Vol.2017-MUS-115 No.18
2017/6/17

ケプストラム距離正則化を用いた半教師ありステレオチャネル楽曲音源分離

関　翔悟[1,a)]　戸田　智基[1,b)]　武田　一哉[1,c)]

概要：本研究では，個別に収録/加工された多数の音源から構成されるステレオチャネル楽曲を対象とした，半教師あり音源分離手法を提案する．個別音源から人工的に合成される楽曲は，実環境下で同時収録される楽曲とは異なり，収録時の空間特性を表すチャネル信号間の位相（差）情報を音源分離の手がかりとして利用することが困難である．したがって提案法では，ステレオチャネル楽曲の振幅スペクトログラムに低ランク構造を仮定し，合成楽曲の生成過程を考慮した非負値テンソル因子分解（Non-negative Tensor Factorization: NTF）に基づくモデル化を行う．また，推定される音源が，楽曲内のそれぞれの楽器や歌声のような異なる音色をもつように，提案法では半教師あり音源分離の枠組みを導入し，各音源がそれぞれに固有なスペクトル包絡にしたがうように制約するケプストラム距離正則化（Cepstrum Distance Regularization）を導入する．実験的評価では，実環境で収録された個別音源より合成された楽曲を用いて分離性能を評価し，提案法の有効性を示すとともに，正則化の影響についても調査する．

1. はじめに

デスクトップオーディオやスマートフォン，ポータブルオーディオプレーヤなどで鑑賞される楽曲は，一般的にCD 音源やダウンロード配信により入手することが可能である．楽曲は通常，歌声や複数の楽器音から構成されており，それらが左右両耳に対応する 2 チャネル（ステレオチャネル）信号として表現されている．このようなステレオチャネル楽曲に対する音源分離は，楽曲の自動採譜 [1] や楽曲中のボーカル抽出 [2], [3] など，多様な応用先が期待される．

代表的な音源分離技術として，観測信号である混合音のみから，音源信号を推定し抽出するブラインド音源分離（Blind Source Separation: BSS）が，精力的に研究されている [4]．BSS は，観測される信号のチャネル数と推定する音源数の関係により問題設定が変化する．観測されるチャネル信号が推定する音源数以上である優決定条件下での BSS として，独立成分分析（Independent Conponent Analysis: ICA） [5] がある．ICA では，推定される個々の音源信号の統計的な独立性のみを仮定し，分離行列を推定することで，線形フィルタにより高性能な音源分離が可能である．ICA を拡張させた独立ベクトル分析（Independent Vector Analysis: IVA）[6], [7] では，周波数領域 ICA（Frequency Domain ICA: FDICA） [8] におけるパーミュテーション問題やスケーリング問題が解消され，より高精度な音源分離が可能である．一方で，観測信号数が音源数より少数である劣決定条件下での BSS の場合には，線形フィルタの設計が困難であるため，ICA 及び IVA による十分な分離性能を得ることは困難である．

代表的な劣決定 BSS 技術として，観測される混合音の振幅/パワースペクトログラムを非負値の行列とみなして振幅/パワースペクトル領域での加法性を仮定し，二つの行列の積として近似する非負値行列因子分解 (Non-negative Matrix Factorization: NMF) [9]–[11] に基づく音源分離手法が提案されている．NMF では，近似によって得られるモノラルチャネル信号の各時間周波数スロットに対して事前 SN 比を推定し，これを元に推定されるウィーナーフィルタを設計することで，劣決定条件下の BSS を解くことが可能である．この NMF に対して，マイクロフォンアレーなどの複数の観測信号を考慮するよう拡張した手法として，マルチチャネル NMF(Multichannel NMF: MNMF) [12], [13] が提案されている．MNMF では，NMF における基底行列およびアクティベーション行列で表される音源信号に関する音源情報に加えて，マイクロフォン間の配置から生じる空間情報を分離の手がかりとして導入することが可能である反面，高い初期値依存性が確認され

[1] 名古屋大学
Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, 464–8603, Japan
[a)] seki.shogo@g.sp.m.is.nagoya-u.ac.jp
[b)] tomoki@icts.nagoya-u.ac.jp
[c)] kazuya.takeda@nagoya-u.jp

Table 1: 従来法との比較

手法	観測チャネル数	BSS の条件	位相差情報
ICA	シングル	優決定	利用
IVA	マルチ	優決定	利用
NMF	シングル	劣決定	不要
MNMF	マルチ	劣決定	利用
IRLMA	マルチ	優決定	利用
提案法	マルチ	劣決定	不要

ている．これに対して，優決定条件下での BSS に限定した独立低ランク行列分析 (Independent Low-Rank Matrix Analysis: ILRMA) [14] では，MNMF において推定される空間情報に対して音源毎のチャネル相関に rank-1 構造を導入することで，MNMF における初期値依存性を解消し，高精度な分離性能を実現することが可能となる．

マルチチャネル信号に対する音源分離が可能となっている反面，分離の際には観測信号の位相情報を必要とする手法が大半であり，これを用いずに劣決定条件の音源分離問題を解くことは依然として困難である．例えば，個別に収録/加工された音源から構成される一般的な楽曲（合成楽曲とよぶ）を対象として音源分離を行う場合がこの条件に相当する．これはコンサートホール等で録音された楽曲（収録楽曲とよぶ）とは異なり，収録環境の空間特性を手がかりとすることができず，振幅情報のみが利用できる．

振幅情報のみが利用可能な音源分離問題において，分離の際に推定される音源に対して事前になんらかの情報が手に入れば，これを音源推定の手がかりとすることが可能である．BSS に対して，学習データ（事前知識）を全てもしくは一部を利用する音源分離は，それぞれ教師あり音源分離，半教師あり音源分離と呼ばれる [16]．教師あり音源分離の一つとして，教師あり NMF（Supervised NMF: SNMF）がある [17]．SNMF では，学習データから音源の音高情報（スペクトル調波構造）と音色情報（スペクトル包絡情報）を併せて学習，固定することで，学習データと類似する振幅スペクトル構造をもつ音源信号を高精度に分離することが可能である．一方で，学習データと評価データの対応する音源について，音高情報または音色情報のいずれかが異なる場合には，推定される音源に対応することが困難であり十分な分離性能が得られない．半教師あり音源分離の一つである半教師あり NMF(Semi-supervised NMF: SSNMF) [17], [18] では，教師データのスペクトル構造を利用し，変化させることで学習データと評価データにおいて対応する音源の差異を吸収することが可能である．さらに SSNMF において，ケプストラム距離正則化（Cepstrum Distance Regularization: CDR）が提案されている [19]．CDR では，推定される音源のスペクトル包絡が教師データのスペクトル包絡にしたがうようソフトに制約することで，分離性能を向上することが可能である．

本研究では，多数の音源から構成されるステレオチャネル楽曲に対する音源分離の実現を目的とする．提案法では，NMF と同様にステレオチャネル楽曲の振幅スペクトログラムに低ランク構造を仮定するとともに，楽曲の生成過程を考慮した非負値テンソル因子分解（Non-negative Tensor Factorization: NTF） [15] に基づくモデル化を行う（Table. 1）．また，半教師あり音源分離の枠組みを適用し，CDR による各音源の音色情報のみに関するソフトな制約を導入する．

2. 音源分離に関する従来研究

2.1 非負値行列因子分解（NMF）

NMF は観測行列に対する，非負制約つき低ランク行列表現手法である．観測信号の振幅/パワースペクトログラム $\boldsymbol{X} \in \mathbb{R}_{\geq 0}^{K \times N}$ に対して NMF を適用することで，少数のスペクトルパターンを表す基底行列 $\boldsymbol{T} \in \mathbb{R}_{\geq 0}^{K \times B}$ と対応するゲインの時間変化を表すアクティベーション行列 $\boldsymbol{U} \in \mathbb{R}_{\geq 0}^{B \times N}$ との積，

$$\boldsymbol{X} \approx \hat{\boldsymbol{X}} = \boldsymbol{T}\boldsymbol{U} \tag{1}$$

へと近似される．ここで K 及び N はそれぞれスペクトログラムの総周波数ビン数，総フレーム数を表す．このとき，$\boldsymbol{X}$ の周波数ビン $k \in \{1, \ldots, K\}$，時間フレーム $n \in \{1, \ldots, N\}$ の要素 x_{kn} に対応する $\hat{x}_{kn}$ は B 個の基底スペクトルの線形和として以下で表される．

$$\hat{x}_{kn} = \sum_{b=1}^{B} t_{kb} u_{bn} \tag{2}$$

2.2 音声/特徴量強調のためのケプストラム距離正則化

ケプストラム距離正則化では，音声強調として提案され，推定される強調音声が特徴量空間において，学習データの音声がもつ分布に従うよう制約する正則化項であり，以下で表される．

$$\mathcal{K}\left(\hat{\boldsymbol{\mathcal{X}}}\right) = -\log \prod_{m} \sum_{p} w_p \prod_{q} \mathcal{N}\left(E_{qm} : \mu_{pq}, \sigma_{pq}^2\right) \tag{3}$$

$$E_{qm} = \sum_{r} c_{qr} \log \sum_{k} f_{rk} \hat{x}_{kn} \tag{4}$$

ただし，E_{qm} は (2) で表される $\hat{x}_{kn}$ のメル周波数ケプストラム係数（Mel-Frequency Cepstrum Coefficients: MFCC）であり，$f_{r,k}$ は $r(\in \{1, \ldots, R\})$ 番目のメルフィルタバンク係数，$\{c_{q,r}\}_{0 \leq q \leq Q-1, 1 \leq r \leq R}$ は逆離散コサイン変換係数である．式 (3) はパラメータ $\left\{w_p, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p\right\}_{1 \leq p \leq P}$ の混合ガウス分布 (Gaussian Mixture Model: GMM) の対数尤度を表す．ただし，w_p，$\boldsymbol{\mu}_p = (\mu_{p,0}, \ldots, \mu_{p,Q-1})^{\mathsf{T}}$，$\boldsymbol{\Sigma}_p = \mathrm{diag}\left(\sigma_{p,0}^2, \ldots, \sigma_{p,Q-1}^2\right)$ は p 番目のガウス分布の重み，平均及び分散を表す．

ケプストラム距離正則化項において，パラメータ $\left\{w_p, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p\right\}_{1 \leq p \leq P}$ は，強調対象音声の学習データより

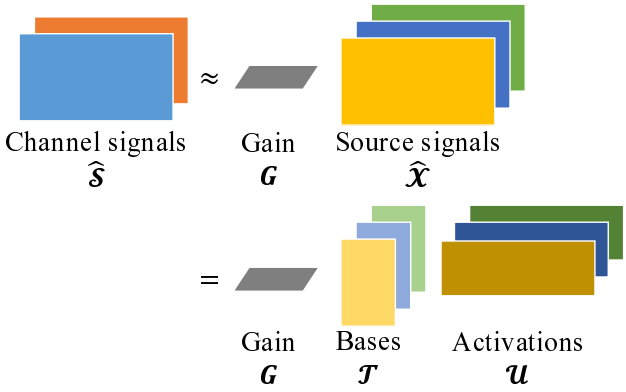


Figure 1: Source separation framework for stereophonic music signals

事前に学習されたものを固定して用いる．したがって推定される強調音声は，それぞれの学習データが特徴量空間上でとる分布にしたがうよう保証されることが期待される．またこの正則化項は，スペクトル包絡の確率モデルとしてGMM を用いていることから音声の特徴量空間におけるソフトなクラスタリング規準を与えると考えられる．

3. 提案法

3.1 NTF に基づく楽曲生成モデル

合成楽曲では，チャネル間の位相情報が手がかりとして利用することは困難である．したがって提案法では，観測されるステレオチャネル信号は，複数の音源に対して左右チャネルへの音量操作（以下，パンニングと表現する）を行い，それらを重畳することにより得られると仮定する．このとき，チャネル $c \in \{1,\ldots,C\}$ の信号の複素スペクトログラム $\boldsymbol{S}_c \in \mathbb{C}^{K\times N}$ および音源 $m \in \{1,\ldots,M\}$ の信号の複素スペクトログラム $\boldsymbol{X}_c \in \mathbb{C}^{K\times N}$ の集合をそれぞれ $\boldsymbol{\mathcal{S}}_c \in \mathbb{C}^{K\times N\times C}$，$\boldsymbol{\mathcal{X}}_c \in \mathbb{C}^{K\times N\times M}$ と表す．また，パンニングを行うゲインを $\boldsymbol{G} \in \mathbb{R}^{M\times C}_{\geq 0}$ とする．ここで，C は総チャネル数（$C=2$）を表す．提案法において仮定するステレオチャネル楽曲の生成過程は次式で表される．

$$\boldsymbol{\mathcal{S}}_c = f\left(\boldsymbol{G}, \boldsymbol{\mathcal{X}}_c\right) \tag{5}$$

(5) で示される混合過程に対して，振幅スペクトル領域における線形演算による以下の近似を導入する．

$$\boldsymbol{\mathcal{S}} \approx \hat{\boldsymbol{\mathcal{S}}} = f(\boldsymbol{G}, \hat{\boldsymbol{\mathcal{X}}}) \tag{6}$$

ここで $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{K\times N\times C}_{\geq 0}$，$\hat{\boldsymbol{\mathcal{S}}} \in \mathbb{R}^{K\times N\times C}_{\geq 0}$ および $\hat{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^{K\times N\times M}_{\geq 0}$ はそれぞれ，観測チャネル信号，推定チャネル信号及び推定音源信号の振幅スペクトログラムを表す．

推定チャネル信号 $\hat{\boldsymbol{\mathcal{S}}}$ が，ゲイン $\boldsymbol{G}$ および推定音源信号 $\hat{\boldsymbol{\mathcal{X}}}$ の線形演算によって得られるとき，音源分離は図 1 上で表されるように，ステレオチャネル信号をゲインおよび音源信号に分解する NMF と同様な枠組みとして解釈できる．提案法では，各音源信号 $\hat{\boldsymbol{X}}$ に対してさらに従来

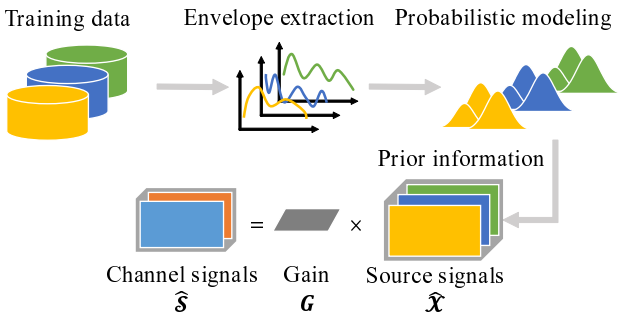


Figure 2: Overview of intrducing prior information for spectral envelopes of source signals

の NMF を適用することで，NMF をテンソルへと拡張した，NTF の枠組みとして各音源信号に分離する．各音源信号 $\hat{\boldsymbol{X}}$ が $\boldsymbol{T}$ および $\boldsymbol{U}$ で表されるとき，音源信号集合 $\hat{\boldsymbol{\mathcal{X}}}$ は基底集合 $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{K\times B\times M}_{\geq 0}$ およびアクティベーション集合 $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{B\times N\times M}_{\geq 0}$ で表すことができる．

以上より，チャネル信号の要素 $\hat{s}_{knc}$ および音源信号の要素 $\hat{x}_{knm}$ はそれぞれ以下で表される．

$$\begin{cases} \hat{s}_{knc} &= \sum_m g_{mc}\hat{x}_{knm} \\ \hat{x}_{knm} &= \sum_b t_{kbm}u_{bnm} \end{cases} \tag{7}$$

ただし g_{mc}，x_{knm}，t_{kbm} および u_{bnm} はそれぞれ，ゲイン $\boldsymbol{G}$，推定音源信号 $\hat{\boldsymbol{\mathcal{X}}}$，基底 $\boldsymbol{\mathcal{T}}$ およびアクティベーション $\boldsymbol{\mathcal{U}}$ の要素を表す．また g_{mc} および t_{kbm} に関しては，以下を満たすとする．

$$\sum_c g_{mc} = 1 \tag{8}$$

$$\sum_k t_{kbm} = 1 \tag{9}$$

3.2 音色情報に基づく正則化の導入

NTF に基づく楽曲生成モデルでは，従来の NMF に基づく音源分離と同様に，対象とする音源を表す基底スペクトルに異なる楽器音や歌声を表現する基底スペクトルが混入することが想定されるため，教師データを用いた半教師あり音源分離を行う．音色情報また，ケプストラム距離正則化を用いて，推定される音源集合 $\hat{\boldsymbol{\mathcal{X}}}$ が個別の音源に対応する教師データの各構成音源と類似した特徴（音色情報）を持つような制約を導入する．したがって，ステレオチャネル楽曲から構成音源を分離する提案法は，目的関数

$$\mathcal{I}(\boldsymbol{\theta}) = \sum_{k,n,c} \mathcal{D}(s_{knc}|\hat{s}_{knc}) + \lambda\mathcal{K}(\hat{\boldsymbol{\mathcal{X}}}) \tag{10}$$

を最小化する最適化問題として定式化することができる．ここで $\boldsymbol{\theta} = \{\boldsymbol{G}, \boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{U}}\}$ は推定する未知パラメータの集合であり，$\mathcal{D}(\cdot)$ は規準情報量を表し，本稿では Kullback-Libler (KL) ダイバージェンス

$$\mathcal{D}_{\mathrm{KL}}(y|x) = y\log\frac{y}{x} - (y-x) \tag{11}$$

を用いる．λ は正則化パラメータを表し，ケプストラム距離正則化項

$$\mathcal{K}\left(\hat{\boldsymbol{\mathcal{X}}}\right) = -\log\prod_{m,n}\sum_{p} w_{pm}\prod_{q}\mathcal{N}\left(E_{qnm}:\mu_{pqm},\Sigma_{pqm}\right) \tag{12}$$

$$E_{qnm} = \sum_{r} c_{qr}\log\sum_{k} f_{rk}\hat{x}_{knm} \tag{13}$$

である．

3.3 パラメータ推定

$\mathcal{I}(\boldsymbol{\theta})$ を最小化するパラメータ $\boldsymbol{G}$，$\boldsymbol{\mathcal{T}}$，$\boldsymbol{\mathcal{U}}$ を解析的に得ることは困難であるが，補助関数法 [20] に基づき，$\mathcal{I}(\boldsymbol{\theta})$ の停留点への収束が保証された反復更新アルゴリズムを導くことが可能である．補助関数法による目的関数を最小化するパラメータ推定では，補助変数 $\bar{\boldsymbol{\theta}}$ を導入した $\mathcal{I}(\boldsymbol{\theta}) = \min_{\bar{\boldsymbol{\theta}}}\mathcal{I}^{+}(\boldsymbol{\theta},\bar{\boldsymbol{\theta}})$ を満たす補助関数を設計し，以下をを交互に反復することで $\mathcal{I}(\boldsymbol{\theta})$ を局所最適化する $\boldsymbol{\theta}$ を得ることができる．

(1) $\bar{\boldsymbol{\theta}} = \mathrm{argmin}_{\bar{\boldsymbol{\theta}}}\mathcal{I}^{+}(\boldsymbol{\theta},\bar{\boldsymbol{\theta}})$
(2) $\boldsymbol{\theta} = \mathrm{argmin}_{\boldsymbol{\theta}}\mathcal{I}^{+}(\boldsymbol{\theta},\bar{\boldsymbol{\theta}})$

したがって補助関数法により，ゲインの更新式は以下となる．

$$g_{mc} = \frac{\sum_{k,b,n} s_{knc}\alpha_{kbnmc}}{t_{kbm}u_{bnm}} \tag{14}$$

ただし，α_{kbnmc} は $\sum_{b,m}\alpha_{kbnmc} = 1$ を満たす非負の補助変数であり，以下で表される．

$$\alpha_{knmc} = \frac{g_{mc}t_{kbm}u_{bnm}}{\hat{s}_{knc}} \tag{15}$$

同様に，基底およびアクティベーションの更新式は以下となる．

$$t_{kbm} = \frac{-\mathsf{b}_{kbm} + \sqrt{\mathsf{b}_{kbm}^2 - 4\mathsf{a}_{kbm}\mathsf{c}_{kbm}}}{2\mathsf{a}_{kbm}} \tag{16}$$

$$u_{bnm} = \frac{-\mathsf{f}e_{bnm} + \sqrt{\mathsf{e}_{bnm}^2 - 4\mathsf{d}_{bnm}\mathsf{f}_{bnm}}}{2\mathsf{d}_{bnm}} \tag{17}$$

ここで a_{kbm}，b_{kbm}，c_{kbm}，d_{bnm}，e_{bnm}，f_{bnm} はそれぞれ，

$$\mathsf{a}_{kbm} = \sum_{n,c} g_{mc}u_{bnm} + \lambda\sum_{r,n}\Bigg[A_{rnm}p(\xi_{rnm})f_{rk}u_{bnm} + \delta_{B_{rnm}\geq 0}|B_{rnm}|\frac{f_{rk}u_{bnm}}{\zeta_{rnm}}\Bigg] \tag{18}$$

$$\mathsf{b}_{kbm} = -\sum_{n,c} s_{knc}\alpha_{kbnmc} - \lambda\sum_{r,n}\delta_{B_{rnm}<0}\left|B_{rnm}\right|\psi_{rkbnm} \tag{19}$$

$$\mathsf{c}_{kbm} = -\lambda\sum_{r,n} A_{rnm}\frac{\phi_{rkbnm}^2}{f_{rk}u_{bnm}} \tag{20}$$

$$\mathsf{d}_{bnm} = \sum_{k,c} g_{mc}t_{kbm} + \lambda\sum_{r,k}\Bigg[A_{rnm}p(\xi_{rnm})f_{rk}t_{kbm} + \delta_{B_{rnm}\geq 0}|B_{rnm}|\frac{f_{rk}t_{kbm}}{\zeta_{rnm}}\Bigg] \tag{21}$$

$$\mathsf{e}_{bnm} = -\sum_{k,c} s_{knc}\alpha_{kbnmc} - \lambda\sum_{r,k}\delta_{B_{rnm}<0}\left|B_{rnm}\right|\psi_{rkbnm} \tag{22}$$

$$\mathsf{f}_{bnm} = -\lambda\sum_{r,k} A_{rnm}\frac{\phi_{rkbnm}^2}{f_{rk}t_{kbm}} \tag{23}$$

となり，$A_{rnm}, B_{rnm}, p(\xi_{rnm})$ は以下である．

$$A_{rnm} = \sum_{p,q}\frac{\beta_{pnm}c_{qr}^2}{2\sigma_{pqm}^2\omega_{pqrnm}}, \tag{24}$$

$$B_{rnm} = -\sum_{p,q}\frac{\beta_{pnm}c_{qr}\gamma_{pqrnm}}{\sigma_{pqm}^2\omega_{pqrnm}}, \tag{25}$$

$$p(\xi_{rnm}) = (\log\xi_{rnm})^2 - 2\log\xi_{rnm} - \frac{2}{\xi_{rnm}} \tag{26}$$

ただし，β_{pnm}, γ_{pqrnm}, ξ_{rnm}, ζ_{rnm}, ϕ_{rkbnm}, ψ_{rkbnm} は補助変数であり，

$$\beta_{pnm} = \frac{w_{pm}\prod_{q}\mathcal{N}(E_{qnm};\mu_{pqm}\sigma_{pqm}^2)}{\sum_{p'}w_{p'm}\prod_{q'}\mathcal{N}(E_{q'nm};\mu_{p'q'm},\sigma_{p'q'm}^2)} \tag{27}$$

$$\gamma_{pqrnm} = c_{qr}\log\varsigma_{rnm} + \omega_{pqrnm}(\mu_{pqm} - E_{qnm}) \tag{28}$$

$$\xi_{rnm} = \zeta_{rnm} = \sum_{k,b} f_{rk}t_{kbm}u_{bnm}(=\varsigma_{rnm}) \tag{29}$$

$$\phi_{rkbnm} = \psi_{rkbnm} = \frac{f_{rk}t_{kbm}u_{bnm}}{\sum_{k',b'}f_{rk'}t_{k'b'm}u_{b'nm}} \tag{30}$$

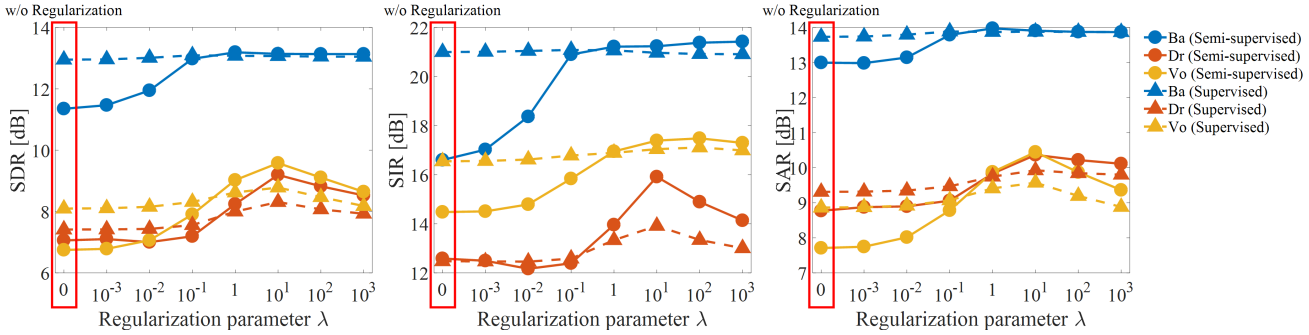
を満たす．また，w_{pqrnm} は $\sum_{r} w_{pqrnm} = 1$ を満たす任意の正の定数である．

(14), (16), (17) によりパラメータを交互に更新していき，推定値を得た後，各チャネル信号における事前 SN 比を推定し，ウィナーフィルタを作成，ステレオチャネル信号に対して適用することで，各音源に対する分離信号を得る．ただし，ゲイン及び基底集合の初期値は学習データを用いた SSNMF により求めることで，異なる音源を表現する基底スペクトルの発生，および音源間のパーミュテーションの発生を防止する．

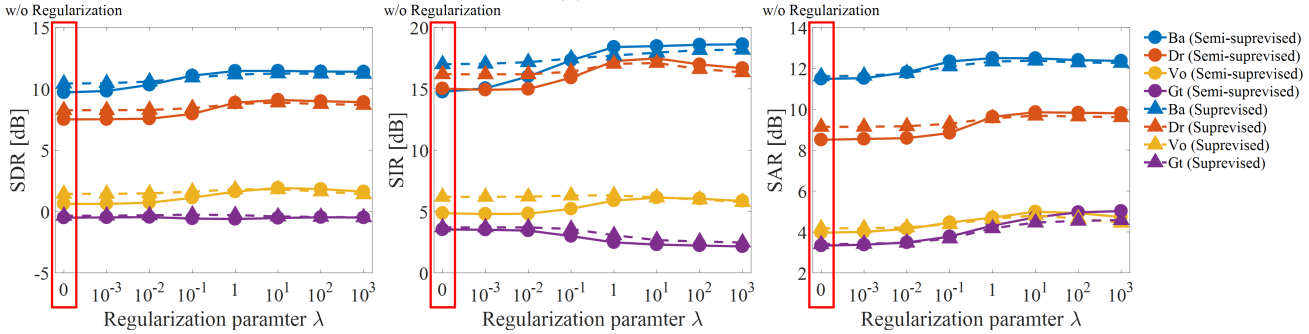
4. 実験的評価

4.1 使用データ

Cambridge Music Technology [21] で配布されている楽曲データを用いる．各楽曲データは音源ごとにサンプリング周波数 44.1 kHz で個別収録されており，本稿では Bass, Drums, Vocal, Guitar, Other という 5 パートへと個別の音源を混合・統合する．実験においては 5 パートの内，Bass, Drums，Vocal, Guitar の計 4 パートを各音源信号とする．各パート音源に対してチャネル間の平均をとることでモノ



(a) 3-source separation



(b) 4-source separation

Figure 3: Separation performance for each source signal in semi-supervised and supervised separation. The cepstrum regularization is not used if the regularization parameter λ set to 0.

Table 2: Music list

ID	Artist	Title	Duration
1	Actions	Devil's Words	3'17"
2	Actions	One Minute Smile	2'44"
3	Actions	South of The Water	3'11"

ラル信号とする．利用する楽曲リストを表 2 に示す．表 2 に示される，楽曲 1 を評価データ及び開発データとする．各パート音源の冒頭 30-45 s を開発データとして利用する．パートごとにモノラル化された音源信号 Bass, Drums, Vocal, Guitar に対して左右にそれぞれ，2:1, 1:2, 1:1, 2:1 とパンニングを適用し，得られたステレオチャネル信号を評価楽曲として作成する．評価楽曲のうち，冒頭 50-65 s を評価データとする．また，楽曲 2,3 のパート音源を学習データとする．

4.2 実験条件

楽曲はサンプリング周波数を 16 kHz へとダウンサンプリングしたのちに利用する．スペクトログラム分析ではフレームサイズ 32 ms, シフトサイズ 16 ms とする．各音源に対応する基底数は 50 とし，パラメータはそれぞれ 200 更新を行い，ケプストラム距離正則化におけるフィルタバンク数は 64 とする．また，ケプストラム距離正則化で利用する GMM のパラメータについては，予備実験を行い各音源ごとに開発データに最適なものを選択する．すなわち，MFCC の次元数 Q 及び GMM 混合数 P は音源 m 毎に異なる $(Q \to Q(m), P \to P(m))$.

正則化項の影響を調査するため，正則化パラメータ λ に対して複数の場合（$\lambda = 0, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}$）を設定し評価する．提案法をケプストラム距離正則化を用いた半教師あり音源分離，従来法をケプストラム距離正則化を用いた教師あり音源分離として比較を行う．また，いずれの場合もケプストラム距離正則化を導入しない場合についても評価する．提案法では，全てのパラメータについて更新を行い，従来法では各音源の基底のみ固定し他のパラメータの更新を行う．評価尺度は，パンイング後の各音源に対するステレオチャネル信号との信号対歪み比（Signal-to-Distortion Ratio: SDR），信号対干渉比（Signal-to-Interference Ratio: SIR），信号対加工比（Signa-to-Artifact Ratio: SAR）を用い，それぞれ BSS Toolbox [22] より算出し，ステレオチャネルの平均をとる．SDR，SIR, SAR はそれぞれ，分離音の音質，分離音に含まれる非目的音の抑圧度合，分離処理により生じる歪みの少なさを表し，いずれにおいても大きい値となる場合高性能である．乱数初期値の影響を考慮し，各実験条件に対して 10 回ずつ分離を行いその平均を評価する．

4.3 実験結果

実験結果を Fig. 3 に示す．Fig. 3(a) は 3 音源の合成楽曲の場合，Fig. 3(b) は 4 音源の合成楽曲の場合のそれぞれの分離性能を表し，左から SDR，SIR，SAR を表す．それぞれの図において，実線が提案法，破線が従来法を示し，赤枠部は正則化がない場合（$\lambda = 0$）の結果を表す．

正則化パラメータを適当に設定することにより（$\lambda = 1 - 10^2$ 程度），提案法および提案法のいずれにおいてもで分離性能の向上が確認される．特に提案法では従来法に比べて，大幅な性能向上がみられる．このことから，従来法が一部のパラメータを固定して最適化しており，学習データと評価データの差異に対処が困難で分離性能が制限される一方，提案法がデータ間の違いに柔軟に対応していることがわかる．また，正則化を行わない場合と比較し，提案法では大幅な性能向上がみられることから，ケプストラム正則化がパラメータ更新に有効な影響を与えていることがわかる．4 音源分離の場合には，3 音源の場合と異なり大幅な性能向上はみられないものの，提案法が従来法を上回ることが確認される．

5. おわりに

本研究では，個別に収録/加工された多数の音源から構成されるステレオチャネル楽曲を対象とした，半教師あり音源分離手法を提案した．提案法では，ステレオチャネル楽曲の振幅スペクトログラムに低ランク構造を仮定し，合成楽曲の生成過程を考慮した NTF に基づくモデル化を行うことで，観測チャネル信号間の位相情報が不要である．また，推定される音源が，楽曲内のそれぞれの楽器や歌声のような異なる音色をもつように，提案法では半教師あり音源分離の枠組みを導入するとともに，各音源がそれぞれに固有なスペクトル包絡にしたがうように制約するケプストラム距離正則化を導入した．実験的評価において，3 音源もしくは 4 音源のステレオチャネル楽曲を用いて，分離性能と評価したところ，従来の教師あり音源分離アプローチが学習・評価データ間のミスマッチに対処することが困難な一方，提案法はミスマッチに柔軟に対応することが可能であり，大幅な分離性能の向上がみられ，有効性が示された．また，正則化を適切に導入することで，分離性能向上が確認された．

謝辞 本研究の一部は，JSPS 科研費 17H01763 により実施したものである．

参考文献

[1] Paris Smaragdis and Judith C Brown, “Non-negative matrix factorization for polyphonic music transcription,” in *Proc. of WASPAA*, pp. 177–180, 2003.

[2] Shankar Vembu and Stephan Baumann, “Separation of vocals from polyphonic audio recordings,” in *Proc. of ISMIR*, pp. 337–344, 2005.

[3] Yukara Ikemiya, Kazuyoshi Yoshii, and Katsutoshi Itoyama, “Singing voice analysis and editing based on mutually dependent f0 estimation and source separation,” in *Proc. of ICASSP*, pp. 574–578, 2015.

[4] Naik, Ganesh R., and Wenwu Wang, *Blind source separation*, Springer, 2014.

[5] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja, *Independent component analysis*, John Wiley & Sons, 2004.

[6] Taesu Kim, Torbjørn Eltoft, and Te-Won Lee, “Independent vector analysis: An extension of ICA to multivariate components,” in *Proc. of ICA*, pp. 165–172, 2006.

[7] Atsuo Hiroe, “Solution of permutation problem in frequency domain ICA, using multivariate probability density functions,” in *Proc. of ICA*, pp. 601–608, 2006.

[8] Hiroshi Saruwatari, Toshiya Kawamura, and Kiyohiro Shikano, “Blind source separation for speech based on fast-convergence algorithm with ICA and beamforming,” *IEEE Trans. on ASLP*, vol. 14, no. 2, pp. 666–678, 2006.

[9] Daniel D Lee and H Sebastian Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[10] Hirokazu Kameoka, Nobutaka Ono, Kunio Kashino, and Shigeki Sagayama, “Complex NMF: A new sparse representation for acoustic signals,” in *Proc. of ICASSP*, pp. 3437–3440, 2009.

[11] Paris Smaragdis, “Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs,” in *Proc. of ICA*, pp. 494–499, 2004.

[12] Alexey Ozerov and Cédric Févotte, “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation,” *IEEE Trans. on ASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[13] Hiroshi Sawada, Hirokazu Kameoka, Shoko Araki, and Naonori Ueda, “Multichannel extensions of non-negative matrix factorization with complex-valued data,” *IEEE Trans. on ASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[14] Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari, “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,” *IEEE/ACM Trans. on ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[15] Cichocki Andrzej, Zdunek Rafal, Phan Anh Huy and Amari Shun-ich, “Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation.” John Wiley & Sons, 2009.

[16] Paris Smaragdis, Raj Bhiksha, and Shashanka Madhusudana, “Supervised and semi-supervised separation of sounds from single-channel mixtures,”in *Proc. of ICA*, pp. 414–421, 2007.

[17] Bryan Nicholas J., and Gautham J. Mysore, “Interactive refinement of supervised and semi-supervised sound source separation estimates,” in *ICASSP*, pp. 883–887, 2013.

[18] Augustin Lefevre, Francis Bach and Cédric Févotte, “Semi-supervised {NMF} with time-frequency annotations for single-channel source separation,” in *Proc. of ISMIR*, pp. 115–120, 2012.

[19] Li Li, Hirokazu Kameoka, Takuya Higuchi, and Hiroshi Saruwatari, “Semi-supervised joint enhancement of spectral and cepstral sequences of noisy speech,” in *Proc. of Interspeech*, pp. 3753–3757, 2016.

[20] Daniel D Lee and H Sebastian Seung, “Algorithms for non-negative matrix factorization,” in *Proc. of NIPS*, pp. 556–562, 2001.

[21] “Cambridge music technology,” `http://cambridge-mt.com/ms-mtk.htm`, Accessed: 2017-05-27.

[22] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. on ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.