日本音響学会誌 72 巻 8 号（2016），pp. 462–470

論　　文

43.60.+d

アドホックマイクロホンアレーにおける時間チャネル領域での非負値行列因子分解を用いた振幅ベースの音声強調*

千葉大将*1　小野順貴*2,*3　宮部滋樹*1
髙橋　祐*4　山田武志*1　牧野昭二*1

［要旨］ 本論文では，時間チャネル領域の非負値行列因子分解（NMF）による，非同期分散型録音の目的音強調手法について述べる。複数の録音機器による多チャネル信号は，機器ごとのサンプリング周波数の微小なずれが引き起こす位相差のドリフトのため，位相情報を用いるアレー信号処理は適さない。位相に比べると振幅の分析はドリフトの影響を大きく受けないことに着目し，戸上らが提案した時間チャネル領域の NMF によるチャネル間ゲイン差の分析（伝達関数ゲイン基底 NMF）に基づく時間周波数マスクを用いる。また，基底数よりも十分大きなチャネル数が得られない条件の音声強調のための，基底を事前に学習する教師あり NMF について議論する。

キーワード　音声強調，アドホックマイクロホンアレー，サンプリング周波数ミスマッチ，非負値行列因子分解，時間周波数マスキング
Speech enhancement, Ad-hoc microphone array, Sampling frequency mismatch, Non-negative matrix factorization, Time-frequency masking

1. はじめに

マイクロホンアレーは，複数のマイクロホンでの観測により空間的な情報を取得し，単一のマイクロホンでは困難な，音源定位，音源分離，目的音強調などの処理を行うフレームワークであり，音を使用する様々なアプリケーションへの応用が期待される。しかし，マイクロホンアレーでは空間的な情報としてマイク間での微小な時間差などを手がかりとしているため厳密な同期録音が必要であり，また，性能を向上させるにはマイク数を増やしたりマイクを広範囲に配置することが求められる。このように録音機器に大きなコストが生じることが実用における問題点とされてきた。

このような中，携帯電話やボイスレコーダなどの非同期録音機器を分散配置して一つのマイクロホンアレーとして使用する非同期分散型マイクロホンアレーへの関心が高まりつつある。非同期分散型マイクロホンアレーでは，身の回りの非同期録音機器を用いるため，低コストでマイクロホンアレーを実現することが期待される。また，マイク数や配置等の構成の自由度が高く，目的話者の近くにマイクを配置できるため優れた SN 比での収音が期待できる。

しかし，非同期機器間での録音開始時刻の差やサンプリング周波数の微細なずれ（サンプリング周波数ミスマッチ）がアレー信号処理時に問題となる [1, 2]。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共に変化させるため，従来の位相情報に依存しているアレー信号処理の性能は劣化してしまう [3, 4]。そのため，非同期録音に対する同期補正 [5, 6] が研究されているが，計算量が多く，また，同期誤差にアレー信号処理の性能が左右される。一方，非同期録音に直接適用できる効率的な音声強調手法として，SN 比最大化ビームフォーマ [7] を振幅情報のみで行う振幅スペクトルビームフォーマ [8] のような位相情報に依存しない振幅ベースの手法が提案されている。

本論文は，後者のような非同期録音に対して直接適用可能な，振幅ベースの目的音強調の検討を目的とする。具体的には，チャネル間の振幅差（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）[9, 10] が非同期録音に対して有効な目的音強調手法であることを検証する。

この伝達関数ゲイン基底 NMF は，伝達関数ゲイ

---

* Amplitude-based speech enhancement with non-negative matrix factorization in time-channel domain for ad hoc microphone array,
by Hironobu Chiba, Nobutaka Ono, Shigeki Miyabe, Yu Takahashi, Takeshi Yamada and Shoji Makino.

*1 筑波大学システム情報工学研究科
*2 国立情報学研究所
*3 総合研究大学院大学
*4 ヤマハ株式会社
（問合先：牧野昭二　e-mail: maki@tara.tsukuba.ac.jp）

（2015 年 10 月 5 日受付，2016 年 2 月 26 日採録決定）

ンの比がマイクごとに明確に異なるという非同期分散型マイクロホンアレーの特性を有効に利用できることから，観測信号間で位相ずれが起こる非同期録音において頑健で効率性の高い強調手法であることが期待される。しかし，時間周波数領域での NMF と異なり，時間チャネル領域ではマイク数は基底数と同程度であり，このような条件では一般的な距離最小化規準での NMF による分離は困難である。そこで，本論文では時間チャネル領域における NMF において解決策の考察を行い，非同期録音におけるブラインドでの強調手法として適用可能か調査を行う。

評価では，戸上らの罰則付き伝達関数ゲイン基底 NMF [9] と，著者らの先行研究である，事前に単一音源区間において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 [10] による目的音強調を行う。そして，人工的にサンプリング周波数ミスマッチを生成した非同期録音及び実際の非同期機器を用いた音声録音を用いて，伝達関数ゲイン基底 NMF による時間周波数マスキングの音声強調性能が，サンプリング周波数ミスマッチの有無に依存していないことを確認する，また，罰則付き伝達関数ゲイン基底 NMF 及び教師あり伝達関数ゲイン基底 NMF と，従来の位相情報を用いる音声強調手法や振幅スペクトルビームフォーマの強調性能を比較する。

本論文の構成を以下に示す。第 2 章では，非同期分散型録音における振幅ベースの混合モデルについて概説する。第 3 章では，伝達関数ゲイン基底 NMF を用いた時間周波数マスキングについて述べる。第 4 章では，伝達関数ゲイン基底 NMF による目的音強調に必要な制約について述べる。第 5 章では，同期録音・非同期録音における目的音強調性能の評価実験について説明し，実験結果の提示と考察を行い，第 6 章では，非同期機器による実録音における強調性能の評価実験について説明し，実験結果の提示と考察を行う。第 7 章では，本論文の結論を述べる。

2. 非同期分散型録音における振幅ベースの混合モデル

2.1 従来の同期録音における線形混合モデル

本論文では時間周波数領域での信号を扱う。また，i 行 j 列に複素数 X_{ij} を成分として持つ $I \times J$ の行列 $\boldsymbol{X}$ を $\boldsymbol{X} = [X_{ij}]_{ij} \in \mathbb{C}^{I \times J}$ と表すこととする。

Fig. 1 のように，同期されたマイクロホンアレーによるマルチチャネルの観測は以下のように表される。

$$\boldsymbol{X}(\omega) = \boldsymbol{A}(\omega)\boldsymbol{S}(\omega) \tag{1}$$

$$\boldsymbol{X}(\omega) = [X_{mn}(\omega)]_{mn} \in \mathbb{C}^{M \times N} \tag{2}$$

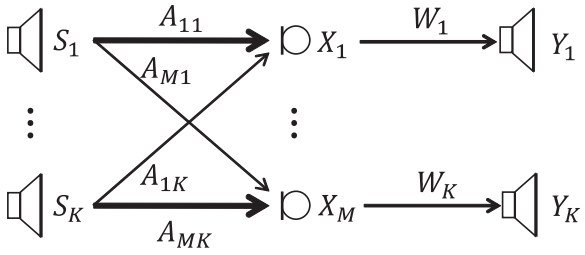

Fig. 1　Mixture model in STFT domain.

$$\boldsymbol{A}(\omega) = [A_{mk}(\omega)]_{mk} \in \mathbb{C}^{M \times K} \tag{3}$$

$$\boldsymbol{S}(\omega) = [S_{kn}(\omega)]_{kn} \in \mathbb{C}^{K \times N} \tag{4}$$

ここで，周波数ビン番号を ω，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を $X_{mn}(\omega)$，k 番目の音源から m 番目のマイクまでの伝達関数を $A_{mk}(\omega)$，n 番目の時間フレームにおける k 番目の音源信号を $S_{kn}(\omega)$ 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。式 (1) で表される従来の線形混合モデルでは，音源及びマイクは静止していると仮定しているため，$\boldsymbol{A}(\omega)$ は線形時不変の混合行列である。そのため，マイクロホンアレー信号処理においては，厳密な同期録音が必要不可欠である。これは，マイクロホンアレー信号処理では，各マイクで録音される信号間の微小な時間差（例えば，経路長 3.4 cm に対して 100 μs）が音源の空間情報の主要な手がかりとなっているためである。よって従来は，各チャネルを正確に同期させるために，すべてのチャネルは多チャネル A/D 変換機に接続され，同一クロックによりサンプリングされる必要があり，これがマイクロホンアレーの多素子化や分散配置などに対して大きなコストを生じる主要因の一つとなっていた [1, 2]。

非同期分散型マイクロホンアレーによる観測信号では，デバイス間のサンプリング周波数ミスマッチによって位相のドリフトが起こり，混合行列 $\boldsymbol{A}(\omega)$ が時変となるため式 (1) のモデルは不適当である。従って，非同期録音において音声強調を行うには非同期録音において成立する混合モデルを考える必要がある。

2.2 振幅領域での混合モデル

複数の録音機器を用いて音響信号を録音する場合，録音機器の公称サンプリング周波数が同じ場合でも，実際のサンプリング周波数は一般に，水晶振動子の個体差や電源電圧の影響などにより，機器ごとにわずかに異なっている（サンプリング周波数ミスマッチ）。

同期録音と比較して非同期録音が持つ性質としては，特に録音開始時刻のずれと，非同期機器のサンプリング周波数ミスマッチがある [1, 2]。録音開始時刻のずれに関しては相互相関関数によって時間シフトすること

により，録音開始時刻の差を十分小さくできる。

いま，録音機器 1, 録音機器 2 のサンプリング周波数を f_1，f_2 で表すと，$\epsilon = f_2/f_1 - 1$ が，録音機器間の相対的なサンプリング周波数ミスマッチを表す無次元量となる。ϵ は，10 ppm（ppm は parts per milion で 10^6 を表す）の数倍程度に収まることが多い。

こうしたわずかなミスマッチがアレイ信号処理に与える影響は大きい。例えば，録音機器 1，2 のサンプリング周波数を 16,000 Hz，16,001 Hz とし，これらを 40 cm 離した配置で，正面方向の音源信号を 10 秒間録音を行う。このときサンプリング周波数ミスマッチは，62.5 ppm である。10 秒間の音響信号は，録音機器 1 では 160,000 サンプル，録音機器 2 では 160,010 サンプルに相当する。すなわち，正面方向から到来する音波は録音機器 1，2 に同時に届くが，サンプリング周波数ミスマッチにより，10 秒間の録音信号の最後では 10 サンプルずれることになる。一方，30 度方向から到来する音波に対する到来時間差は，$0.4 \times \sin 30^\circ / 340 \times 16{,}000 \simeq 9.4$ で，約 9.4 サンプルに相当する。つまり，62.5 ppm というサンプリング周波数のずれが引き起こす時間差は，10 秒間に音源が正面から 30 度方向に移動したのと区別がつかないことになる。当然のことながら，このままでは，チャネル間の時間差からは音源位置情報が得られない。また，多くの線形アレイ信号処理においては，音源からマイクロホンまでの伝達関数は線形時不変であることが仮定されているため，こうしたチャネル間の時間差のドリフトは，音源分離などにも深刻な破綻を引き起こす。

サンプリング周波数ミスマッチ ϵ により，片方の信号には，ある時間長 T に対して $T\epsilon$ の時間差ドリフトが生じる。ここで，本論文で扱う録音信号では，時間差ドリフト $T\epsilon$ と短時間フーリエ変換（STFT: Short-time Fourier transform）フレーム長 L において，

$$T\epsilon \ll L \tag{5}$$

という関係が成り立つことを仮定する。すなわち，無次元量 $T\epsilon/L$ が 1 よりも十分に小さい値であれば，STFT フレーム長 L に対してサンプリング周波数ミスマッチ ϵ が十分に小さいと考えられる。そこで，以上の非同期録音の性質から，非同期録音でも成立する，位相情報に依存しない混合モデルを考える。

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が，この位相ずれのサンプル数が STFT フレーム長より十分に小さい場合，$\boldsymbol{A}$ の振幅（伝達関数ゲイン）は時不変であると仮定できる。従って，観測信号の振幅スペクトルを音源信号のそれの積和として近似する以下の線形混合モデルが成立する。

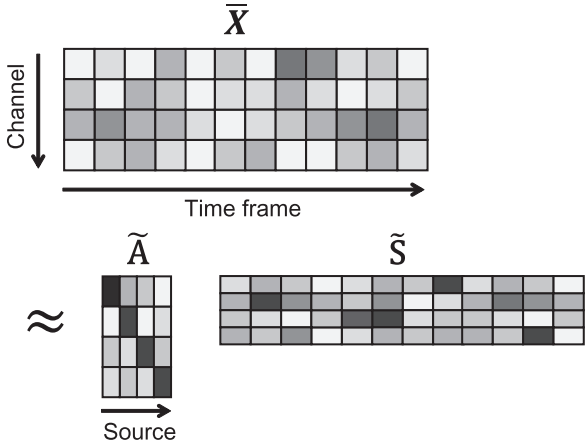

Fig. 2 Channel-time domain representation of observed signals for each frequency bin.

$$\bar{\boldsymbol{X}}(\omega) \approx \bar{\boldsymbol{A}}(\omega)\bar{\boldsymbol{S}}(\omega) \tag{6}$$

$$\bar{\boldsymbol{X}}(\omega) = \left[\bar{X}_{mn}(\omega)\right]_{mn} \tag{7}$$

$$= \left[|X_{mn}(\omega)|\right]_{mn} \in \mathbb{R}_+^{M\times N} \tag{8}$$

$$\bar{\boldsymbol{A}}(\omega) = \left[|A_{mk}(\omega)|\right]_{mk} \in \mathbb{R}_+^{M\times K} \tag{9}$$

$$\bar{\boldsymbol{S}}(\omega) = \left[|S_{kn}(\omega)|\right]_{kn} \in \mathbb{R}_+^{K\times N} \tag{10}$$

ここで，$\bar{X}_{mn}(\omega)$，$\bar{A}_{mk}(\omega)$，$\bar{S}_{kn}(\omega)$ はそれぞれ $X_{mn}(\omega)$，$A_{mk}(\omega)$，$S_{kn}(\omega)$ の振幅を表し，$\mathbb{R}_+$ は非負の実数の集合を表す。また，$|\cdot|$ は絶対値を返す演算子である。このようなパワー領域もしくは振幅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 [11]。そこで，本論文では **Fig. 2** で表されるように，時間チャネル領域における NMF（伝達関数ゲイン基底 NMF）[9] によって観測信号の振幅スペクトル $\bar{\boldsymbol{X}}(\omega)$ から伝達関数ゲイン $\bar{\boldsymbol{A}}(\omega)$ と音源アクティベーション $\bar{\boldsymbol{S}}(\omega)$ を推定する。

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。また，NMF により推定された伝達関数ゲインを音源アクティベーションをそれぞれ $\tilde{\boldsymbol{A}}$，$\tilde{\boldsymbol{S}}$ と表す。

3. 伝達関数ゲイン基底 NMF を用いた時間周波数マスキング

3.1 時間チャネル領域における罰則付き伝達関数ゲイン基底 NMF

本論文では，β ダイバージェンス規準 NMF [12, 13] を採用する。すなわち，以下の目的関数，

$$\mathcal{J}(\bar{\boldsymbol{X}}, \tilde{\boldsymbol{A}}\tilde{\boldsymbol{S}}) = \mathcal{D}_\beta(\bar{\boldsymbol{X}}|\tilde{\boldsymbol{A}}\tilde{\boldsymbol{S}}) \tag{11}$$

に対して補助関数法によって導出される乗法型更新式，

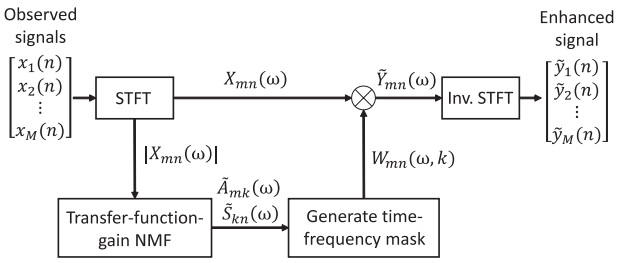

Fig. 3 Block diagram of amplitude-based speech enhancement with NMF in channel-time domain.

$$\tilde{A}_{mk} \leftarrow \tilde{A}_{mk} \left(\frac{\sum_n |X_{mn}| \tilde{X}_{mn}^{\beta-2} \tilde{S}_{kn}}{\sum_n \tilde{X}_{mn}^{\beta-1} \tilde{S}_{kn}} \right)^{\psi(\beta)} \quad (12)$$

$$\tilde{S}_{kn} \leftarrow \tilde{S}_{kn} \left(\frac{\sum_m |X_{mn}| \tilde{X}_{mn}^{\beta-2} \tilde{A}_{mk}}{\sum_m \tilde{X}_{mn}^{\beta-1} \tilde{A}_{mk}} \right)^{\psi(\beta)} \quad (13)$$

を更新することで，伝達関数ゲインと音源アクティベーションのそれぞれの推定値，$\tilde{\boldsymbol{A}} = [\tilde{A}_{mk}]_{mk} \in \mathbb{R}_+^{M \times K}$，$\tilde{\boldsymbol{S}} = [\tilde{S}_{kn}]_{kn} \in \mathbb{R}_+^{K \times N}$ を得る。ここで，$\tilde{X}_{mn} = \sum_k \tilde{A}_{mk} \tilde{S}_{kn}$ は更新ごとに計算される観測信号の推定振幅スペクトルを表す。また，

$$\mathcal{D}_\beta(y|x) = \frac{1}{\beta(\beta-1)} (y^\beta + (\beta-1)x^\beta - \beta y x^{\beta-1}) \quad (14)$$

$$\psi(\beta) = \begin{cases} \dfrac{1}{2-\beta} & \beta < 1 \\ 1 & 1 \le \beta \le 2 \\ \dfrac{1}{\beta-1} & \beta > 2 \end{cases} \quad (15)$$

であり，$\beta = 1$ では I ダイバージェンス規準での更新式となる。なお，式 (12)，(13) の更新ごとに，

$$\tilde{A}_{mk} \leftarrow \frac{\tilde{A}_{mk}}{\sum_i \tilde{A}_{mi}}, \quad (16)$$

$$\tilde{S}_{kn} \leftarrow \left(\sum_i \tilde{A}_{mi} \right) \tilde{S}_{kn} \quad (17)$$

として基底を正規化する。以上の操作により伝達関数ゲインと音源アクティベーションを推定することで得られる，目的音と非目的音の推定振幅スペクトルを用いて，**Fig. 3** で表されるように時間周波数マスクを生成することで目的音強調を行う。なお，基底が複数の場合，周波数ビンごとに基底の順番が入れ替わるパーミュテーション問題が発生する。そのため，パワーの相関や基底ベクトルの類似度によるパーミュテーション解決手法や基底の共通化が提案されている [9]。

3.2 ウィーナフィルタによる時間周波数領域マスキング

式 (6) の混合モデルでは振幅の重ね合わせを仮定しているが，時間チャネル領域においてはモデル誤差が大きい。従って，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 $\tilde{\boldsymbol{A}}$ と音源アクティベーション $\tilde{\boldsymbol{S}}$ を用いた時間周波数領域でのウィーナマスクによる強調を行う。本論文では，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マスクをかけ，各音源の強調信号 $\tilde{\boldsymbol{Y}} = [\tilde{Y}_{kn}]_{kn} \in \mathbb{C}^{K \times N}$ を得る。具体的には，振幅領域で適用した場合は，

$$\tilde{Y}_{kn} = \frac{(\tilde{A}_{kk} \tilde{S}_{kn})^2}{\sum_i (\tilde{A}_{ki} \tilde{S}_{in})^2} X_{kn} \quad (18)$$

として強調信号を得る。

4. 伝達関数ゲイン基底 NMF による目的音強調に必要な制約

4.1 時間チャネル領域での NMF による伝達関数ゲイン推定の問題点と解決策

時間チャネル領域における NMF では，チャネル数と基底（音源）数との差が小さい条件下においては距離最小化規準で加法性の構成成分への分解を行う NMF は低ランク近似としての拘束力が弱く，音源を分離しない解が最適解となってしまう。従って，チャネル数が音源数を十分に上回っていない条件下において，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。

その制限の一つとして，戸上らの時間フレームごとの音源アクティベーションにおける振幅の重ね合わせに対する罰則の導入がある。戸上らは音源アクティベーションに対して $L_{0.5}$ ノルムによるスパースネス制約を導入した罰則付き NMF により，強調効果を優決定系のブラインドな観測データで確認している。このスパースネス制約では，振幅の重ね合わせに対してペナルティを与え，低ランク近似としての拘束力を高めることで音源が分離される解を得ることができる。

また，各音源が分離されたアクティベーションが推定できる制限として，事前に伝達関数ゲイン基底を学習する教師あり NMF が考えられる [14]。具体的には，ある音源のみアクティブな区間（単一音源区間）で基底数 1 として伝達関数ゲイン基底 NMF を行う。この場合，他の音源との振幅の重ね合わせが起こらず，振幅の加法性を仮定した混合モデルと実際の観測の誤差が小さい。また，非同期分散型マイクロホンアレーでは，音源が多少動いても伝達関数ゲインは時不変と見なすことができるため，少ないマイク数でも伝達関数ゲインを高い精度で推定できることが期待される。

以下の節では，時間チャネル領域における罰則付き NMF と教師あり NMF について説明する。

4.2 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 [12, 13] において，時間フレームごとの音源アクティベーション $\tilde{\boldsymbol{S}}$ に対するスパースネス制約を評価する関数 $g(\tilde{\boldsymbol{S}})$ 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。

$$\mathcal{J}(\bar{\boldsymbol{X}}, \tilde{\boldsymbol{A}}\tilde{\boldsymbol{S}}, \lambda) = \mathcal{D}_\beta(\bar{\boldsymbol{X}}|\tilde{\boldsymbol{A}}\tilde{\boldsymbol{S}}) + \lambda g(\tilde{\boldsymbol{S}}) \quad (19)$$

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，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べきである。本論文では，スパースネス制約として $L_{0.5}$ ノルム を使用し，距離尺度として $\beta = 1$，すなわち I ダイバージェンスを使用する。従って，式 (12) と以下の罰則付き乗法型更新式 [15] を用いて局所解を求める。

$$\tilde{S}_{kn} \leftarrow \tilde{S}_{kn} \frac{\Sigma_m \frac{\bar{X}_{mn}\tilde{A}_{mk}}{\tilde{X}_{mn}}}{\Sigma_m \tilde{A}_{mk} + \lambda \frac{\Sigma_k \sqrt{\tilde{S}_{kn}}}{\sqrt{\tilde{S}_{kn}}}} \quad (20)$$

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スペクトルを表す。

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パーミュテーション問題が起こる。そこで，本論文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 $\tilde{\boldsymbol{A}}$ の初期値を，

$$\tilde{A}_{mk} = \begin{cases} 1 & (m = k) \\ \alpha & (m \neq k) \end{cases} \quad (21)$$

として与える。ここで，パラメータ 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。更に，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。

4.3 教師あり伝達関数ゲイン基底 NMF

伝達関数ゲイン基底の学習は，まず，各音源ごとに伝達関数ゲイン基底ベクトルを学習する。ここでは単一音源区間は人手で与えられることを仮定するが，各マイクレベルの差などを利用して自動検出することも可能と考えられる。音源 k のみの単一音源区間において，式 (12)，(13) の更新式による伝達関数ゲイン基底 NMF を行い，音源 k におけるランク 1 の伝達関数ゲイン基底行列 $\bar{\boldsymbol{a}}_k = \left[\bar{A}_{mk}\right]_{m1} \in \mathbb{R}_+^{M\times 1}$ を得る。そして，音源 k ごとに得られた伝達関数ゲイン基底行列を結合することで，伝達関数ゲイン基底行列 $\bar{\boldsymbol{A}} = (\bar{\boldsymbol{a}}_1 \cdots \bar{\boldsymbol{a}}_K)$ を学習する。目的音強調区間では，伝達関数ゲイン基底行列を事前学習した基底に固定し，音源アクティベーション行列のみ式 (13) で更新する。なお，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 $\bar{\boldsymbol{A}}$ の学習における初期値を，

$$\bar{A}_{mk} = \begin{cases} 1 & (m = k) \\ \alpha & (m \neq k) \end{cases} \quad (22)$$

として与える。ここで，パラメータ 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。

5. 同期録音・非同期録音における目的音強調性能の評価

5.1 実験条件

非同期分散型マイクロホンアレーによる会議録音を想定した，同期録音と人工的に生成した非同期録音データによって伝達関数ゲイン基底 NMF の目的音強調性能を評価した。録音データは，**Fig. 4** のようなマイク・話者配置とし，同期型マイクロホンアレーにより話者ごとに **Table 1** のような環境で収録した。録音後，マイクごとに **Table 2** 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 3** 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ため，本論文では比較的長いフレーム長を採用する。評価尺度は SDR（Source to Distortion Ratio）と SIR（Source to Interference Ratio）を用いた [16]。SDR は出力音のひ

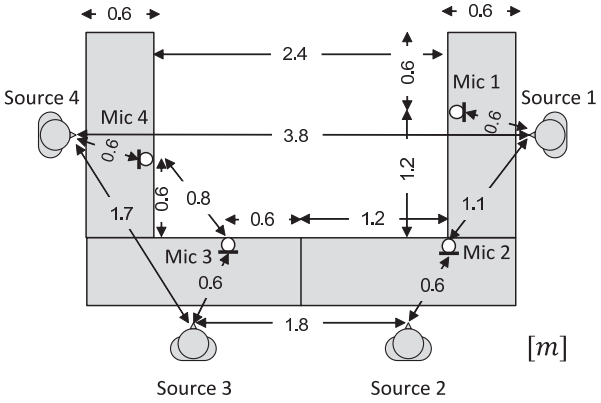

Fig. 4 Arrangement of microphones used in evaluation.

Table 1　Recording environment.

Microphone	SHURE SM57
Power amp.	YAMAHA XM4080
AD/DA	Steinberg UR824

Table 2　Sampling frequency of each microphone in asynchronous recording.

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

Table 3　Experimental condition.

Source	4 people
Frame length	4,096 samples
Frame shift	2,048 samples
Signal length for supervised training	10 s
Signal length for evaluation	10 s
β-divergence	$\beta = 1.0$
Initialization parameter α	$\alpha = 0.1$
Number of NMF iterations	200

ずみ, SIR は非目的信号の抑圧率を評価する尺度となっており，値が大きいほど目的音強調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。評価値は，未処理の観測信号（Unproc.）と，SN 比最大化ビームフォーマ（mSNRbf）[7], 振幅スペクトルビームフォーマ（ASbf）[8], 独立ベクトル分析（IVA）[17], 罰則付き教師なし伝達関数ゲイン基底 NMF（PNMF）[9], 教師あり伝達関数ゲイン基底 NMF（SNMF）[10]，における各音源の強調信号に対して算出した。ここで，IVA と mSNRbf は位相情報を用いるアレー信号処理手法であり，サンプリング周波数ミスマッチがあるような観測に対しては目的音強調性能が劣化することが予想される。なお，mSNRbf, ASbf, SNMF は単一音源区間での事前学習が必要な手法であり，強調区間とは異なる区間で事前学習を行った。

5.2 評価結果

同期録音の SDR を **Fig. 5** に，非同期録音の SDR を **Fig. 6** に示す。また，同期録音の SIR を **Fig. 7** に，非同期録音の SIR を **Fig. 8** に示す。なお，PNMF では罰則項の大きさの調整により，SDR の平均が最も高くなったときの評価値を掲載している。実際の話者を用いた実験であるため，話者ごとの発話音量差が存在し，入力 SNR が大きくばらついている。この様子は，Unproc. による SDR に表れている。位相情報を用いる mSNRbf, IVA では非同期録音において SDR, SIR ともに大きく低下していることから，観測信号間での

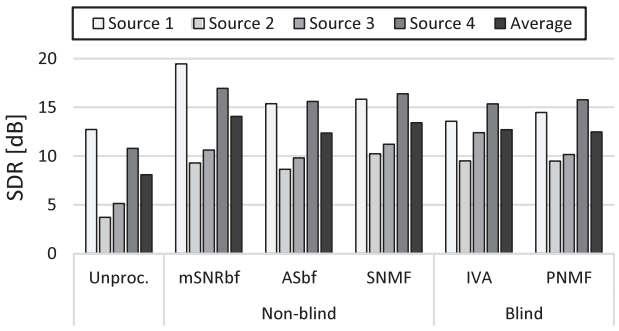

Fig. 5　The comparison of SDR between each method in a synchronous recording case.

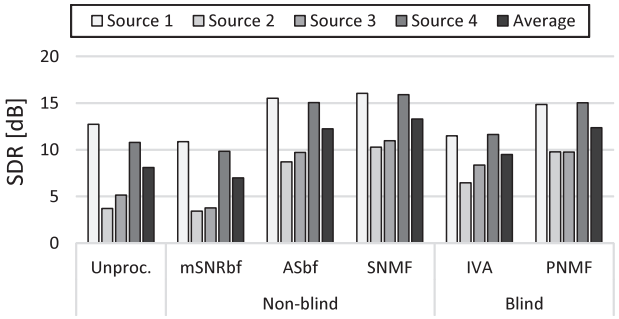

Fig. 6　The comparison of SDR between each method in an asynchronous recording case.

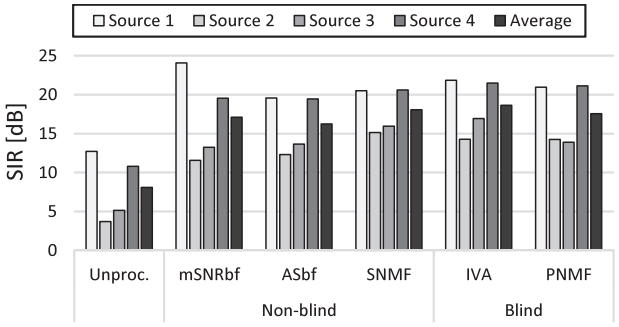

Fig. 7　The comparison of SIR between each method in a synchronous recording case.

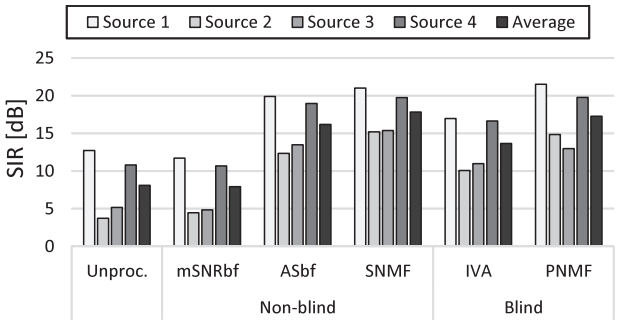

Fig. 8　The comparison of SIR between each method in an asynchronous recording case.

位相ずれによって目的音強調性能が劣化したと考えられる。一方，ASbf, PNMF, SNMF による強調では，未処理の観測信号（Unproc.）と比較して，SDR, SIR ともに大きく向上している。同期録音と非同期録音の

結果を比較したところ，対応する各音源での評価値に差が見られない。ノンブラインドの手法である ASbf と SNMF を比較すると，SNMF のほうが目的音強調性能が高いことが確認できる。

以上より，伝達関数ゲイン基底 NMF が機器間の位相差ずれに頑健であり，非同期分散型マイクロホンアレーにおける目的音強調に適した手法であることが確認できた。

6. 非同期機器による実録音における強調性能の評価

6.1 実験条件

実際に非同期機器を用いて収録した録音データによって伝達関数ゲイン基底 NMF の目的音強調性能を評価した。録音データは，**Fig. 9** のようなマイク・話者配置とし，非同期分散型マイクロホンアレーにより話者ごとに 1 人ずつ **Table 4** のような環境で収録した。収録後，観測信号の録音開始時刻を相互相関を用いて揃え，混合することで混合信号を生成した。収録は全話者を変更して 2 回行った。実験条件を **Table 6** に示す。評価する手法は，前章と同じく，未処理の観測信号（Unproc.）と，SN 比最大化ビームフォーマ（mSNRbf），振幅スペクトルビームフォーマ（ASbf），独立ベクトル分析（IVA），罰則付き教師なし伝達関数ゲイン基底 NMF（PNMF），教師あり伝達関数ゲイン基底 NMF（SNMF）である。これらの手法において各音源の強調信号に対して SDR を算出した。

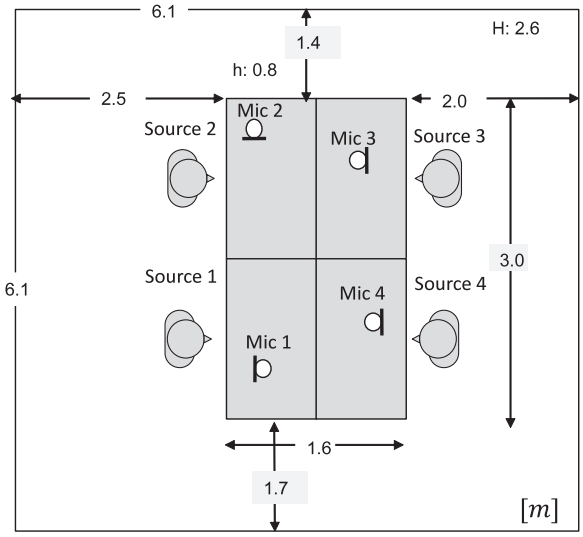

Fig. 9 Arrangement of microphones used in evaluation.

Table 4 Recording environment.

Reverberation time	0.65 s
Noise level	31.5 dBA

Table 5 Recording devices.

Mic 1	Apple iPhone 5S
Mic 2	Apple iPhone 5S
Mic 3	Sony Mobile Xperia
Mic 4	Apple iPhone 5

6.2 評価結果

まず，**Fig. 10** で示す，ある話者の発話に対して相互相関関数を適用した結果を **Fig. 11** に示す。このように，位相のドリフトが存在している観測信号におい

Table 6 Experimental condition.

Source	4 people×2
Codec	PCM
Sampling-rate	16 kHz
Bit-depth	16-bit
Frame length	4,096 samples
Frame shift	2,048 samples
Signal length for supervised training	10 s
Signal length for evaluation	10 s
β-divergence	$\beta = 1.0$
Initialization parameter α	$\alpha = 0.1$
Number of NMF iterations	200

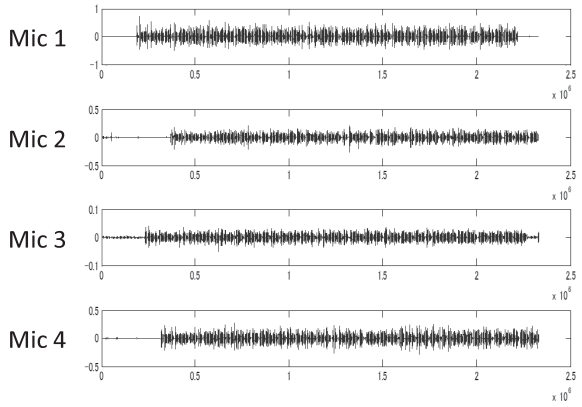

Fig. 10 The observed signals on recording a certain source.

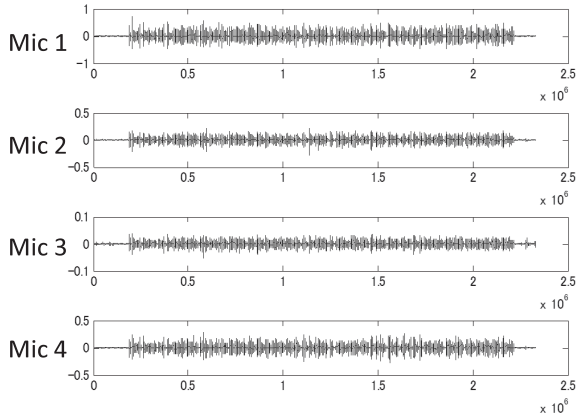

Fig. 11 Apply cross-correlation to the observed signals on recording a certain source.

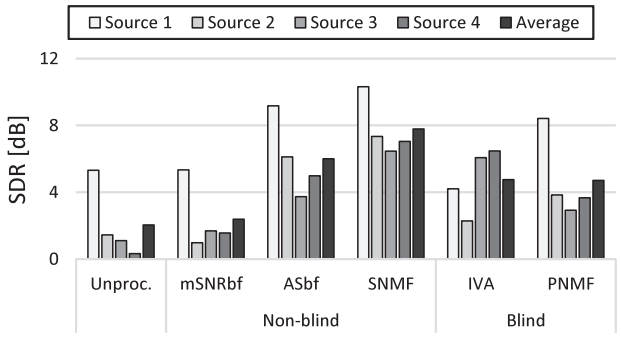

Fig. 12 The comparison of SDR between each method in an asynchronous recording (Case 1).

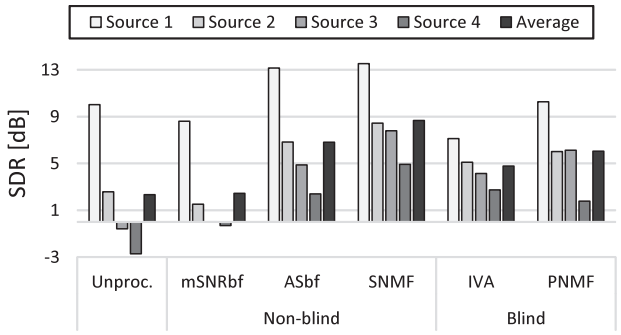

Fig. 13 The comparison of SDR between each method in an asynchronous recording (Case 2).

ても，録音開始時刻を大まかにそろえることが可能であることを確認した。

各手法により生成した強調信号の SDR を **Fig.** 12, 13 に示す。ここで，Case 1 と Case 2 は話者を全員変更して録音した観測である。なお，前章の実験と同様に，PNMF では罰則項の大きさの調整により，SDR の平均値が最も高くなったときの評価値を掲載している。実際の非同期機器を用いた実験であるため，本論文で着目している録音開始時刻やサンプリング周波数ミスマッチ以外に，マイク感度や指向性などの録音機器個体差，話者ごとの発話音量差が存在している。そのため，入力 SNR が大きくばらついている。この様子は，Unproc. による SDR に表れている。

Fig. 12, 13 から，位相情報を用いる mSNRbf 及び IVA は音源によっては強調性能が劣化していることが分かる。一方，PNMF 及び SNMF による強調信号と未処理の観測信号を比較すると，すべての音源に対して音声強調効果があることが確認できる。

以上の実験結果から，振幅ベースの音声強調手法である ASbf，PNMF，SNMF による目的音強調は実際の非同期機器で録音した観測信号においても適用可能であることが確認できた。

7. おわりに

本論文では，非同期録音に頑健な目的音強調手法として，チャネル間の振幅差に基づく目的音強調手法である伝達関数ゲイン基底 NMF を用いた時間周波数マスキングの性能を検証した。評価では戸上らの罰則付き伝達関数ゲイン基底 NMF [9] と，著者らの先行研究 [10] である，単一音源区間で伝達関数ゲイン基底を事前学習する教師あり伝達関数ゲイン基底 NMF を用いた。実環境での録音を用いた実験結果より，サンプリング周波数ミスマッチによらず，これらの伝達関数ゲイン基底 NMF が非同期録音に対して適用可能な，位相ずれに頑健な強調手法であることを確認した。また，教師あり伝達関数ゲイン基底 NMF の優れた目的音強調効果を確認した。

謝　辞

本論文は，科学研究費補助金基盤研究（B）(25280069) 及びセコム科学技術振興財団の助成を受けたものです。また，本論文を進める上で有益なご助言をいただきました，日立製作所中央研究所の戸上真人氏に感謝申し上げます。

文　献

[1] 小野順貴, 宮部滋樹, 牧野昭二, “非同期分散マイクロホンアレイに基づく音響信号処理,” 音響学会誌, 70, 391–396 (2014).
[2] 小野順貴, Trung-Kien Le, 宮部滋樹, 牧野昭二, “アドホックマイクロホンアレー—複数のモバイル録音機器で行う音響信号処理—,” 信学会 *Fundam. Rev.*, 7, 336–347 (2014).
[3] E. Robledo-Arnuncio, T.S. Wada and B.-H. Juang, “On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation,” *Proc. WASPAA*, pp. 34–37 (2007).
[4] Z. Liu, “Sound source separation with distributed microphone arrays in the presence of clock synchronization errors,” *Proc. IWAENC*, pp. 1–4 (2008).
[5] S. Miyabe, N. Ono and S. Makino, “Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain,” *Proc. ICASSP*, pp. 674–678 (2013).
[6] R. Sakanashi, N. Ono, S. Miyabe, T. Yamada and S. Makino, “Speech enhancement with ad-hoc microphone array using single source activity,” *Proc. APSIPA*, pp. 1–6 (2013).
[7] S. Araki, H. Sawada and S. Makino, “Blind speech separation in a meeting situation with maximum SNR beamformers,” *Proc. ICASSP*, pp. 41–45 (2007).
[8] 加古達也, 小林和則, 大室　仲, “非同期分散マイクアレーのための振幅スペクトルビームフォーマの提案,” 音講論集, pp. 829–830 (2013.3).
[9] M. Togami, Y. Kawaguchi, H. Kokubo and Y. Obuchi, “Acoustic echo suppressor with multichannel semi-blind non-negative matrix factorization,” *Proc. APSIPA*, pp. 522–525 (2010).
[10] H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada and S. Makino, “Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording,” *Proc. IWAENC*, pp. 204–208 (2014).
[11] D.D. Lee and H.S. Seung, “Algorithms for non-negative matrix factorization,” *Adv. Neural Inf. Process. Syst.*, 13, 556–562 (2001).

[12] R. Kompass, “A generalized divergence measure for nonnegative matrix factorization,” *Neural Comput.*, **19**, 780–791 (2007).
[13] M. Nakano, H. Kameoka and J.L. Roux, “Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence,” *Proc. MLSP*, pp. 283–288 (2010).
[14] P. Smaragdis, B. Raj and M. Shashanka, “Supervised and semi-supervised separation of sounds from single-channel mixtures,” *Proc. ICA*, pp. 414–421 (2007).
[15] A. Cichocki, R. Zdunek and S. Amari, “New algorithms for non-negative matrix factorization in applications to blind source separation,” *Proc. ICASSP*, pp. 621–624 (2006).
[16] E. Vincent, R. Gribonval and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio Speech Lang. Process.*, **14**, 1462–1469 (2006).
[17] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” *Proc. WASPAA*, pp. 189–192 (2011).

千葉　大将

2013 筑波大・情報・情報科卒。2015 同大学大学院・シス情・CS 博士前期課程了。修士（工学）。同年パイオニア株式会社に入社。在学中，アレー信号処理，音声符号化に関する研究に従事。日本音響学会会員。

小野　順貴

2001 東大博士後期課程修了。同年 同大学助手。2005 同大学講師。2011 国立情報学研究所 准教授。アレイ信号処理，音源定位，音源分離などの音響信号処理の研究に従事。博士（工学）。IEEE Senior member，日本音響学会，電子情報通信学会，情報処理学会，計測自動制御学会，各会員。

宮部　滋樹

2007 奈良先端大博士後期課程了。2008 米ジョージア工科大学客員研究員。2009 東大特任研究員。2010 同大助教。2011 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，各会員。

髙橋　　祐

2010 奈良先端大博士後期課程了。同年ヤマハ株式会社に入社。音声強調・音源分離処理に関する研究に従事。博士（工学）。日本音響学会，IEEE，人工知能学会，各会員。

山田　武志

1999 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。

牧野　昭二

1981 東北大大学院修士課程了。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。文部科学大臣表彰（科学技術賞 研究部門），ICA Unsupervised Learning Pioneer Award，IEEE Signal Processing Society Best Paper Award 受賞。IEEE Distingushed Lecturer。IEEE Fellow。電子情報通信学会 Fellow。