伝達関数ゲイン基底NMFにおける マイク数・マイク配置と目的音強調性能の関係*

☆村瀬慶和, 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

近年、会議録音において、携帯電話やボイスレコーダーなどの非同期録音機器を使用したマイクロホンアレー信号処理への期待が高まりつつある。非同期録音機器を使用することができれば、マイクの増設や配置などの構成が柔軟にできるため、高いSN比での収音が期待できる。しかし、非同期録音機器を用いたアレー信号処理では、各機器間のサンプリング周波数のずれによって、ディジタル信号処理時に各機器間の位相差が時間とともに変化するため、従来のアレー信号処理では性能が劣化してしまう [1, 2]。そのため、従来のマイクロホンアレー信号処理を適用することを目的とした非同期録音機器に対する同期補正 [3, 4] が研究されているが、同期時の誤差にアレー信号処理の性能が左右される。そこで、同期誤差に頑健なマイクロホンアレー信号処理として位相情報を使用しない振幅ベースの目的音強調手法が提案されている。振幅ベースの目的音強調手法は現在までに非目的音と目的音のパワー比を最大化する振幅スペクトルビームフォーマ [5] や伝達関数の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) による時間周波数マスキング [6, 7] が提案されている。振幅スペクトルビームフォーマは各音源の単一音源区間での学習を行う必要があるのに対して、伝達関数ゲイン基底NMFはそのような学習が必ずしも必要ではなく、よりブラインドな目的音強調手法となり得るため、本研究では伝達関数ゲイン基底NMFによる目的音強調に焦点をあてる。

本稿では、伝達関数ゲイン基底NMFの目的音強調性能の向上を目指し、有効なマイクの数と配置について調査する。伝達関数ゲイン基底NMFによる目的音強調の性能は、観測信号の数やSN比に影響を受ける。このため、本研究ではマイクの個数や配置によって観測信号を変化させることで、伝達関数ゲイン基底NMFによる目的音強調の性能がどのように変化するのかを調査する。結果として、観測信号の数が多いほど目的音強調性能は向上したが、性能に限界が存在することを確認した。また、すべての観測信号においてSN比が高い観測信号が得られるマイク配置の場合に性能が向上することを確認した。

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 問題設定

いま、同期された録音機器を用いて M 個のマイクロホンで K 個の音源を録音する。このとき、時間周波数領域の観測信号は以下のように表される。

$$\begin{aligned}\mathbf{X}(\omega) &= [X_{mn}(\omega)]_{mn} \in \mathbb{C}^{M\times N} \\ &= \mathbf{A}(\omega)\mathbf{S}(\omega) \quad (1)\\ \mathbf{A}(\omega) &= [A_{mk}(\omega)]_{mk} \in \mathbb{C}^{M\times K} \quad (2)\\ \mathbf{S}(\omega) &= [S_{kn}(\omega)]_{kn} \in \mathbb{C}^{K\times N} \quad (3)\end{aligned}$$

ここで、$[B_{ij}]_{ij} \in \mathbb{C}^{I\times J}$ は i 行 j 列に B_{ij} を成分として持つ $I \times J$ の行列と表すこととし、ω は離散時間フーリエ変換の角周波数、N は時間フレーム数である。また、$X_{mn}(\omega)$ は m 番目のマイクで観測された n 番目の時間フレームの観測信号、$A_{mk}(\omega)$ は m 番目のマイクと k 番目の音源の間の伝達関数、$S_{kn}(\omega)$ は k 番目の音源の n 番目の時間周波数成分を表す。

また、非同期録音機器を使用する場合には柔軟な構成により $m = k$ 番目のマイクが k 番目の音源の最も近くに置かれ、A_{kk} $(k = 1, \ldots, K)$ は $A_{kj}, j = 1, \ldots, K$ の中で絶対値が最大になるとする。

さらに、非同期録音のサンプリング周波数のずれにより、観測信号 $\mathrm{X}(\omega)$ の位相は上のようなモデルで表される正確なものが得られないものとする。本稿では、このような条件のもとで各音源が最も高いSN比で観測されるマイクの音源強調を行い、振幅 $\bar{y}_{kn} = |A_{kk}S_{kn}|$ を求めることを目的とする。

以降の議論ではすべての処理を周波数ビンごとに行うため、周波数を表す記号 ω は省略する.

2.2 NMFによる目的音強調

本節では、振幅加法性混合モデルを採用し、Fig. 1に図示するNMFを用いて混合モデルのパラメタを推定することによる目的音強調手法について述べる。

時間周波数領域での観測信号の振幅の加法性を仮定することによって、位相情報を用いない混合モデルを以下のように表す。

$$\begin{aligned}\bar{\mathbf{X}} &= \left[\bar{X}_{mn}\right]_{mn} \in \mathbb{R}_+^{M\times N} \\ &\approx \bar{\mathbf{A}}\bar{\mathbf{S}} \quad (4)\\ \bar{\mathbf{A}} &= \left[\bar{A}_{mk}\right]_{mk} \in \mathbb{R}_+^{M\times K} \quad (5)\\ \bar{\mathbf{S}} &= \left[\bar{S}_{kn}\right]_{kn} \in \mathbb{R}_+^{K\times N} \quad (6)\end{aligned}$$

*On microphone arrangement for amplitude spectral array signal processing based on nonnegative matrix factorization by Yoshikazu MURASE, 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)

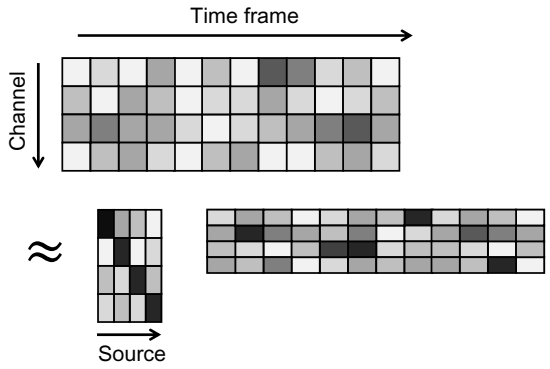

Fig. 1 Channel-time domain representation of observed signals for each frequency bin.

ここで、$[B_{ij}]_{ij} \in \mathbb{R}_+^{I\times J}$ は i 行 j 列に非負の実数 B_{ij} を成分として持つ $I\times J$ の行列を表すこととし、$\bar{X}_{mn}$、$\bar{A}_{mk}$、$\bar{S}_{kn}$ はそれぞれ振幅スペクトル領域における観測信号、伝達関数、音源の絶対値振幅を表す。そして、このモデルにおける $\bar{\mathbf{A}}$ と $\bar{\mathbf{S}}$ を NMF によって推定することで目的音強調を行う。

NMF は非負行列を 2 つの非負行列の積として

$$\bar{\mathbf{X}} \approx \tilde{\mathbf{X}} = \tilde{\mathbf{A}}\tilde{\mathbf{S}} \tag{7}$$

のように低ランク近似する手法である。NMF で最小化する行列の距離尺度には様々なものを用いることができ、問題にあわせて適切なものを選択する。このような低ランク近似は非負の制約により、解がスパースなものに限定され、適切な条件のもとでは基底 $\tilde{\mathbf{A}}$ が伝達関数ゲイン $\bar{\mathbf{A}}$ の同定となり、アクティベーション $\tilde{\mathbf{S}}$ が音源の絶対値振幅 $\bar{\mathbf{S}}$ の推定となるような行列分解が得られる。

本手法では、周波数領域 ICA などの従来の周波数独立の音源分離手法と同様にそれぞれの周波数ビンにおいて分離信号の各周波数成分が異なる順番で現れるというパーミュテーション問題が発生する。そこで、A_{kk} $(k = 1, \ldots, K)$ の絶対値が $A_{kj}, j = 1, \ldots, K$ の中で最大となる仮定から基底の初期値を

$$\bar{A}_{mk} = \begin{cases} 1 & (m = k) \\ \alpha & (m \neq k) \end{cases} \tag{8}$$

とすることでパーミュテーション問題の抑制を行う。ここで、α は非目的音に対する初期値であり、$\alpha < 1$ となる任意の正の実数である。

k 番目の音源を強調した強調信号 $\tilde{Y}_{kn}$ は SN 比が最も高い $m = k$ 番目のマイクで収音した観測信号 X_{kn} と k 番目の音源を強調するウィーナーマスクとの積によって、

$$\tilde{Y}_{kn} = X_{kn} \frac{\left(\tilde{A}_{kk}\tilde{S}_{kn}\right)^2}{\sum_{i=1}^{K}\left(\tilde{A}_{ki}\tilde{S}_{in}\right)^2} \tag{9}$$

と表される。ここで、ウィーナーフィルタは観測信号の音源の重ね合わせによるモデル誤差を緩和するために使用する。

本手法では以上のような操作によって目的音強調を行うが、音源数に対してマイク数が同程度である場合は強調効果が低いことが確認されている。その解決法として、アクティベーション行列をスパースにするための罰則項を導入した音源のチャネル間振幅差を基底ベクトルとする音源分離 [6] や、基底行列を各音源の単一音源区間によって学習した伝達関数ゲイン基底 NMF[7] などの手法が提案されている。

3 伝達関数ゲイン基底 NMF による目的音強調可能な条件について

NMF は非負の制約によってスパースな解に誘導するため、目的音強調のための有用な伝達関数ゲインに近づけることができる。しかし、非負の制約だけでは目的音強調に不十分であり、教師なし学習で目的音強調を行うためには付加的な条件が必要となる。そこで本節では、定性的ではあるが目的音強調を行うための付加的な条件について、マイク数と配置の観点から議論する。ただしアクティベーションをスパースにするような正則化項の導入は本稿では議論の対象外とし、正則化項のない単純な NMF の学習に焦点を当てて議論する。

3.1 マイク数の影響

目的音強調可能な条件をマイク数 M と音源数 K の大小関係に分けて議論する。

$M = K$ の場合は、NMF による基底 $\tilde{\mathbf{A}}$ とアクティベーション $\tilde{\mathbf{S}}$ の推定において以下のような解が存在する。

$$\tilde{\mathbf{A}} = \mathbf{E},\ \tilde{\mathbf{S}} = \bar{\mathbf{X}} \tag{10}$$

ここで、$\mathbf{E} \in \mathbb{R}_+^{M\times M}$ は単位行列を表す。このような解は一例にすぎず、$\tilde{\mathbf{X}} = \tilde{\mathbf{A}}\tilde{\mathbf{S}} = \bar{\mathbf{X}}$ を満たして NMF の低ランク近似の誤差を厳密にゼロとする解は無数に存在する。そして、このような解は一般に伝達関数ゲインと音源の絶対値振幅の推定とはかけ離れたものとなるため、目的音強調ができないと考えられる。

$M < K$ の場合は、$M = K$ と同様に $\tilde{\mathbf{X}} = \tilde{\mathbf{A}}\tilde{\mathbf{S}} = \bar{\mathbf{X}}$ となる解が無数に存在するため目的音強調できないと考えられる。

$M > K$ の場合は、マイク数の増加にともなってアクティベーションの任意性が制限されるため、無意味な解から離れて性能が単調的に改善すると考えられる。

3.2 マイク配置の影響

目的音強調可能な条件をマイクの配置によって変化する SN 比をもとに議論する。観測信号のスパース性が十分に保たれない場合には、NMF の最適解が任意性を持ってしまい、目的音強調に有用な基底とアクティベーションが得られないということが、多重音解析の分野で広く議論されている [9]。強調する音源の SN 比が低い配置の場合には、観測信号の十分なスパース性が確保されず、目的音強調の性能が低下してしまう。一方で、SN 比が高い配置の場合には観測信

Table 1　Experimental conditions.

Number of sources	4
Sampling frequency	16 kHz
Frame length	4096 samples
Frame shift	2048 samples
Signal length for evaluation	10 sec
Signal length for supervised NMF training	10 sec
Divergence	I-divergence
α (initialization parameter)	0.25
Number of NMF iterations	200
Reverberation time	0.3 sec

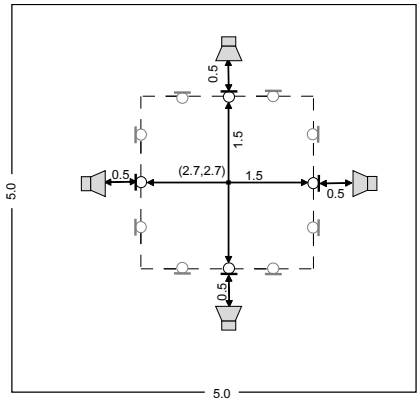

Fig. 2　Arrangement of speakers and microphones.

号がよりスパースになるため解の任意性が減り、所望の伝達関数ゲインと音源の絶対値振幅に近づく解を求めることができる。このため、SN 比が高くなるような配置であれば目的音強調が可能であると考えられる。このようなマイク配置を変えた場合の SN 比の変化は教師なし学習、教師あり学習のどちらにも影響を与える。しかし、教師あり学習では事前に基底が学習されるため、任意性が性能差に与える影響は教師なし学習ほど生じないと考えられる。

4　マイク数、マイク配置を変化させた場合の目的音強調性能評価

本実験では、3 章で議論したマイク数、マイク配置を変化させた場合の目的音強調性能に対する実際の影響について実験的に調査する。

4.1　マイク数：評価実験方法

実験条件を Table 1 に示す。評価に用いる信号は、鏡像法によって生成したインパルス応答を音源信号に畳み込み生成した [10]。また、すべてのマイクは無指向性マイクとしてシミュレーションを行った。マイク配置は、Fig. 2 のような空間で破線上にマイクが均等に並ぶような 6 パターンの配置で実験を行った。それぞれのパターンは、一辺の破線上にそれぞれ 1, 3, 5, 9, 17, 33 個のマイクが等間隔に並ぶ。ただし、各音源を強調をするための観測信号を収音するマイクは各音源の前に固定される。また、ある 1 つの音源が支配的な観測信号を得るために、各辺の両端にはマイクを設置しない。そして、計 4, 12, 20, 36, 68, 132 個のマイクによって得られた観測信号を用いて、目的音強調を行った。

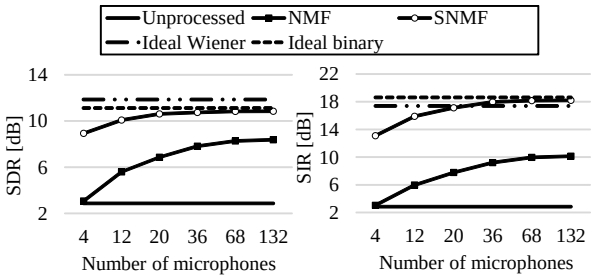

Fig. 3　SDRs and SIRs of increasing microphones with NMF and SNMF.

時間領域の強調信号は、ウィーナーマスクによって求めた振幅スペクトルに観測信号の位相を与えて離散時間フーリエ逆変換することによって求めた。

評価尺度は、Signal-to-distortion ratio (SDR) と Source-to-interference ratio (SIR) を用いた [11]。これらは正答となる元の音源信号と強調信号によって算出される。SDR は強調信号の歪み、SIR は非目的信号の抑圧率を表し、ともに値が大きいほど目的音強調性能が良いことを示す。そして、未処理の観測信号（Unprocessed）、教師なし学習による伝達関数ゲイン基底 NMF（NMF）、教師あり学習による伝達関数ゲイン基底 NMF（SNMF）、各音源のパワー比から作成した理想ウィーナーマスク（I-Wiener）、理想バイナリマスク（I-binary）の 5 つの手法に対して評価を行った。

4.2　マイク数：評価結果

各マイク数における SDR と SIR の評価結果を Fig. 3 に示す。NMF では、マイクが 4 個の場合に観測信号の評価値と同等であり、音源とマイクの数が同程度である場合には目的音強調効果は期待できない事が確認できた。これは 3.2 節で議論したように無意味な解に収束しているためだと考えられる。一方、SNMF では、マイク数が 20 個の場合に I-binary の強調性能近くまで達していることを確認した。この結果から、SNMF ではマイクを 20 個程度使用することで、限界性能に達していることがわかる。また、SNMF、NMF ともにマイク数に対して性能が単調増加するが、マイク数の増加に対する性能向上の飽和がみられた。以上より、両者の性能差はマイク数が増加しても変わらず、マイク数を無限に増やしても教師なし学習は教師あり学習に追いつくことはないと予想される。

4.3　マイク配置：実験方法

Fig. 4 のような 6 つのパターンでマイクを配置し、目的音強調性能の比較を行った。全てのパターンにおいて、マイク数は全て 20 個で統一し、各音源の前に目的音強調処理を行う観測信号を録音するマイクが共通の位置に 1 つ固定されている。そして、それ以外の 16 個のマイク配置を変えることで各パターンを作成した。パターン 0 は 4.1 節の実験におけるマイク数

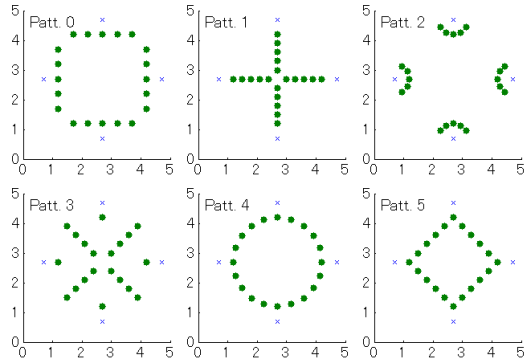

Fig. 4　Microphone arrangement of each pattern.

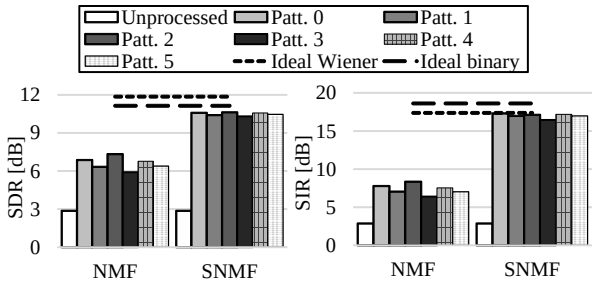

Fig. 5　SDRs and SIRs on each pattern with NMF and SNMF.

が 20 個の場合と同じ配置、パターン 1 は 2 つの音源間にマイクが並ぶ配置、パターン 2 は各音源から等しい距離に扇状に並ぶ配置、パターン 3 は対角線上に並ぶ配置、パターン 4, 5 は中心点から円形と菱形になるようにマイクを配置している。また、パターン 1, 3 は 2 つの音源でマイクまでの距離が等しく、得られる観測信号の SN 比が低い配置で、パターン 2 はすべてのマイクで得られる観測信号の SN 比が高い配置、パターン 0, 4, 5 は観測信号の SN 比が徐々に変化するような配置となっている。評価は Unprocessed、NMF、SNMF、I-Wiener、I-binary の 5 つの手法に対して行った。また、マイクの個数と配置以外の実験条件は 3 章と同様である。

4.4　マイク配置：評価結果

各パターンによって得られた目的音強調性能を Fig. 5 に示す。NMF において、SDR、SIR ともにパターン 2 の性能が最も高く、パターン 4, パターン 1, パターン 3 の順に性能が低下している。また、NMF ほどの性能差は見られないが、SNMF においても同様の順で性能が高いことを確認した。この結果より、パターン 1, パターン 3 のように SN 比が低い配置よりも、パターン 2 のように強調する観測信号以外のマイクでも SN 比が高くなるようなマイク配置では、伝達関数ゲイン基底 NMF による目的音強調性能が向上することを確認した。

5　まとめ

本稿では、非同分散型マイクロホンアレーの使用を想定した伝達関数ゲイン基底 NMF による目的音強調において、マイクの個数や配置によって強調性能がどのように変化するかの調査を行った。まず、伝達関数ゲイン基底 NMF のモデル化と学習の原理について述べ、望ましい観測の条件について議論した。そして、鏡像法を用いたシミュレーション実験により、マイク数による強調性能では、教師あり学習、教師なし学習のどちらの伝達関数ゲイン基底 NMF においても、マイク数に対して性能が単調増加するが、限界があることを確認した。また、マイク配置による強調性能は各音源に対して SN 比が高くなるような配置の場合に向上することを確認した。

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ある。

参考文献

[1] Robledo *et al.*, *Proc. WASPAA*, pp. 34-37, 2007.

[2] Liu *et al.*, *Proc. IWAENC*, 2008.

[3] Miyabe *et al.*, *Proc. ICASSP*, pp. 674-678, 2013.

[4] Sakanashi *et al.*, *Proc. APSIPA*, pp. 1-6, 2013.

[5] Kako *et al.*, *Proc. Acoustic Society of Japan*, pp. 829-830, Mar. 2013.

[6] Togami *et al.*, *Proc. Acoustical Society of Japan*, pp. 803-804, 2010.

[7] Chiba *et al.*, *Proc. Acoustical Society of Japan*, pp.757-760, Mar. 2014.

[8] Lee and Seung, *Proc. NIPS*, pp. 556-562, 2000.

[9] Rigaud et al., *Proc. WASPAA*, 2013.

[10] Habets, Available: `http://home.tiscali.nl/ehabets/rir_generator.html`

[11] Vincent *et al.*, *Proc. IEEE Trans. on Audio, Speech & Language Processing*, vol.14, no. 4, pp. 1462-1469, 2006.