伝達関数ゲイン基底NMFによる 分散配置非同期録音における目的音強調の検討*

☆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), 高橋 祐 (ヤマハ)

1 はじめに

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する非同期分散型マイクロホンアレーはマイク数や配置等の構成が柔軟であるため、高いSN比での収音が期待できる。しかし、非同期機器間のサンプリング周波数のずれにより、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、非同期録音では従来のマイクロホンアレー信号処理は性能が劣化してしまう。そのため、非同期録音において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に対する同期補正 [1][2] 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手法として、振幅情報のみを用いたSN比最大化ビームフォーマによる目的音強調 [3] のような各音源の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。本稿では、分散配置の容易さを活かし、各音源の近くにマイクロフォンを配置できる場合には観測信号間のゲイン比が音源毎に明確に異なることを利用し、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)[4] による時間周波数マスキングが非同期録音において頑健な目的音強調手法であることを検証する。評価では、非同期分散型マイクロホンアレーを利用した会議録音を想定した録音データを用いて、ある特定の話者のみが発話している単一音源区間より伝達関数ゲイン基底を学習した教師あり NMF[5] による目的音強調性能を確認する。

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 問題設定

本稿では信号を時間周波数領域での複素振幅として表現する。また、i 行 j 列に X_{ij} を成分として持つ $I \times J$ の行列 $\mathbf{X}$ を $\mathbf{X} = [x_{ij}]_{ij} \in \mathbb{C}^{I \times J}$ と表すこととする。

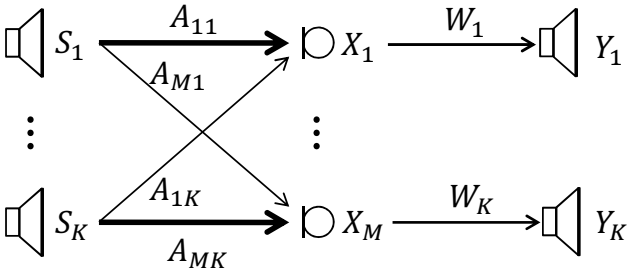

Fig. 1 Mixture model for speech enhancement.

Fig 1 に混合モデルを示す。非同期分散型マイクロホンアレーによる多チャネルの観測信号を、

$$\mathbf{X}(\omega) \approx \mathbf{A}(\omega)\mathbf{S}(\omega) \quad (1)$$
$$\mathbf{X} = [X_{mn}]_{mn} \in \mathbb{C}^{M \times N} \quad (2)$$
$$\mathbf{A} = [A_{mk}]_{mk} \in \mathbb{C}^{M \times K} \quad (3)$$
$$\mathbf{S} = [S_{kn}]_{kn} \in \mathbb{C}^{K \times N} \quad (4)$$

と表す。ここで、ω は周波数ビン番号、N は時間フレーム数、M はマイクロホン素子数、K は音源数であり、音源 k における n 番目の時間フレームの音源を $S_{kn}(\omega)$、マイク m における n 番目の時間フレームの観測信号を $X_{mn}(\omega)$、m 番目のマイクロホン素子と K 番目の音源間の伝達関数を $A_{mk}(\omega)$ とする。時間周波数領域での目的音強調では、音源 k を目的音として収録したマイク m における観測信号 $X_{mn}(\omega)$ より音源 k を強調する時間周波数領域マスク $W_{kn}(\omega)$ を用いて、

$$Y_{kn}(\omega) = W_{kn}(\omega, k)X_{mn}(\omega) \quad (5)$$

として目的信号を強調した信号 $Y_{kn}(\omega)$ を得る。ここで、$0 \leq M_k(\omega) \leq 1$ であり、本稿ではマイク m で観測する信号は音源 $k = m$ が目的信号であると仮定する。以降の議論は本節同様に周波数独立のモデル化と処理を議論し、全ての処理は周波数ビン毎に行うため、周波数を表す記号 ω は省略する。

---

*Speech enhancement by non-negative matrix factorization based on transfer function gain for asynchronous distributed recording. 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), Yu TAKAHASHI (Yamaha)

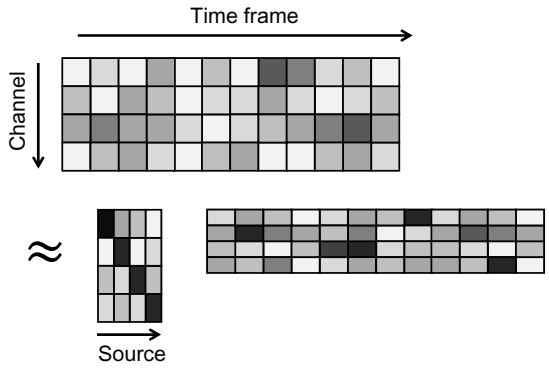

Fig. 2 In channel-time domain representation of observed signals for each frequency bin.

2.2 チャネル時間領域における伝達関数ゲイン基底 NMF

式 (1) のように複素数の積和で表される時間周波数領域の観測信号を、次式のように振幅スペクトル領域の関和で近似し、Fig 2 のような伝達関数ゲインを基底とした NMF により、

$$\bar{\mathbf{X}} \approx \bar{\mathbf{A}}\bar{\mathbf{S}} \tag{6}$$

$$\bar{\mathbf{X}} = [|X_{mn}|]_{mn} \in \mathbb{C}^{M\times N} \tag{7}$$

$$\bar{\mathbf{A}} = [A_{mk}]_{mk} \in \mathbb{C}^{M\times K} \tag{8}$$

$$\bar{\mathbf{S}} = [S_{kn}]_{kn} \in \mathbb{C}^{K\times N} \tag{9}$$

として各パラメタを推定することを考える。ここで $\bar{A}_{mn} \geq 0$ は m 番目のマイクロホン素子と k 番目の音源間の伝達関数ゲインの推定値、$\bar{S}_{kn} \geq 0$ は n 番目の時間フレームにおける第 k 音源の振幅の推定値を表す。本稿では β-divergence 規準 NMF[6] による以下の乗法型更新式、

$$\bar{S}_{kn} \leftarrow \bar{S}_{kn}\left(\frac{\sum_m \bar{X}_{mn}(\sum_k \bar{A}_{mk}\bar{S}_{kn})^{\beta-2}\bar{A}_{mk}}{\sum_m(\sum_k \bar{A}_{mk}\bar{S}_{kn})^{\beta-1}\bar{A}_{mk}}\right)^{\psi(\beta)} \tag{10}$$

$$\bar{A}_{mk} \leftarrow \bar{A}_{mk}\left(\frac{\sum_n \bar{X}_{mn}(\sum_k \bar{A}_{mk}\bar{S}_{kn})^{\beta-2}\bar{S}_{kn}}{\sum_n(\sum_k \bar{A}_{mk}\bar{S}_{kn})^{\beta-1}\bar{S}_{kn}}\right)^{\psi(\beta)} \tag{11}$$

を更新することで局所的最小解を得る。ここで、

$$\psi(\beta) = \begin{cases} \frac{1}{(2-\beta)} & \beta \geq 1 \\ 1 & 1 \leq \beta \leq 2 \\ \frac{1}{(\beta-1)} & \beta \leq 2 \end{cases} \tag{12}$$

であり、$\beta = 0$ の場合は板倉斎藤ダイバージェンス規準、$\beta = 1$ の場合は I ダイバージェンス規準、$\beta = 2$ の場合はフロベニウスノルム規準での更新式となる。また、式 (10)、(11) の更新ごとに基底を、

$$\bar{A}_{mk} \leftarrow \frac{\bar{A}_{mk}}{\sum_m \bar{A}_{mk}} \tag{13}$$

$$\bar{S}_{kn} \leftarrow \left(\sum_m \bar{A}_{mk}\right)\bar{S}_{kn} \tag{14}$$

として正規化する。

以上の NMF アルゴリズムにより伝達関数ゲインを表す基底行列が周波数ビンごとに得られるが、周波数領域 ICA などの従来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。ここで、本稿で扱う分散型マイクロホンアレー配置では、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はるかに小さいと仮定できるため、基底行列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、k 番目の音源を目的信号とするマイク番号を $m = k$ とし、基底行列 $\bar{\mathbf{A}}$ の初期値を、

$$\bar{A}_{mk} = \begin{cases} 1 & (m = k) \\ \alpha & (m \neq k) \end{cases} \tag{15}$$

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、$\alpha \ll 1$ となる任意の正の実数である。

2.3 伝達関数ゲイン基底の学習

前節の伝達関数ゲイン基底 NMF は、基底として各話者ごとの伝達関数ゲインを用いていることから、ある話者のみが発話している単一音源区間において伝達関数ゲイン基底を学習することでより最適解に近い解を得ることができると考えられる。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。すなわち、音源 k のみの単一音源区間において、式 (10)、(11) 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 $\mathbf{a}_k = [a_{mk}]_{mk} \in \mathbb{C}^{M\times 1}$ 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 $\bar{\mathbf{A}} = \left(\bar{\mathbf{a}}_1 \cdots \bar{\mathbf{a}}_K\right)$ を学習する。目的音強調区間では、基底行列の初期値として学習した基底行列 $\bar{\mathbf{A}}$ を与え、アクティベーション行列のみ式 (10) で更新する。

2.4 ウィーナーフィルタによる時間周波数領域マスキング

時間周波数領域でのウィーナーフィルタによる目的音 k を強調するマスク W_{kn} は信号 k の推定パワー値 $\left(\bar{A}_{mk}\bar{S}_{kn}\right)^2$ より、

$$W_{kn} = \frac{\left(\bar{A}_{mk}\bar{S}_{kn}\right)^2}{\sum_k \left(\bar{A}_{mk}\bar{S}_{kn}\right)^2} \tag{16}$$

として求めることができる。

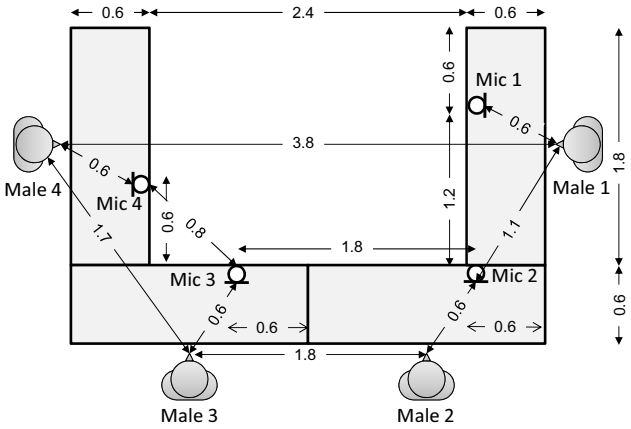

Fig. 3　Arrangement of microphones using on evaluation.

Table 1　Recording environment.

マイクロホン	SHURE SM57
Power amp.	YAMAHA XM4080
AD/DA	Steinberg UR824
机	0.6 m×1.7 m×0.7 m
マイク-話者間距離	0.6m
話者	日本人男性 4 人
気温	18.4 ℃
湿度	37 %

Table 2　Setting artificial sampling frequency mismatches for each microphones 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.

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
音声長	90 sec
基底行列の学習区間	0 - 10 sec
目的音強調区間	70 - 80 sec
β ダイバージェンス	$\beta = 0.0, 1.0, 2.0$
基底行列初期値 α	$\alpha = 0.25$
NMF 更新回数	200 回

3　評価実験

3.1　実験条件

実際の会議録音を想定した録音データにより提案手法の性能を評価した。提案手法が非同期録音で発生するような機器間のサンプリング周波数の違いによる位相差ずれに頑健な手法であることを示すために、同期録音データ及び非同期録音データで評価を行った。録音データは、Fig. 3 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 1 のような環境で録音した。録音後、マイク毎に Table 2 に示すサンプリング周波数でリサンプリングを行って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。評価尺度は SDR (Signal to Distortion Ratio) と SIR (Source to Interference Ratio) を用いた [7]。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となっており、値が大きいほど目的音強調の性能が良いことを示す。評価値は、未処理 (No processing)、板倉斎藤ダイバージェンス規準の教師あり伝達関数ゲイン NMF (SNMF-ISdiv)、I ダイバージェンス規準の教師あり伝達関数ゲイン NMF (SNMF-Idiv)、フロベニウスノルム規準の教師あり伝達関数ゲイン NMF (SNMF-Fnorm) における各観測信号ごとに算出する。

3.2　評価結果

同期録音の SDR を Fig. 4 に、非同期録音の SDR を Fig. 5 に示す。また、同期録音の SIR を Fig. 6 に、非同期録音の SIR を Fig. 7 に示す。未処理の No processing と比較して、教師ありの伝達関数ゲイン基底 NMF では各観測信号において SDR、SIR ともに大きく向上している。また、同期録音と非同期録音の結果を比較したところ、対応する各値に差が見られないことから、提案手法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。なお、実験条件は同じではないが、[1] では SDR が 5 dB 程度の音声に対して位相情報を用いたアレイ信号処理（独立ベクトル分析）で音源分離を行った結果を行った結果、同期の場合は 20 dB 程度分離できる場合でも、サンプリング周波数が最大で ±1.5Hz ほどずれただけで 0 dB 付近まで低下することが示されている。

4　まとめ

本稿では、非同期録音に頑健な目的音強調として複数マイクロホン素子間の伝達関数ゲイン

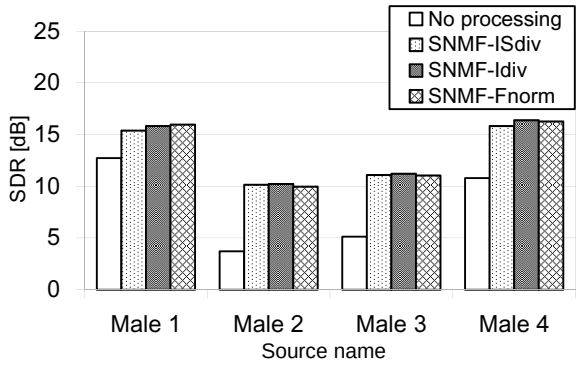

Fig. 4 SDR values of proposal method on a synchronous recording.

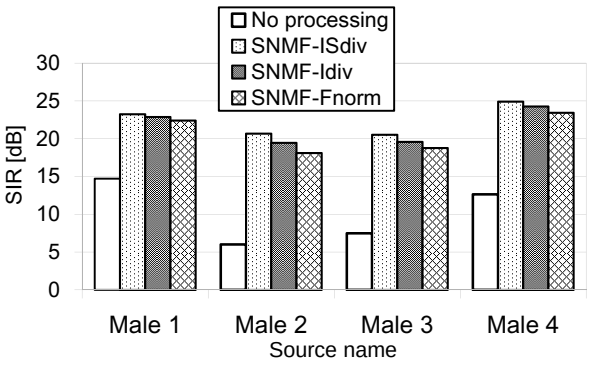

Fig. 6 SIR values of proposal method on a synchronous recording.

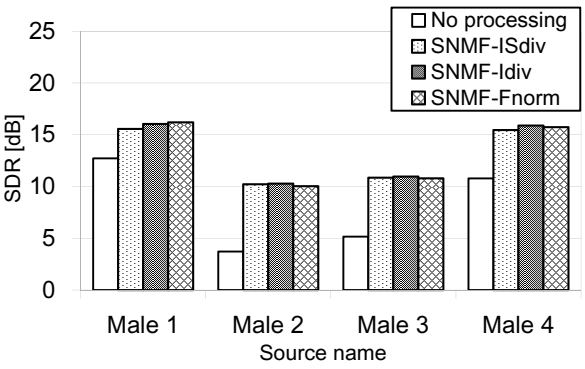

Fig. 5 SDR values of proposal method on an asynchronous recording.

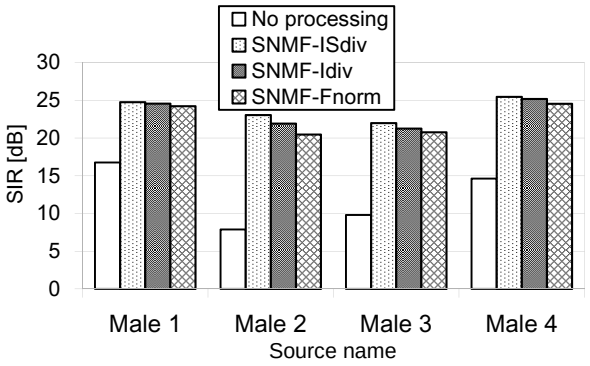

Fig. 7 SIR values of proposal method on an asynchronous recording.

の差に基づく NMF を用いた時間周波数マスキングの性能を評価した。評価の結果、教師あり伝達関数ゲイン NMF によって時間周波数マスキングは元の信号よりも SDR、SIR が大きく向上しており、位相ずれに頑健で非同期録音に対して使用可能な目的音強調手法であることを確認した。

参考文献

[1] 宮部 他, “非整数サンプルシフトのフレーム分析を用いた非同期録音の同期化, ” 音講論 (秋) , pp. 593-596, 2013.

[2] R. Sakanashi et al., “Speech enhancement with ad-hoc microphone array using single source activity, ” in Proc. APSIPA2013, OS.21-SLA.7.5, 6 pages, Oct. 2013.

[3] 加古 他, “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 (春) , pp. 829-830, 2013.

[4] 戸上 他, “音源のチャンネル間振幅差を基底ベクトルとする音源分離,” 音講論 (春) , pp. 803-804, 2010.

[5] P. Smaragdis et al., “Supervised and semi-supervised separation of sounds from single-channel mixtures,” in Proc.ICA 2007, pp. 414-421, 2007.

[6] R. Kompass, “A generalized divergence measure for nonnegative matrix factorization,” Neural Computation,19(3), pp. 780-791, 2007.

[7] E. Vincent et al., “Performance measurement in blind audio source separation,” IEEE Trans. ASLP, 14(6), 1462-1469, 2006.