教師なし伝達関数ゲイン基底 NMF による目的音強調における罰則項の特性評価 *

☆千葉大将 (筑波大)，小野順貴 (NII/総研大)，宮部滋樹，山田武志，牧野昭二 (筑波大)

1 はじめに

非同期分散型マイクロホンアレーはスマートフォンやボイスレコーダなどの身の回りの非同期録音機器を用いてアレー信号処理を行う枠組みであり，機器構成の柔軟性に富み，簡易に優れた SN 比での収音が期待できる．しかし，非同期機器間での観測が同期していないため録音開始時刻の差やサンプリング周波数のずれ（サンプリング周波数ミスマッチ）が起こり，特にサンプリング周波数ミスマッチは各観測信号間での位相差を時間とともに変化させるため，従来の位相情報に依存したアレー信号処理の性能が劣化してしまう [1, 2]．そこで，非同期録音に対する同期補正 [3, 4] が研究されているが，同期誤差がアレー信号処理の性能に影響する．一方，非同期録音用の目的音強調手法として，SN 比最大化ビームフォーマ [5] を振幅情報のみで行う振幅スペクトルビームフォーマ [6] のような，各音源の位相情報に依存しない振幅ベースの強調手法が提案されている．このような振幅情報のみを利用する目的音強調手法の一つとして，チャネル間の振幅もしくはパワー比（伝達関数ゲイン）を基底とする非負値行列因子分解 (NMF: Non-negative Matrix Factorization) (伝達関数ゲイン基底 NMF) による強調手法 [7] がある．前回の発表において，我々は各音源の単一音源区間より伝達関数ゲイン基底を学習する教師あり伝達関数ゲイン基底 NMF による目的音強調手法を提案し，伝達関数ゲイン基底 NMF が非同期録音に対して頑健な手法であることを示した [8]．本研究では，教師なし伝達関数ゲイン基底 NMF の性能を教師ありの場合まで向上させることを目的とし，まず，伝達関数ゲイン基底 NMF において伝達関数ゲイン及び音源アクティベーションを推定するために必要な制約について議論を行う．そして，戸上らが音源アクティベーションに対してスパースネス制約を導入することでブラインド分離を可能にしていることに着目し，実収録した非同期音声録音を用いてスパースネス制約を制御する罰則項の特性評価を行う．評価では，教師なし伝達関数ゲイン基底 NMF に 5 種類のスパースネス制約を導入し，罰則項の強さを変化させて教師ありの場合と強調性能を比較する．また，振幅領域及びパワー領域で評価を行うことで，伝達関数ゲイン基底 NMF による目的音強調においてどちらの領域での適用が適当であるか調査する．

2 伝達関数ゲイン基底 NMF を用いた時間周波数マスキング

2.1 振幅領域での混合モデル

本稿では時間周波数領域での信号を扱う．また，i 行 j 列に非負の実数 B_{ij} を成分として持つ $I \times J$ の行列 $\mathbf{B}$ を $\mathbf{B} = [B_{ij}]_{ij} \in \mathbb{R}_+^{I \times J}$ と表すこととする．なお，$\mathbb{R}_+$，$\mathbb{C}$ はそれぞれ非負の実数集合と複素数の集合を表す．

非同期録音では，機器間のサンプリング周波数ミスマッチによって観測信号間の位相ずれが発生する．この位相ずれのサンプル数は STFT フレーム長より十分に小さいと考えると，伝達関数の振幅は時不変であると仮定できる．そこで，観測信号の振幅もしくはパワースペクトルに対して加法性を仮定した，以下の線形混合モデルを導入する．

$$\begin{aligned} \bar{\mathbf{X}}(\omega) &= \left[\bar{X}_{mn}(\omega)\right]_{mn} \in \mathbb{R}_+^{M \times N} && (1)\\ &\approx \bar{\mathbf{A}}(\omega)\bar{\mathbf{S}}(\omega) && (2)\\ \bar{\mathbf{A}}(\omega) &= \left[\bar{A}_{mk}(\omega)\right]_{mk} \in \mathbb{R}_+^{M \times K} && (3)\\ \bar{\mathbf{S}}(\omega) &= \left[\bar{S}_{kn}(\omega)\right]_{kn} \in \mathbb{R}_+^{K \times N} && (4) \end{aligned}$$

ここで，ω は周波数ビン番号，n 番目の時間フレームにおける m 番目のマイクでの観測信号の振幅もしくはパワーを $\bar{X}_{mn}(\omega)$，k 番目の音源から m 番目のマイクまでの伝達関数の振幅もしくはパワーを $\bar{A}_{mk}(\omega)$，n 番目の時間フレームにおける k 番目の音源信号の振幅もしくはパワーを $\bar{S}_{kn}(\omega)$ と表す．また，K, M, N はそれぞれ音源数，マイク数，時間フレーム数を表す．このような振幅もしくはパワー領域での混合モデルは NMF を用いる際によく定式化されている [9]．そこで，時間チャネル領域における NMF（伝達関数ゲイン基底 NMF）によって，

$$\bar{\mathbf{X}}(\omega) \approx \tilde{\mathbf{X}}(\omega) = \tilde{\mathbf{A}}(\omega)\tilde{\mathbf{S}}(\omega) \qquad (5)$$

として観測信号 $\bar{\mathbf{X}}$ を低ランク近似する伝達関数ゲイン基底 $\tilde{\mathbf{A}}$ と音源アクティベーション $\tilde{\mathbf{S}}$ を得る [7]．

以降では，周波数ビンごとに同様のモデル化と処理を行うため周波数ビン番号を表す記号 ω を省略する．

2.2 時間周波数マスキングによる目的音強調

振幅の重ね合わせによる推定誤差に頑健な強調を行うため，伝達関数ゲイン基底 NMF より得られた伝達関数ゲイン基底 $\tilde{\mathbf{A}}$ と音源アクティベーション $\tilde{\mathbf{S}}$ を用いた時間周波数領域でのウィーナーマスクによ

---

* Characteristic evaluation of sparsity penalties for amplitude-based speech enhancement with unsupervised non-negative matrix factorization. by Hironobu CHIBA (University of Tsukuba), Nobutaka ONO (National Institute of Informatics / The Graduate University for Advanced Studies), Shigeki MIYABE, Takeshi YAMADA, Shoji MAKINO (University of Tsukuba)

る強調を行う．本稿では，k 番目の音源の SN 比が最も高い観測信号である $X_{kn} \in \mathbb{C}$ に対して k 番目の音源を強調するウィーナーマスクをかけ，各音源の強調信号 $\tilde{\mathbf{Y}} = \left[\tilde{Y}_{kn}\right]_{kn} \in \mathbb{C}^{K\times N}$ を得る．具体的には，NMF を振幅領域で適用した場合は，

$$\tilde{Y}_{kn} = \frac{\left(\tilde{A}_{kk}\tilde{S}_{kn}\right)^2}{\sum_i \left(\tilde{A}_{ki}\tilde{S}_{in}\right)^2} X_{kn} \tag{6}$$

であり，また，パワー領域で適用した場合は，

$$\tilde{Y}_{kn} = \frac{\tilde{A}_{kk}\tilde{S}_{kn}}{\sum_i \tilde{A}_{ki}\tilde{S}_{in}} X_{kn} \tag{7}$$

として強調信号を得る．

3 伝達関数ゲイン基底 NMF へのスパースネス制約の導入

3.1 伝達関数ゲイン基底 NMF による目的音強調に必要な制約

NMF では，非負値制約下の距離最小化規準で加法性の構成成分への分解を行う．従って，伝達関数ゲイン NMF による伝達関数ゲイン及び音源アクティベーションの推定では，マイク数と音源数の関係によっては音源を分離しない解が最適解となってしまう．

マイク数より音源数が大きい劣決定系 ($M < K$) やマイク数と音源数が等しい決定系 ($M = K$) では，観測信号の振幅スペクトル $\bar{\mathbf{X}}$ と推定された振幅スペクトル $\tilde{\mathbf{X}}$ の誤差が 0 となる，以下のような自明解 $\hat{\mathbf{A}}$，$\hat{\mathbf{S}}$ が存在する．

$$\bar{\mathbf{X}} = \tilde{\mathbf{X}} = \hat{\mathbf{A}}\hat{\mathbf{S}} \tag{8}$$

さらに，このような自明解が 1 つ存在すれば，以下のような無数の自明解 $\tilde{\mathbf{A}}$，$\tilde{\mathbf{S}}$ が存在する．

$$\tilde{\mathbf{A}} = \hat{\mathbf{A}}\mathbf{P},\ \tilde{\mathbf{S}} = \mathbf{P}^{-1}\hat{\mathbf{S}} \tag{9}$$

ここで，$\mathbf{P}$ は ($\hat{\mathbf{A}}\mathbf{P}$) 及び ($\mathbf{P}^{-1}\hat{\mathbf{S}}$) が非負行列となる $K \times K$ の任意の行列を表す．これらの解は一般に音源分離とはかけ離れたものとなるため，$M \leq K$ では伝達関数ゲイン基底 NMF における伝達関数ゲイン及び音源アクティベーションの推定は原理的に難しい．そこで，伝達関数ゲイン基底 NMF で音源を分離するためには，NMF 自体の非負値制約に加えて，何らかの制約が必要である．

音源数よりマイク数が大きい優決定系 ($M > K$) という制約下では，NMF は少ない基底で観測を表現しようとするため，伝達関数ゲイン基底及び音源アクティベーションの推定が可能となる．しかし，マイク数が音源数に近い場合は，$M \leq K$ における無意味な解に近い解が最適解となるため推定性能は期待できない．また，振幅の重ね合わせによりマイク数が小さいほど音源アクティベーションの任意性が大きくなる．従って，優決定系という制約下で伝達関数ゲイン基底 NMF による強調性能を向上させるには，マイク数が音源数を十分に上回っている必要がある．

以上の議論から，伝達関数ゲイン基底 NMF によって十分な目的音強調効果を得るには，音源を分離しない無意味な解を避けるために音源アクティベーションの任意性を制限する必要がある．従って，振幅の重ね合わせに関する制約をかけることが有効であると考えられる．その一つとして，我々が提案した教師あり伝達関数ゲイン基底 NMF のような単一音源区間を用いて伝達関数ゲイン基底を事前学習する方法が考えられる．単一音源区間において基底数 1 で伝達関数ゲイン基底 NMF を行う場合，振幅の重ね合わせが起こらないため高い精度で伝達関数ゲイン基底を推定できることが確認されている．一方，戸上らは時間フレームごとの音源アクティベーションの $L_{0.5}$ ノルムによるスパースネス制約を導入した教師なし伝達関数ゲイン基底 NMF による目的音強調効果を優決定系の観測データで確認している．このスパースネス制約は，振幅の重ね合わせに対してペナルティを与えるため，NMF の非負値制約と組み合わさることで自然と音源が分離される解が得られる．従って，決定系における無意味な解に近い場合コストが大きくなるため，決定系でも分離が可能となる．そこで，本稿では伝達関数ゲイン基底 NMF において音源を分離するためのスパースネス制約の強さと制約方法について調査を行う．

3.2 罰則付き伝達関数ゲイン基底 NMF

β ダイバージェンス規準 NMF [10, 11] において，時間フレームごとの音源アクティベーション $\tilde{\mathbf{S}}$ に対するスパースネス制約を評価する関数 $g(\tilde{\mathbf{S}})$ に，非負の係数 λ をかけた罰則項を加えた目的関数は以下のように表される．

$$\mathcal{J}(\bar{\mathbf{X}}, \tilde{\mathbf{A}}\tilde{\mathbf{S}}, \lambda) = \mathcal{D}_\beta(\bar{\mathbf{X}}|\tilde{\mathbf{A}}\tilde{\mathbf{S}}) + \lambda g(\tilde{\mathbf{S}}) \tag{10}$$

$$\mathcal{D}_\beta(y|x) = \frac{1}{\beta(\beta-1)}\left(y^\beta + (\beta-1)x^\beta - \beta y x^{\beta-1}\right) \tag{11}$$

このように目的関数に複数の項がある場合，目的関数が入力信号のスケールに対して非依存となるように各項の次元量は一致していることが望ましい．従って，本稿では入力信号のスケールに非依存となるようなダイバージェンス項と罰則項の組み合わせを選択する．罰則項としては Table 1 で表される，次元量が等しい 5 種類のスパースネス制約を用いる．距離規準としては，用いるスパースネス制約の次元量に対応させ，$\beta = 1$，すなわち I ダイバージェンス規準とする．従って，本稿では以下の罰則付き乗法型更新式 [12, 13] を用いて局所解を求める．

$$\tilde{A}_{mk} \leftarrow \tilde{A}_{mk} \frac{\frac{\sum_n \bar{X}_{mn}\tilde{S}_{kn}}{\tilde{X}_{mn}}}{\sum_n \tilde{S}_{kn}} \tag{12}$$

$$\tilde{S}_{kn} \leftarrow \tilde{S}_{kn} \frac{\sum_m \frac{\bar{X}_{mn}\tilde{A}_{mk}}{\tilde{X}_{mn}}}{\sum_m \tilde{A}_{mk} + \lambda \nabla g(\tilde{S}_{kn})} \tag{13}$$

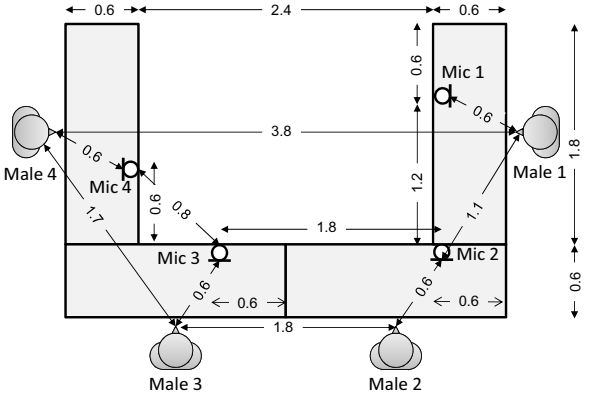

Fig. 1　Arrangement of microphones used in evaluation.

ここで，$\tilde{X}_{mn} = \sum_k \tilde{A}_{mk}\tilde{S}_{kn}$ は更新ごとに推定された観測信号の振幅もしくはパワースペクトルを表す．また，式 (13) の分母の $\nabla g(\tilde{S}_{kn})$ は式 (10) における $g(\tilde{S}_{kn})$ の勾配を表している．なお，式 (12)，(13) の更新ごとに，

$$\tilde{A}_{mk} \leftarrow \frac{\tilde{A}_{mk}}{\sum_k \tilde{A}_{mk}}, \tag{14}$$

$$\tilde{S}_{kn} \leftarrow \left(\sum_k \tilde{A}_{mk}\right)\tilde{S}_{kn} \tag{15}$$

として基底を正規化する．以上の処理により，周波数ビンごとに伝達関数ゲインを表す基底行列が得られるが，周波数ビンごとに基底ベクトルの順番が入れ替わるパーミュテーション問題が起こる．そこで，本稿では伝達関数ゲイン基底の初期値設定によってパーミュテーション問題の発生を抑制する．具体的には，各マイクにおける非目的信号の伝達関数ゲインの値は目的信号の伝達関数ゲインよりも小さいと仮定できるため，k 番目の音源を目的音とするマイク番号は k であるとし，伝達関数ゲイン基底 $\tilde{\mathbf{A}}$ の初期値を，

$$\tilde{A}_{mk} = \begin{cases} 1 & (m = k) \\ \alpha & (m \neq k) \end{cases} \tag{16}$$

として与える．ここで，パラメタ α は非目的信号の伝達関数ゲインの初期値であり，$\alpha < 1$ となる任意の正の実数である．さらに，ハンガリー法を用いて，推定した伝達関数ゲイン基底行列の対角成分を最大化する規準でパーミュテーションを解決する．

4　目的音強調性能の罰則項による特性評価

4.1　実験条件

非同期分散型マイクロホンアレーを用いた音声録音データを用いて，教師なし伝達関数ゲイン基底 NMF による目的音強調の罰則項による特性評価を行った．録音データは，Fig. 1 のようなマイク・話者配置とし，同期された分散型マイクロホンアレーにより話者ごとに Table 2 のような環境で録音した．録音後，マイク毎に Table 3 に示すサンプリング周波数でリサンプリングを行い人工的に非同期録音データを生成した．実験条件を Table 4 に示す．一般的にフレーム長が長いほど位相ずれに頑健であることから，本稿では比較的長いフレーム長を採用した．罰則項のチューニングパラメタ λ は対数等間隔で最適な値を探索した．評価尺度は SDR (Signal to Distortion Ratio) と SIR (Source to Interference Ratio) を用いた [14]．SDR は出力音の歪み，SIR は非目的信号の抑圧率を評価する尺度であり，値が大きいほど目的音強調の性能が良いことを示す．なお，SDR, SIR の算出に必要なリファレンスソースは話者ごとの録音データを利用した．評価する手法は，未処理の観測信号（Baseline），Table 1 のスパースネス制約による罰則付き教師なし伝達関数ゲイン基底 NMF (L1，CL0.75，CL0.5，CL0.25，GM)，そして比較手法として教師あり NMF (SNMF) [8] を用いた．SNMF は目的音強調区間と同じ時間長である学習用区間において話者ごとの録音より伝達関数ゲイン基底を学習した．

Table 1　Sparsity penalty functions used for estimation of source-activations in NMF.

Acronym	Type of the penalty	$g(\tilde{\mathbf{S}})$
L1	L_1 norm	$\sum_n \sum_k \tilde{S}_{kn}$
CL0.75	Column $L_{0.75}$ norm	$\sum_n (\sum_k \sqrt[4]{\tilde{S}_{kn}^3})^{\frac{4}{3}}$
CL0.5	Column $L_{0.5}$ norm	$\sum_n (\sum_k \sqrt{\tilde{S}_{kn}})^2$
CL0.25	Column $L_{0.25}$ norm	$\sum_n (\sum_k \sqrt[4]{\tilde{S}_{kn}})^4$
GM	Geometric Mean	$\sum_n (\prod_k \tilde{S}_{kn})^{\frac{1}{K}}$

Table 2　Recording environment.

マイクロホン	SHURE SM57
Power amp.	YAMAHA XM4080
AD/DA	Steinberg UR824

Table 3　Sampling frequencies for each microphone.

Mic 1	16,000 Hz
Mic 2	16,001 Hz
Mic 3	16,002 Hz
Mic 4	16,003 Hz

Table 4　Experimental conditions.

マイクロホン素子数	4
音源数	4
サンプリング周波数	16 kHz
フレーム長	4096 samples
シフト長	$\frac{1}{2}$ フレーム
目的音強調区間	10 sec
距離規準	I ダイバージェンス
基底行列初期値 α	振幅領域: $\alpha = 0.1$
	パワー領域: $\alpha = 0.01$
NMF 更新回数	200 回
チューニングパラメタ λ	$\lambda = 0, 10^{-3} \leq \lambda \leq 10^1$

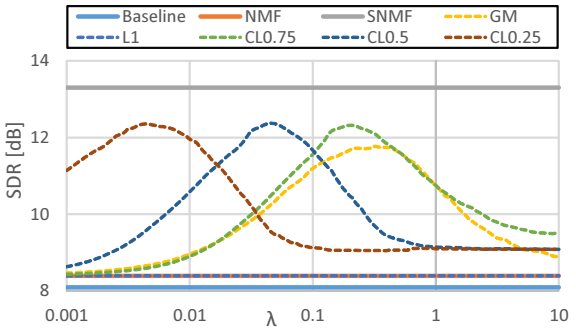

Fig. 2　SDRs depending on strength of each sparsity penalty in amplitude domain.

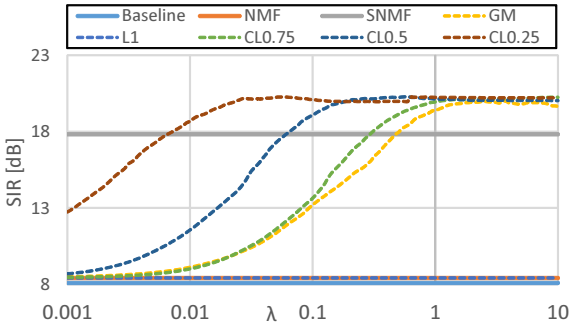

Fig. 3　SIRs depending on strength of each sparsity penalty in amplitude domain.

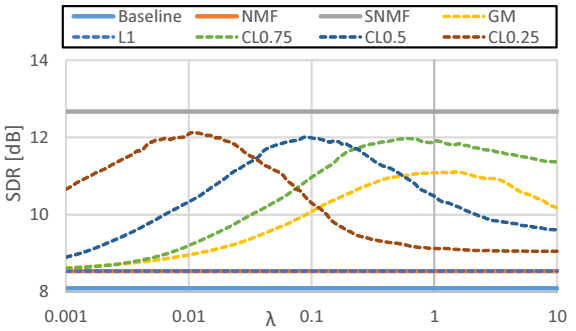

Fig. 4　SDRs depending on strength of each sparsity penalty in power domain.

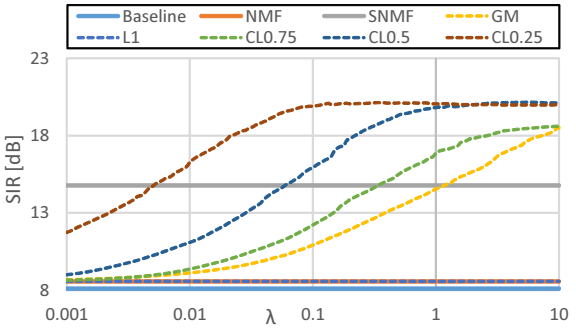

Fig. 5　SIRs depending on strength of each sparsity penalty in power domain.

4.2　評価結果

振幅領域での SDR，SIR による評価結果を，それぞれ Fig. 2, 3 に示す．また，パワー領域での SDR, SIR による評価結果を，それぞれ Fig. 4, 5 に示す．ここで，横軸はスパースネス制約の強さを制御するチューニングパラメタ λ の値を示しており，プロットした SDR, SIR は観測信号ごとに算出された評価値の平均値である．未処理の観測信号（Baseline）と比較すると，スパースネス制約を加えていない伝達関数ゲイン基底 NMF (NMF) と L_1 ノルム制約 (L1) では目的音強調効果は期待できない．一方，L1 以外のスパースネス制約では，チューニングパラメタ λ に依存するが SDR と SIR は Baseline から大きく向上しており，目的音強調効果が確認できた．特に，$p < 1$ である L_p ノルムによるスパースネス制約 (CL0.75, CL0.5, CL0.25) は教師あり (SNMF) に近い，優れた強調効果が得られた．なお，パワー領域と振幅領域での各強調手法の SDR，SIR を比較すると，振幅領域のほうが高い目的音強調効果が得られる傾向を確認した．従って，伝達関数ゲイン NMF を適用する領域としては，振幅領域が有効である可能性が示された．

5　まとめ

本稿では，非同期分散型マイクロホンアレーによる非同期録音に対して頑健な目的音強調手法である伝達関数ゲイン NMF において，5 種類のスパースネス制約の特性評価を行った．実収録した決定系の非同期音声録音を用いた評価の結果，振幅領域における $p < 1$ である L_p ノルム罰則付きの教師なし伝達関数ゲイン基底 NMF において，教師あり伝達関数ゲイン NMF に近い，優れた目的音強調効果を確認した．

謝辞　本研究は，科学研究費補助金基盤研究 (B)(25280069) の助成を受けたものです．また，本研究を進める上でご助言を頂きました日立製作所中央研究所の戸上真人氏に心から感謝申し上げます．

参考文献

[1] E. Robledo-Arnuncio, *et al.*, Proc. WASPAA, pp. 34-37, 2007.
[2] Z. Liu, Proc. IWAENC, 2008.
[3] S. Miyabe, *et al.*, Proc. ICASSP, pp. 674-678, 2013.
[4] R. Sakanashi, *et al.*, Proc. APSIPA, pp. 1-6, 2013.
[5] S. Araki *et al.*, Proc. ICASSP, vol. I, pp. 41-45, 2007.
[6] 加古 他, 音講論 (春), pp. 829-830, 2013.
[7] 戸上 他, 音講論 (春), pp. 803-804, 2010.
[8] 千葉 他, 音講論 (春), pp. 757-760, 2013.
[9] D. D. Lee *et al.*, in Neural Information Proc. Syst., vol.13, pp.556-562, 2001.
[10] R. Kompass, Neural Computation, vol. 19, no. 3, pp. 780-791, 2007.
[11] M. Nakano, *et al.*, Proc. IEEE MLSP, pp. 283-288, 2010.
[12] C. Joder, *et al.*, Proc. ICASSP, page 858-862, 2013.
[13] A. Cichocki, *et al.*, Proc. ICASSP, pp. 621-624, 2006.
[14] E. Vincent et al., IEEE Trans. ASLP, 14(6), 1462-1469, 2006.