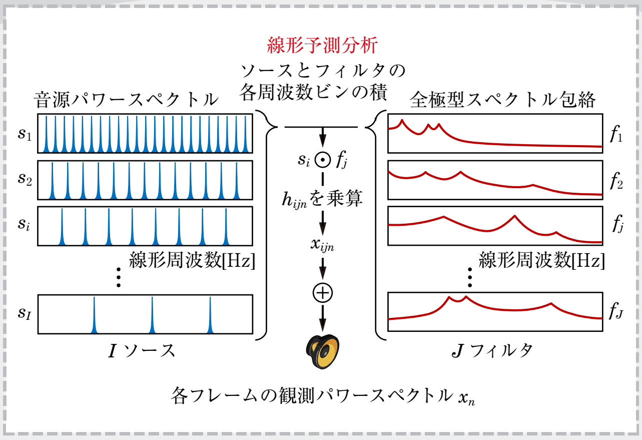

図３　混合音に対する複合自己回帰モデル（CAR）

には等しいパワーを持つという式 (5) の仮定との乖離が大きくなるためである．一方，NMFを音声・音楽信号の音源分離に適用すると，異なる音高ごとに基底スペクトル $\boldsymbol{w}_k$ が割当てられるため（参考：図2における基底スペクトル），式 (21) を用いると混合音が音高ごとに分離されるだけで，楽器パート（音色）ごとに分離することはできなかった．

上記問題を解決するため，CARではスペクトル包絡（音色を表現）と音源スペクトル（音高を表現）をNMFと同様の枠組みで同時推定する．音源スペクトルが調波構造をもつよう制約を加えて，音楽音響信号の音高推定と楽器パート分離を同時に行う拡張も可能である[6]．

4.1　コスト関数最小化として定式化

最初に，3.1項のNMFの枠組みと照らして，CARの定式化について示しておく．図3に示す通り，CARは混合音のパワースペクトログラムをI個の微細構造（ソース）とJ個の全極型スペクトル包絡（フィルタ）とに分解することができるソース・フィルタNMFである[5]．いま，観測パワースペクトログラム$\boldsymbol{X} \in \mathbb{R}^{M \times N}$中の各非負ベクトル$\boldsymbol{x}_n$の三因子への分解を考える．

$$\boldsymbol{x}_n \approx \sum_{i=1}^{I}\sum_{j=1}^{J} h_{ijn}(\boldsymbol{s}_i \odot \boldsymbol{f}_j) \stackrel{\text{def}}{=} \boldsymbol{y}_n \tag{23}$$

ここで，$\boldsymbol{s}_i \in \mathbb{R}^M$はソース$i$のパワースペクトル，$\boldsymbol{f}_j \in \mathbb{R}^M$はフィルタ$j$のパワースペクトル，$h_{ijn}$はフレーム$n$におけるソース$i$・フィルタ$j$の組合せのパワーを表す．式 (23) はNMFと同様に，$\boldsymbol{s}_i$および$\boldsymbol{f}_j$は定常（時不変）であり，その重みh_{ijn}のみが時間変化すると仮定している．観測ベクトル$\boldsymbol{x}_n$と再構成ベクトル$\boldsymbol{y}_n$との間の誤差$\mathcal{D}(\boldsymbol{x}_n|\boldsymbol{y}_n)$を評価する尺度として，式 (13) で定義されるISダイバージェンスを用いるのが適切であることを次項で示す．

4.2　確率モデルの定式化

CARでは，音源信号のガウス性は仮定するが，白色性は仮定しない．式 (5) とは異なり，音源信号（ソース）iの複素スペクトル$\tilde{\boldsymbol{s}}_i=\{S_i(e^{i\omega_m})\}_{m=1}^{M}$は，周波数ビン$m$ごとに異なる分散パラメータ$\boldsymbol{s}_i=\{s_{im}\}_{m=1}^{M}$を持つ独立な複素ガウス分布に従うことを仮定する．

$$\tilde{\boldsymbol{s}}_i \sim \mathcal{N}_c(0, \mathrm{diag}(\boldsymbol{s}_i)) \tag{24}$$

一方，調音フィルタに関してはLPCと同様に全極型を仮定する．すなわち，フィルタjの複素スペクトル$\tilde{\boldsymbol{f}}_j=\{F_j(e^{i\omega_m})\}_{m=1}^{M}$は次式で与えられる．

$$\tilde{f}_{jm} = \frac{1}{\sum_{p=0}^{P} a_{jp} e^{-i\omega_m p}} \tag{25}$$

ここで，$\boldsymbol{a}_j=\{a_{jp}\}_{p=0}^{P}$はソース$j$の線形予測係数である．式 (9) と同様に，フィルタjの非負のパワースペクトルを$\boldsymbol{f}_j=\tilde{\boldsymbol{f}}_j \odot \tilde{\boldsymbol{f}}_j^*$としておく．

いま，あるフレームnにおけるソースiとフィルタjの組合せに起因する複素スペクトル$\tilde{\boldsymbol{x}}_{ijn}$は，式 (4) 同様

$$\tilde{\boldsymbol{x}}_{ijn} = a_{ijn}(\tilde{\boldsymbol{s}}_i \odot \tilde{\boldsymbol{f}}_j) \tag{26}$$

で与えられる．ここで，a_{ijn}はスケーリング係数（直感的には音量に対応）である．式 (24) を用いると，

$$\tilde{\boldsymbol{x}}_{ijn} \sim \mathcal{N}_c(0, h_{ijn}\mathrm{diag}(\boldsymbol{s}_i \odot \boldsymbol{f}_j)) \tag{27}$$

を得る．ここで，$h_{ijn}=a_{ijn}^2 \geq 0$とした．観測される混合音の複素スペクトル$\tilde{\boldsymbol{x}}_n=\{X_n(e^{i\omega_m})\}_{m=1}^{M}$は，あらゆる$i$と$j$の組合せの重畳$\tilde{\boldsymbol{x}}_n=\sum_{ij}\tilde{\boldsymbol{x}}_{ijn}$であると考え，複素ガウス分布の再生性に着目すると，

$$\tilde{\boldsymbol{x}}_n \sim \mathcal{N}_c(0, \boldsymbol{y}_n) \tag{28}$$

を得る．式 (28) は，$|\tilde{x}_{nm}|^2=x_{nm} \geq 0$とすると，

$$x_{nm} \sim \mathrm{Exponential}(y_{nm}) \tag{29}$$

と等価であり，式 (23) におけるコスト関数としてISダイバージェンスが適切であることを示している．

CARはLPCやIS-NMFをその特殊な場合として含む．音源スペクトルがガウス性白色雑音であり（$\boldsymbol{s}_i=\sigma^2\boldsymbol{I}$），ソース・フィルタの個数が$S=J=1$の場合，CARはLPCに帰着する．一方，全周波数帯域でフラットなフィルタが一つだけ存在する場合（$J=1$かつ$\{a_{jp}=0\}_{p=0}^{P}$），CARはIS-NMFに帰着する．

4.3　乗法更新アルゴリズムに基づく最適化

観測パワースペクトログラム$\boldsymbol{X}=\{\boldsymbol{x}_n\}_{n=1}^{N}$が与えられたときに，式 (28) の尤度を最大化する音源スペクトル$\{\boldsymbol{s}_i\}_{i=1}^{I}$，スペクトル包絡$\{\boldsymbol{f}_j\}_{j=1}^{J}$（すなわち$\{\boldsymbol{a}_j\}_{j=1}^{J}$），それらの組合せの時間方向のパワー変化$\{\boldsymbol{h}_{ij}\}_{i=1,j=1}^{I,J}$を求めたい．まず，式 (28) の対数を取ると，式 (13) で示されるISダイバージェンスと定数を除いて等しくなる．これを最小化するには，EMアルゴリズム[5]や2.3項および3.2項で紹介した乗法更新アルゴリズムを組合せて用いればよい[6]．

5．半正定値テンソル分解

本節では，NMFの正統的な拡張である半正定値テンソ

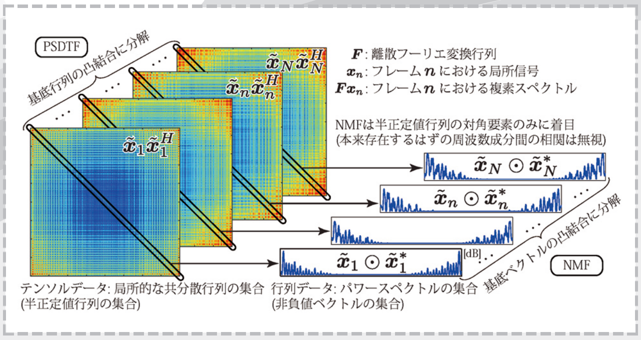

図４　音源分離のための半正定値テンソル分解（PSDTF）

ル分解[3)4)]（PSDTF）について解説する．図４の通り，PSDTFでは，各時刻における複素スペクトル$\tilde{\boldsymbol{x}}_n$の直積$\boldsymbol{X}_n=\tilde{\boldsymbol{x}}_n\tilde{\boldsymbol{x}}_n^H$，すなわち半正定値行列を少数の半正定値基底行列の和に分解する．一方，NMFでは，上記行列の対角成分（パワースペクトル）$\boldsymbol{x}_n=\tilde{\boldsymbol{x}}_n\odot\tilde{\boldsymbol{x}}_n^*$，すなわち非負値ベクトルを少数の非負値基底ベクトルの和に分解する．行列の半正定値性は，ベクトルの非負性の自然な拡張概念である．従来の非負値テンソル分解（NTF）は，非負値データのみを取り扱う点でNMFの単純な拡張であり，PSDTFとは本質的に異なっている．

PSDTFでは，式（16）を保持しながら，観測スペクトログラム$\tilde{\boldsymbol{X}}$から音源スペクトル$\tilde{\boldsymbol{X}}_k$の位相を適切に復元することで，高品質な音源分離を実現する．音源信号の周期と短時間フーリエ変換の窓長Mが異なる場合には，音源信号の巡回定常性の仮定が成り立たなくなるため，周波数ビン間に相関が生じる問題を取り扱える利点は大きい．一方，NMFでは，$\tilde{\boldsymbol{X}}_k$の位相は$\boldsymbol{X}$と同じものをそのまま再利用していたため，分離品質に限界があった．

5.1　コスト関数最小化としての定式化

PSDTFでは，観測データとして3階のテンソル$\boldsymbol{X}=[\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]\in\mathbb{C}^{M\times M\times N}$に対する分解を行う．各要素$\boldsymbol{X}_n\in\mathbb{C}^{M\times M}$は半正定値行列とする．いま，各$\boldsymbol{X}_n$を$K$個の半正定値行列$\{\boldsymbol{V}_k\}_{k=1}^K$（基底行列）の錐（すい）結合

$$\boldsymbol{X}_n \approx \sum_{k=1}^{K} h_{kn}\boldsymbol{V}_k \overset{\text{def}}{=} \boldsymbol{Y}_n \tag{30}$$

で近似したい．ここで，$h_{kn}\geq 0$は$\boldsymbol{X}_n$におけるk番目の基底行列$\boldsymbol{V}_k$の重みである．観測行列$\boldsymbol{X}_n$と再構成行列$\boldsymbol{Y}_n$との間の誤差$\mathcal{D}(\boldsymbol{X}_n|\boldsymbol{Y}_n)$を評価する尺度として，本稿では以下で定義されるLog-Determinant（LD）ダイバージェンス[13)]に着目する．

$$\mathcal{D}_{\mathrm{LD}}(\boldsymbol{X}_n|\boldsymbol{Y}_n)=\mathrm{tr}(\boldsymbol{X}_n\boldsymbol{Y}_n^{-1})-\log|\boldsymbol{X}_n\boldsymbol{Y}_n^{-1}|-M \tag{31}$$

これは，式（13）のISダイバージェンスの自然な拡張である．全体のコスト関数$\mathcal{D}_{\mathrm{LD}}(\boldsymbol{X}|\boldsymbol{Y})=\sum_n\mathcal{D}_{\mathrm{LD}}(\boldsymbol{X}_n|\boldsymbol{Y}_n)$を最小化する$\boldsymbol{H}=[\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K]\in\mathbb{R}^{N\times K}$および$\boldsymbol{V}=[\boldsymbol{V}_1, \ldots, \boldsymbol{V}_K]\in\mathbb{C}^{M\times M\times K}$を求めるLD-PSDTFに対しては，乗法更新アルゴリズム[3)]が提案されている．

5.2　乗法更新アルゴリズムに基づく最適化

本項では，3.2項と同様に，補助関数法に基づく乗法更新アルゴリズムを紹介する．導出は文献3）4）に詳しいので，本稿では結果のみ記すと，乗法更新則は

$$\boldsymbol{V}_k \leftarrow \boldsymbol{V}_k\boldsymbol{L}_k(\boldsymbol{L}_k^T\boldsymbol{V}_k\boldsymbol{P}_k\boldsymbol{V}_k\boldsymbol{L}_k)^{-\frac{1}{2}}\boldsymbol{L}_k^T\boldsymbol{V}_k \tag{32}$$

$$h_{kn} \leftarrow h_{kn}\left(\frac{\mathrm{tr}(\boldsymbol{Y}_n^{-1}\boldsymbol{V}_k\boldsymbol{Y}_n^{-1}\boldsymbol{X}_n)}{\mathrm{tr}(\boldsymbol{Y}_n^{-1}\boldsymbol{V}_k)}\right)^{\frac{1}{2}} \tag{33}$$

となる．ここで，$\boldsymbol{L}_k$はコレスキー分解$\boldsymbol{Q}_k=\boldsymbol{L}_k\boldsymbol{L}_k^T$で求まる下三角行列であり，$\boldsymbol{P}_k$および$\boldsymbol{Q}_k$は次式で求まる．

$$\boldsymbol{P}_k=\sum_{n=1}^{N}h_{kn}\boldsymbol{Y}_n^{-1}\quad \boldsymbol{Q}_k=\sum_{n=1}^{N}h_{kn}\boldsymbol{Y}_n^{-1}\boldsymbol{X}_n\boldsymbol{Y}_n^{-1} \tag{34}$$

したがって，h_{kn}の非負性と$\boldsymbol{V}_k$の半正定値性は自然に保たれているが，$\mathrm{tr}(\boldsymbol{V}_k)=1$を満たすよう，反復ごとに$\boldsymbol{V}_k$および$\boldsymbol{h}_k$をスケーリングしておく．式（32）および式（33）は，式（14）および式（15）の自然な拡張である．

5.3　音源分離への応用

式（16）を満たすように，$\tilde{\boldsymbol{x}}_n$を音源スペクトル$\{\tilde{\boldsymbol{x}}_{kn}\}_{k=1}^K$の和に分解したい．まず，潜在変数$\tilde{\boldsymbol{x}}_{kn}$が共分散行列$\boldsymbol{Y}_{kn}$をもつ複素ガウス分布に従うことを仮定する．

$$\tilde{\boldsymbol{x}}_{kn}\sim\mathcal{N}_c(\boldsymbol{0},\boldsymbol{Y}_{kn}) \tag{35}$$

ここで，式（18）のように共分散行列を対角行列に限定しないことで，周波数ビン間の相関を考慮している．式（16）に着目すると，複素ガウス分布の再生性から

$$\tilde{\boldsymbol{x}}_n\sim\mathcal{N}_c(\boldsymbol{0},\boldsymbol{Y}_n) \tag{36}$$

を得る．ただし，$\boldsymbol{Y}_n=\sum_k\boldsymbol{Y}_{kn}$である．ここで，式（36）の対数をとって符号反転させると，式（31）と定数項を除いて等しい．したがって，式（36）の最大化は式（31）の最小化と等価であり，LD-PSDTFを用いて$\boldsymbol{Y}_n$や$\boldsymbol{Y}_{kn}$を求めることができる．

最終的に，式（35）および式（36）から，$\tilde{\boldsymbol{x}}_n$が与えられたときの$\tilde{\boldsymbol{x}}_{kn}$の事後分布は複素ガウス分布になることがわかり，その平均と分散は次式で求めることができる．

$$\mathbb{E}[\tilde{\boldsymbol{x}}_{kn}|\tilde{\boldsymbol{x}}_n]=\boldsymbol{Y}_{kn}\boldsymbol{Y}_n^{-1}\tilde{\boldsymbol{x}}_n \tag{37}$$

$$\mathbb{V}[\tilde{\boldsymbol{x}}_{kn}|\tilde{\boldsymbol{x}}_n]=\boldsymbol{Y}_{kn}-\boldsymbol{Y}_{kn}\boldsymbol{Y}_n^{-1}\boldsymbol{Y}_{kn} \tag{38}$$

このウィナーフィルタリングでは，式（21）とは異なり，$\tilde{\boldsymbol{X}}_k$の位相は$\tilde{\boldsymbol{X}}$の位相とは異なる点に注意する．ISNMFのように各周波数ビンn, mごとではなく，各フレームnごとに一挙に分離を行うことで，周波数ビン間の相関を考慮しながら高品質な分離が可能となる．

6．統計的音響信号処理の最先端

最後に，LPCとPSDTFを確率的に統合することで，LPC, NMF, CARをすべて包含した統一的な確率モデルを

構成できることを示す．以降，4.2項で示したCARの流れに沿って説明する．まず，ソースiの複素スペクトル$\tilde{\boldsymbol{s}}_i = \{S_i(e^{i\omega_m})\}_{m=1}^M$は，周波数ビン間の相関を考慮した複素ガウス分布に従うことを仮定する．

$$\tilde{\boldsymbol{s}}_i \sim \mathcal{N}_c(0, \boldsymbol{V}_i) \tag{39}$$

ここで，$\boldsymbol{V}_i$は共分散行列であり，CARにおける式(24)のように対角行列とは限らない．あるフレームnのソースiとフィルタjの組合せに起因する複素スペクトル$\tilde{\boldsymbol{x}}_{ijn} = \{X_{ijn}(e^{i\omega_m})\}_{m=1}^M$は，式(26)の線形性からやはり複素ガウス分布に従うことがわかる．

$$\tilde{\boldsymbol{x}}_{ijn} \sim \mathcal{N}_c(0, h_{ijn}(\mathrm{diag}(\tilde{\boldsymbol{f}}_j)\boldsymbol{V}_i \mathrm{diag}(\tilde{\boldsymbol{f}}_j)^H)) \tag{40}$$

ここで，$\tilde{\boldsymbol{f}}_j = \{F_j(e^{i\omega_m})\}_{m=1}^M$は，式(24)で定まるフィルタ$j$の複素スペクトルであり，分散行列を$\boldsymbol{Y}_{ijn} = h_{ijn}(\mathrm{diag}(\tilde{\boldsymbol{f}}_j)\boldsymbol{V}_i \mathrm{diag}(\tilde{\boldsymbol{f}}_j)^H)$としておく．混合音の複素スペクトル$\tilde{\boldsymbol{x}}_n = \{X_n(e^{i\omega_m})\}_{m=1}^M$は，あらゆるソース$i$とフィルタ$j$の組合せの和であるから，複素ガウス分布の再生性に着目すると，

$$\tilde{\boldsymbol{x}}_n \sim \mathcal{N}_c(0, \boldsymbol{Y}_n) \tag{41}$$

を得る．ここで，$\boldsymbol{Y}_n = \sum_{ij} \boldsymbol{Y}_{ijn}$とした．これは基底数が$\boldsymbol{IJ}$個のPSDTFに対して，LPCが確率的に組み入れられた統合モデルとなっている．これを最小化するには，2.3項および5.2項で紹介した乗法更新アルゴリズムを組合せて用いることができる．

さらに，各変数に事前分布を導入することで確率モデルのベイズ的な取り扱いも可能である[14]．このとき，近年着目されているノンパラメトリックベイズ理論を用いて，理論的には無限の複雑さを持つベイズモデルを構成できることもできる[15)6)]．具体的には，ソースの重み$\{\theta_i\}_{i=1}^I$およびフィルタの重み$\{\phi_i\}_{j=1}^J$を考え，$\boldsymbol{I}, \boldsymbol{J} \to \infty$の極限でほとんどの要素がゼロとなるようなスパースな学習を行うことにより，データに合わせて自動的に実行的な複雑さを決定することができる．

7．むすび

本稿では，ISダイバージェンス最小化という統一的な視点から，信号処理分野発祥のLPC，画像処理分野発祥のNMF，さらに機械学習技術を取り込みつつ音響信号処理分野で独自の進化を遂げたCARやPSDTFなどの最新の統計的音響信号処理技術について述べた．温故知新の言葉通り，古典的な音響理論を現代風に確率モデルとして再定式化することで，最先端の確率モデルのパーツとして組み入れるアプローチは非常に有望である．好例として，音声のF0の動きをよく説明できる藤崎モデル（1980年代に発表）をHMMの枠組みで再定式化を行い[16]，音響信号に対するF0推定のための確率モデルに組み入れた研究が挙げられる[17]．このアプローチは決して音響信号処理分野に限定されるものではなく，本解説が自然言語処理や画像処理などの他のメディア情報処理分野のさらなる発展に役立てば幸いである．

謝辞　本研究の一部は，JSPS科研費26700020，24220006，24700168およびJST CREST「OngaCRESTプロジェクト」の支援を受けた．貴重なアドバイスをくださった亀岡弘和氏（東京大学/NTT）に感謝する．

（2014年10月25日受付）

〔文　献〕

1) F. Itakura and S. Saito: "Analysis synthesis telephony based on the maximum likelihood method", Int. Cong. on Acoust., pp.C17-C20 (1968)
2) C. Févotte et al.: "Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis", Neural Computation, 21, 3, pp.793-830 (2009)
3) K. Yoshii et al.: "Infinite positive semidefinite tensor factorization for source separation of mixture signals", ICML, pp.576-584 (2013)
4) K. Yoshii et al.: "Beyond NMF: Time-domain audio source separation without phase reconstruction", ISMIR, pp.369-374 (2013)
5) H. Kameoka and K. Kashino: "Composite autoregressive system for sparse source-filter representation of speech", ISCAS, pp.2477-2480 (2009)
6) K. Yoshii and M. Goto: "Infinite composite autoregressive models for music signal analysis", ISMIR, pp.79- 84 (2012)
7) 亀岡弘和ほか："マルチカーネル線形予測モデルによる音声分析"，日本音響学会春季研究発表会，pp.499-502 (2010)
8) K. Yoshii and M. Goto: "Infinite kernel linear prediction for joint estimation of spectral envelope and fundamental frequency", ICASSP, pp.463-467 (2013)
9) 鹿野清宏ほか："音声認識システム"，オーム社 (2001)
10) R. Hennequin et al.: "NMF with time-frequency activations to model nonstationary audio events", IEEE TASLP, 19, 4, pp.744-753 (2011)
11) P. Smaragdis and J.C. Brown: "Non-negative matrix factorization for polyphonic music transcription", WASPAA, pp.177-180 (2003)
12) 亀岡弘和："非負値行列因子分解の音響信号処理への応用"，日本音響学会誌，68，11，pp.559-565 (2012)
13) B. Kulis, M. Sustik and I. Dhillon: "Low-rank kernel learning with Bregman matrix divergences", JMLR, 10, pp.341-376 (2009)
14) A.T. Cemgil: "Bayesian inference for nonnegative matrix factorisation models", Comp. Int. and Neurosci (2009)
15) M. Hoffman et al.: "Bayesian nonparametric matrix factorization for recorded music", ICML, pp.439-446 (2010)
16) 亀岡弘和ほか："音声F0パターン生成過程の確率モデル"，日本音響学会秋季研究発表会，pp.207-210 (2010)
17) 亀岡弘和："全極型声道モデルとF0パターン生成過程モデルを内部にもつ統一的音声生成モデル"，日本音響学会秋季研究発表会，pp.211-214 (2010)

吉井（よしい）　和佳（かずよし）　2008年，京都大学大学院情報学研究科博士後期課程修了．同年，産業技術総合研究所情報技術研究部門に入所．2014年，京都大学大学院情報学研究科講師に着任．音楽情報処理，統計的音響信号処理の研究に従事．博士（情報学）．

糸山（いとやま）　克寿（かつとし）　2011年，京都大学大学院情報学研究科博士後期課程修了．同年，同大学研究科助教に着任．音楽情報処理，音楽鑑賞インタフェース等の研究に従事．博士（情報学）．