教師あり多チャネルNMFと統計的音声強調を用いた柔軟索状ロボットにおける音源分離*

☆高田一真 (東大), 北村大地 (総研大), 中嶋広明 (東大), 小山翔一 (東大), 猿渡洋 (東大), 小野順貴 (NII/総研大), 牧野昭二 (筑波大)

1 はじめに

近年，自然災害等に被災した人々を効率的に捜索する為の遠隔操作可能なロボットの開発が期待されている．その中でも柔軟索状ロボットは，蛇状の特殊な形をしていることから，通常の遠隔操作ロボットでは探索できない環境下での捜索に活用できる．柔軟索状ロボットは，本体の周りに一定間隔でマイクロホンが取り付けられているが．本体形状が刻一刻と変化するため，一般的な固定マイクロホンアレイのように音声強調を行うことができない．また，柔軟索状ロボットは，駆動する際に本体自身からノイズが発生し，観測する音に直接混合されてしまう，そのため，音声等の目的音源の推定精度が劣化してしまう問題がある．本稿では，このようなノイズを「エゴノイズ」と呼ぶ．

本稿では，柔軟索状ロボットで得られた多チャネル観測信号から，エゴノイズと目的音声を分離する手法を提案する．特に，非負値行列因子分解（non-negative matrix factorization: NMF）[1] を多チャネル観測信号に拡張したランク 1 多チャネル NMF（rank-1 multichannel NMF: Rank-1 MNMF）[2]，ベイズ型ポストフィルタの一種である一般化平均二乗誤差最小化短時間振幅スペクトル推定器（MMSE estimation with Optimizable Speech model and Inhomogeneous Error criterion: MOSIE）[3]，及び全極モデルを用いた基底変形 [4] を用いる．実際に柔軟索状ロボットが観測する音を模擬的に再現し，従来の音源分離手法よりも高い音声強調が実現できることを示す．

Fig. 1　Hose-shaped rescue robot.

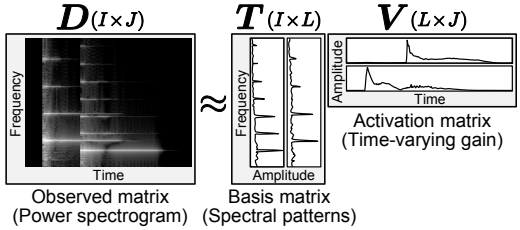

Fig. 2　Decomposition model of simple NMF ($L=2$).

2 柔軟索状ロボットと従来の音源分離手法

2.1 柔軟索状ロボットとエゴノイズ

柔軟索状ロボットは，災害等での被災者の捜索を目的とした救助ロボットである（Fig. 1 参照）．また，全体には繊毛が巻き付けられており，その節々にマイクロホンとバイブレータが複数個搭載されている．このバイブレータが繊毛とロボット全体を振動させることで，他の動力無しに前進することができる特徴を持つ．しかし，動力源がバイブレータであることに起因して，自身の発するエゴノイズが各マイクロホンに収録されてしまい，音声などの重要な目的音の検知や推定に悪影響を及ぼしてしまう問題がある．このようなエゴノイズは，バイブレータ自身の駆動音の他に，ロボットが地面と接する振動音等も含まれているため，ロボットの姿勢や地面の状況等によって，各マイクロホンが観測するエゴノイズは異なり，それらが時間的にも変化する．

2.2 音源の瞬時混合モデル

音源数と観測チャネル数をそれぞれ N, M とし，各時間周波数における多チャネル音源信号，多チャネル観測信号，及び分離信号をそれぞれ

$$\boldsymbol{s}_{\omega,\tau} = (s_{\omega,\tau,1} \dots s_{\omega,\tau,N})^{\mathrm{t}} \tag{1}$$

$$\boldsymbol{x}_{\omega,\tau} = (x_{\omega,\tau,1} \dots x_{\omega,\tau,M})^{\mathrm{t}} \tag{2}$$

$$\boldsymbol{y}_{\omega,\tau} = (y_{\omega,\tau,1} \dots y_{\omega,\tau,N})^{\mathrm{t}} \tag{3}$$

とする（要素は全て複素数）．ここで，$\omega = 1, \dots, \Omega$ は周波数インデクス，$\tau = 1, \dots, T_1$ は時間フレームインデクス，$n = 1, \dots, N$ は音源インデクス，$m = 1, \dots, M$ はチャネルインデクスを示し，$^{\mathrm{t}}$ は転置を表す．ここでは，目的音源（音声）やエゴノイズが各マイクロホンに伝わる際の伝達系が，周波数領域における瞬時混合系であると仮定する．この仮定を用いると，観測音と音源の関係式を次式で表すことができる．

$$\boldsymbol{x}_{\omega\tau} = \boldsymbol{A}_{\omega}\boldsymbol{s}_{\omega\tau} \tag{4}$$

ただし，$\boldsymbol{A}_{\omega} = (\boldsymbol{a}_{\omega,1} \dots \boldsymbol{a}_{\omega,N})$ は混合行列を示し，$\boldsymbol{a}_{\omega,n}$ は各音源のステアリングベクトル（n 番目の音源から各マイクロホンまでの伝達系を示す複素ベクトル）である．$M = N$ で $\boldsymbol{A}_{\omega}$ がフルランクであれば，分離行列 $\boldsymbol{W}_{\omega} = \boldsymbol{A}_{\omega}^{-1}$ が定義でき，観測音と分離音の関係は以下のようになる．

$$\boldsymbol{y}_{\omega\tau} = \boldsymbol{W}_{\omega}\boldsymbol{x}_{\omega\tau} \tag{5}$$

2.3 NMF

NMF は，観測音のパワースペクトログラムを，非負の基底行列と非負のアクティベーション行列に近似的に分解する手法であり，有意なスペクトル特徴量を抽出することができる．これは，非負制約付きの低ランク近似分解と解釈できる．NMF の分解モデルは次

---

* "Source separation with supervised multichannel NMF and statistical speech enhancement method for hose-shaped rescue robot," by Kazuma Takata (The University of Tokyo), Daichi Kitamura (SOKENDAI), Hiroaki Nakazima (The University of Tokyo), Shoichi Koyama (The University of Tokyo), Saruwatari Hiroshi (The University of Tokyo), Nobutaka Ono (NII/SOKENDAI) and Shoji Makino (University of Tsukuba).

式で表される．

$$\boldsymbol{D} \approx \boldsymbol{T}\boldsymbol{V} \tag{6}$$

$\boldsymbol{D}$ ($\in \mathbb{R}_{\geq 0}^{\Omega \times T_1}$) は要素が非負のパワースペクトルグラムの行列，$\boldsymbol{T}$ ($\in \mathbb{R}_{\geq 0}^{\Omega \times L}$) は $\boldsymbol{D}$ に含まれる頻出スペクトルパターンを列ベクトルに L 本持つ基底行列，$\boldsymbol{V}$ ($\in \mathbb{R}_{\geq 0}^{L \times T_1}$) は $\boldsymbol{T}$ 中の各基底に対応する時間的な強度変化を行ベクトルとして持つアクティベーション行列である．Figure 2 に NMF の分解モデル図を示す．パワースペクトログラム中に頻出している 2 個のスペクトルパターンが基底として得られ，それらの時間的な強度変化がアクティベーションに表れていることが分かる．

2.4 Rank-1 MNMF

Rank-1 MNMF は，独立ベクトル分析（independent vector analysis: IVA）[5] と同様に周波数領域での瞬時混合（式（4））を仮定した線形分離手法である．これは，従来の多チャネル NMF [6] における空間相関行列のランクを 1 に制約した空間モデルを持つ．IVA の音源モデルは，球対称多変量分布を用いたシンプルなものであったが，Rank-1 MNMF は音源モデルが NMF による基底分解に拡張されている．より厳密なスペクトログラムを捉えることができるため，分離精度の向上につながる他，事前学習したスペクトル基底を与えることで，教師ありの線形分離手法に拡張できる．

2.5 MOSIE

MOSIE は，ベイズ型ポストフィルタの一つであり，目的音声の推定成分（音源分離手法の出力）と各音源の事前分布を用いる音声強調手法である．事前分布が良く適合する場合には，前段の音源分離よりも高精度な音声強調及び雑音抑圧が可能となる．

MOSIE は，各周波数ビンにおいて，時間方向の振幅成分に対する事前分布を仮定する．目的音声成分に対してはカイ分布を仮定し，それ以外の音源には零平均の複素ガウス分布を仮定することで，それらのモデルに適合するように分離が強調される．ある周波数ビンにおける確率密度関数は以下で与えられる．

$$P(\chi) = \frac{2}{\Gamma(\rho)} \left(\frac{\rho}{\mathrm{E}(\rho^2)} \right)^{\rho} \chi^{2\rho-1} \exp\left(-\frac{\rho}{\mathrm{E}(\chi^2)} \right) \chi^2 \tag{7}$$

$$P(\zeta|\chi, \theta) = \frac{1}{\pi \mathrm{E}(\alpha^2)} \exp\left(-\frac{|\zeta - \chi e^{j\theta}|^2}{\mathrm{E}(\alpha^2)} \right) \tag{8}$$

ここで，χ は目的音声の振幅スペクトルを表す確率変数，ρ は確率変数 χ の確率密度関数（カイ分布）の形状母数，E[·] は期待値演算を示す．また，α はエゴノイズの振幅スペクトルを表す確率変数であり，ζ は周波数ビン ω，時間フレーム τ，チャネル m での観測音の振幅スペクトルを表す確率変数，θ は目的音声の位相スペクトルの確率変数，j は虚数単位である．従って，式（7）は目的音声がカイ分布に従い，式（8）はエゴノイズが零平均複素ガウス分布に従うという仮定をそれぞれ示している．θ は χ とは独立で，一様分布に従うと仮定している．エゴノイズの確率変数の分散は，事前に適用する半教師あり Rank-1 MNMF より与えることができる．また，形状母数の ρ は，文献 [7] で提案されている手法で推定できる．式（7）と式（8）より目的音の振幅の事後分布を求めることができる．文献 [3] より，目的音声の振幅の最適な推定値はゲイン関数を用いて次式のように書くことができる．

$$\hat{\chi}_{\omega,\tau} = G_{\omega,\tau} \zeta_{\omega,\tau} \tag{9}$$

$$G_{\omega,\tau} = \frac{\sqrt{\nu_{\omega,\tau}}}{\gamma_{\omega,\tau}} \left(\frac{\Gamma(\frac{\beta}{2} + \rho) M(1 - \frac{\beta}{2} - \rho, 1; -\nu_{\omega,\tau})}{\Gamma(\rho) M(1 - \rho, 1; -\nu_{\omega,\tau})} \right)^{\frac{1}{\beta}} \tag{10}$$

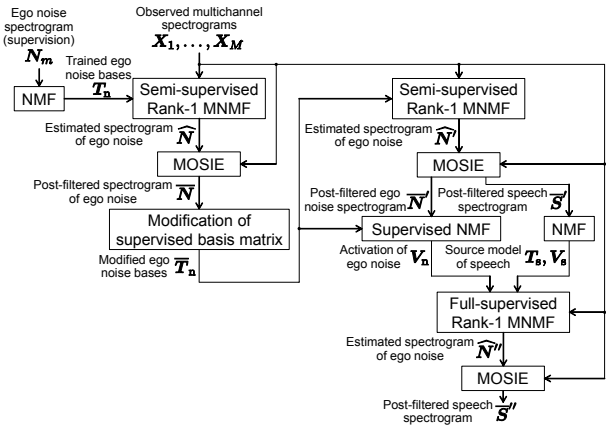

Fig. 3　Signal flow of proposed method.

ここで，$\Gamma(\cdot)$ はガンマ関数，$M(\cdot)$ は合流型超幾何関数である．$\nu_{\omega,\tau}$ は

$$\nu_{\omega,\tau} = \frac{\gamma_{\omega,\tau} \xi_{\omega,\tau}}{1 + \xi_{\omega,\tau}} \tag{11}$$

と定義され，$\xi_{\omega,\tau}$ は事前 SN 比，$\gamma_{\omega,\tau}$ は事後 SN 比で次式で表される．

$$\xi_{\omega,\tau} = K \gamma_{\omega,\tau-1} G_{\omega,\tau-1}^2 + (1 - K) \max[\gamma_{\omega,\tau} - 1, 0] \tag{12}$$

$$\gamma_{\omega,\tau} = \frac{|\zeta|_{\omega,\tau}^2}{\mathrm{E}[\alpha^2]} \tag{13}$$

K は忘却係数である．

3 提案手法

3.1 動機

従来の IVA 等の音源分離手法では，多変量分布に基づく音源モデルを仮定しているので，教師あり手法に拡張することができない．しかし，柔軟索状ロボットでは，無音声区間からエゴノイズのみが含まれる信号が容易に入手できるため，事前学習を伴う半教師あり音源分離が可能である．このようなアプローチは，音声強調あるいは雑音抑圧の性能を大幅に向上することが期待できるため，本稿では，Rank-1 MNMF を半教師ありに拡張した多チャネル音源分離を適用することを提案する．

3.2 処理の流れ

まずエゴノイズのみが含まれるパワースペクトログラムを用意し，それを従来の NMF で分解する．この事前学習によって，エゴノイズのみを表現するスペクトル基底（教師基底）が得られる．エゴノイズの教師基底を Rank-1 MNMF の音源モデルに用いることで，半教師あり多チャネル音源モデルを実現する．ただし，音声の音源モデルは依然として，ブラインドで推定されることに注意する．このような半教師あり手法においては，事前学習で得られるエゴノイズ基底が正確であることが極めて重要である．しかし，エゴノイズのスペクトルは，ロボットの姿勢や接地状況等に強く依存する上，時間的にも多少変化するため，事前学習のみで音源分離に必要十分な教師基底を得ることは通常困難である．そこで提案法では，Rank-1 MNMF の後段に，MOSIE による音声強調と教師基底の再学習（変形適応）手法 [4] を組み合わせることで，事前学習の曖昧性を解消し，エゴノイズの音源モデルを適応的に学習する高精度な音声強調法を提案する（Fig. 3 参照）．以下では各信号処理部の詳細について述べる．

3.2.1 半教師あり Rank-1 MNMF

適当なチャネル（m 番目のチャネル）で観測したエゴノイズのみの信号のパワースペクトログラム $\boldsymbol{N}_m$ に対して NMF を適用し，エゴノイズを表現する教師基底 $\boldsymbol{T}_{\mathrm{n}}$ ($\in \mathbb{R}^{\Omega\times L_1}$) を得る．次に，多チャネルの観測信号 $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_M$ に対して，教師基底 $\boldsymbol{T}_{\mathrm{n}}$ をエゴノイズの音源モデルに用いた半教師あり Rank-1 MNMF を適用し，エゴノイズの分離推定信号 $\widehat{\boldsymbol{N}}$ を得る．

3.2.2 エゴノイズの基底変形

まず，半教師あり Rank-1 MNMF のエゴノイズ推定信号 $\widehat{\boldsymbol{N}}$ の 2 乗を E[α^2] とし，観測信号の任意チャネル $\boldsymbol{X}$ を用いて，MOSIE（式 (9)-(13)）を実行することにより，次式のように分離精度の高い音声成分 $\overline{\boldsymbol{S}}$ とエゴノイズ成分 $\overline{\boldsymbol{N}}$ が得られる．

$$\overline{S}_{\omega,\tau} = G_{\omega,\tau} X_{\omega,\tau} \tag{14}$$

$$\overline{N}_{\omega,\tau} = (1 - G_{\omega,\tau}) X_{\omega,\tau} \tag{15}$$

$$G_{\omega,\tau} = \frac{\sqrt{\frac{|X|^2_{\omega,\tau}}{\widehat{N}^2_{\omega,\tau}} \frac{\xi_{\omega,\tau}}{1+\xi_{\omega,\tau}}}}{\frac{|X|^2_{\omega,\tau}}{\widehat{N}^2_{\omega,\tau}}} \left(\frac{\Gamma(\frac{\beta}{2}+\rho) M(1-\frac{\beta}{2}-\rho, 1; -\frac{|X|^2_{\omega,\tau}}{\widehat{N}^2_{\omega,\tau}} \frac{\xi_{\omega,\tau}}{1+\xi_{\omega,\tau}})}{\Gamma(\rho) M(1-\rho, 1; -\frac{|X|^2_{\omega,\tau}}{\widehat{N}^2_{\omega,\tau}} \frac{\xi_{\omega,\tau}}{1+\xi_{\omega,\tau}})} \right)^{\frac{1}{\beta}} \tag{16}$$

$\overline{S}_{\omega,\tau}, \overline{N}_{\omega,\tau}, X_{\omega,\tau}, \widehat{N}_{\omega,\tau}$ は、それぞれ $\overline{\boldsymbol{S}}, \overline{\boldsymbol{N}}, \boldsymbol{X}, \widehat{\boldsymbol{N}}$ の周波数 ω、時間フレーム τ の成分である．次に，$\overline{\boldsymbol{N}}$ とゲイン関数 $G_{\omega,\tau}$ を要素にもつ行列 $\boldsymbol{G}$ ($\in \mathbb{R}^{\Omega\times T_1}$) を用いて文献 [4] で提案されている全極フィルタによる教師基底変形を $\boldsymbol{T}_{\mathrm{n}}$ に施す．エゴノイズの基底変形は以下の最小化問題を解くことに帰着する．

$$\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{V}_{\mathrm{a}}} = \arg\min_{\boldsymbol{A},\boldsymbol{V}_{\mathrm{a}}} D_{\mathrm{KL}}(\boldsymbol{I} \circ \overline{\boldsymbol{N}} | \boldsymbol{I} \circ (\boldsymbol{A}\boldsymbol{T}_{\mathrm{n}}\boldsymbol{V}_{\mathrm{a}})) \tag{17}$$

ただし $D_{\mathrm{KL}}(\cdot)$ は一般化 KL ダイバージェンスである．ここで，バイナリマスク $\boldsymbol{I}$ ($\in \mathbb{R}^{\Omega\times T_1}$) は確信度を表す時間周波数行列であり，MOSIE ゲイン関数（式 (10)）がある閾値以上の時間周波数グリッドにおいては 1，かつそれ以外のグリッドでは 0 となる．$\boldsymbol{A}$ は対角成分に全極モデルによるスペクトル重みを持つ行列を表す．$\boldsymbol{V}_{\mathrm{a}}$ はアクティベーション行列である。$\boldsymbol{A}$ の対角成分は，次のように全極モデルで与える．

$$A_{\omega,\omega} = \frac{1}{1 - \Sigma_{k=1}^{p} \alpha_k \exp(-\pi j k \frac{\omega}{\Omega})} \tag{18}$$

ここで，p は全極モデルの次数で，α_k は全極モデルの係数である．$A_\omega = 1 - \Sigma_{k=1}^{p} \alpha_k \exp(-\pi j k \frac{\omega}{\Omega})$ と置くと，スペクトル重みの学習に用いる NMF の一般化 KL ダイバージェンスにおける目的関数 J および更新式は次式で与えられる．

$$J = \Sigma_{\omega,\tau} i_{\omega,\tau} [-\overline{N}_{\omega,\tau} + \frac{\Sigma_k t_{\omega,k} v_{k,\tau}}{A_\omega} + \overline{N}_{\omega,\tau} \log \frac{\overline{N}_{\omega,\tau}}{\Sigma_k t_{\omega,k} v_{k,\tau} / |A_\omega|} \tag{19}$$

$\overline{N}_{\omega,\tau}, t_{\omega,k}, v_{k,\tau}$ は，それぞれ $\overline{\boldsymbol{N}}, \boldsymbol{T}_{\mathrm{n}}, \boldsymbol{V}_{\mathbf{a}}$ に対応する行列の要素値である．この目的関数を最小化する際に更新するものは A_ω 及び、$v_{k,\tau}$ である．$k = 1,\ldots,L_1$ はエゴノイズ基底のインデックスである．補助関数法により $v_{k,\tau}$ 更新式は次式になる．

$$v_{k,\tau} \leftarrow v_{k,\tau} \frac{\Sigma_\omega i_{\omega,\tau} \overline{N}_{\omega,\tau} t_{\omega,k} / (\Sigma_k t_{\omega,k} v_{k,\tau})}{\Sigma_\omega i_{\omega,\tau} t_{\omega,k} / |A_\omega|} \tag{20}$$

ここで $R_{k,q}$ を要素に持つ行列を $\boldsymbol{R}$，r_q を要素に持つベクトル $\boldsymbol{r}$，α_k を要素に持つ $\boldsymbol{\alpha}$ を用いれば，次のようになる．

$$\boldsymbol{R}\boldsymbol{\alpha} = \boldsymbol{r} \tag{21}$$

$R_{k,q}$ と r_q の更新式は次のようになる．

$$\begin{aligned} R_{k,q} &= \Sigma_{\omega,\tau} [i_{\omega,\tau} (\Sigma_k t_{\omega,k} v_{k,t} \frac{1}{|A_\omega|^3} + \overline{N}_{\omega,\tau} \frac{1}{2|A_\omega|^2}) \\ &\quad (\exp(-\pi j \frac{\omega}{\Omega}(k-q) + \exp(\pi j \frac{\omega}{\Omega}(k-q)))) \end{aligned} \tag{22}$$

$$\begin{aligned} r_q &= \Sigma_{\omega,\tau} i_{\omega,\tau} [(\Sigma_k t_{\omega,k} v_{k,\tau} \frac{1}{|A_\omega|^3} + \overline{N}_{\omega,\tau} \frac{1}{2|A_\omega|^2}) \\ &\quad (\exp(-\pi j \frac{\omega}{\Omega} q + \exp(\pi j \frac{\omega}{\Omega})) \\ &\quad - \frac{3}{|A_\omega|^2} \Sigma_k t_{\omega,k} v_{k,\tau} \Re[\frac{A^*_\omega}{A_\omega} \exp(-\pi j \frac{\omega}{\Omega} q)])] \end{aligned} \tag{23}$$

ここで，$\Re[c]$ は複素数 c の実部のみをとり出す演算を示す．

3.2.3 目的音声の基底、アクティベーションとエゴノイズのアクティベーションの初期化

López らは，事前推定した音源モデルのパワースペクトログラムを IVA に与えることで性能が向上することを示した [8]．これを参考にして，本節では MOSIE の推定音を Rank-1 MNMF に初期値として与える手法を示す．教師基底の変形後は，再び半教師ありの Rank-1 MNMF と MOSIE を適用し，音声成分とエゴノイズ成分のより高精度な分離推定結果（$\overline{\boldsymbol{S}}'$ 及び $\overline{\boldsymbol{N}}'$）を得る．これらの推定成分は，MOSIE による統計的な事前情報（時間方向のカイ分布及び複素ガウス分布モデル）に基づいて推定されており，各成分の真の信号に近いアクティベーションを保持していることが期待できる．そこで，次式のように、これらを NMF で分解し，エゴノイズのアクティベーション $\boldsymbol{V}_{\mathrm{n}}$ ($\in \mathbb{R}^{L_1\times T_1}$)，音声の基底 $\boldsymbol{T}_{\mathrm{s}}$ ($\in \mathbb{R}^{\Omega\times L_2}$)，音声のアクティベーション $\boldsymbol{V}_{\mathrm{s}}$ ($\in \mathbb{R}^{L_2\times T_1}$) を得る．$l = 1,\ldots,L_2$ は目的音声の基底のインデックスである．NMF での分解は以下の最小化問題を解 くことに帰着する．

$$\widetilde{\boldsymbol{T}_{\mathrm{s}}}, \widetilde{\boldsymbol{V}_{\mathrm{s}}} = \arg\min_{\boldsymbol{T}_{\mathrm{s}},\boldsymbol{V}_{\mathrm{s}}} D_{\mathrm{IS}}(\overline{\boldsymbol{S}}' \circ \overline{\boldsymbol{S}}' | \boldsymbol{T}_{\mathrm{s}}\boldsymbol{V}_{\mathrm{s}})) \tag{24}$$

$$\widetilde{\boldsymbol{V}_{\mathrm{n}}} = \arg\min_{\boldsymbol{V}_{\mathrm{n}}} D_{\mathrm{IS}}(\overline{\boldsymbol{N}}' \circ \overline{\boldsymbol{N}}' | \overline{\boldsymbol{T}}_{\mathrm{n}}\boldsymbol{V}_{\mathrm{n}})) \tag{25}$$

ただし $D_{\mathrm{IS}}(\cdot|\cdot)$ は板倉斎藤距離基準のダイバージェンスである．式（24）のコスト関数を次式に示す．

$$Q_{NMF} = \Sigma_{\omega,\tau} \left(\frac{d_{\omega,\tau}}{\Sigma_l t_{\omega,l} v_{\omega,\tau}} + \log \Sigma_l t_{\omega,l} v_{l,\tau} \right) \tag{26}$$

ここで，$d_{\omega,\tau}, t_{\omega,l}, v_{l,\tau}$ はそれぞれ $\overline{\boldsymbol{S}}' \circ \overline{\boldsymbol{S}}', \boldsymbol{T}_{\mathrm{s}}, \boldsymbol{V}_{\mathrm{s}}$ の成分である．$\boldsymbol{T}_{\mathrm{s}}, \boldsymbol{V}_{\mathrm{s}}$ に関しての更新式は次式のようになる．

$$t_{\omega,l} \leftarrow t_{\omega,l} \sqrt{\frac{\Sigma_\tau d_{\omega,\tau} v_{l,\tau} (\Sigma_{l'} t_{\omega,l'} v_{l',\tau})^{-2}}{\Sigma_\tau v_{l,\tau} (\Sigma_{l'} t_{\omega,l'} v_{l',\tau})^{-1}}} \tag{27}$$

$$v_{l,\tau} \leftarrow v_{l,\tau} \sqrt{\frac{\Sigma_\omega d_{\omega,\tau} v_{\omega,l} (\Sigma_{l'} t_{\omega,l'} v_{l',\tau})^{-2}}{\Sigma_\omega t_{\omega,l} (\Sigma_{l'} t_{\omega,l'} v_{l',\tau})^{-1}}} \tag{28}$$

式（25）についても同様である．

エゴノイズの変形済み教師基底 $\overline{\boldsymbol{T}}_{\mathrm{n}}$ のみを固定し，その他の音源モデル（$\boldsymbol{V}_{\mathrm{n}}$，$\boldsymbol{T}_{\mathrm{s}}$，$\boldsymbol{V}_{\mathrm{s}}$）は Rank-1 MNMF

の音源モデルの要素であるエゴノイズのアクティベーション，目的音声の基底，目的音声のアクティベーションに対応する部分に初期値として与えて全教師ありの Rank-1 MNMF を実行する．その後，MOSIE を適用し，最終出力（音声成分）$\overline{S}''$ を得る．

4　分離性能の評価と比較

4.1　実験条件

提案手法の有効性を確認する為に，シミュレーションによる分離実験を行う．本実験では，単一チャネルの罰則条件付き半教師あり NMF（penalized semi-supervised NMF: PSNMF）[9]，IVA，従来の（ブラインド手法の）Rank-1 MNMF，及び提案手法の 4 つを比較する．観測信号は，エゴノイズと音声の混合信号を模擬的に作成した．音声に関しては，混合系のインパルス応答を畳み込むことで，多チャネルの観測信号を作成し，エゴノイズに関しては，実際に柔軟索状ロボットを駆動させた際の多チャネルの観測雑音をそのまま用いた．この両多チャネル信号をチャネル毎に足し合わせたものを混合信号としている．また，事前学習に用いるエゴノイズの信号は，適当な 1 チャネルの観測信号のみを用いている．参考として，実際の混合信号に含まれる真のエゴノイズを学習に用いる場合（matched）と，異なる時間区間でのエゴノイズの録音信号を学習に用いる場合（mismatched）の 2 つの評価を行った．従って，mismatched の場合は学習されたエゴノイズ基底と混合信号に含まれるエゴノイズのスペクトルに差異が生じており，教師基底の変形が必要となる．さらに，短時間フーリエ変換の窓関数はハミング窓を用いた．窓長は 512 ms，シフト長は 128 ms を用いた．

実験対象となる柔軟索状ロボットのマイクロホン数は $M = 8$ である．音源数は音声とエゴノイズの 2 つであるが，ロボットの姿勢等の影響で多種多様なエゴノイズが観測されることを考慮し，音声を 1 音源，エゴノイズを残りの 7 音源として仮定した上で，8 音源分離の Rank-1 MNMF を適用している．提案手法は半教師ありアプローチであるため，推定された信号の内，どの成分がエゴノイズであるかは自動的に判別することができる．エゴノイズに対応する 7 つの分離推定成分を全て足し合わせたものを，エゴノイズの推定音とする．比較手法のブラインドな Rank-1 MNMF では，8 つの分離推定音を実際に聞き比べ，手動で音声とエゴノイズにクラスタリングした上で足し合わせて 2 音源の分離推定音を得ている．また比較手法の PSNMF は唯一，単一チャネル（モノラル）信号にのみ適用可能な手法である為，適当な 1 チャネルのみの観測信号に適用する．その他，エゴノイズ用の基底数は 7 つの成分に対して 29 本，音声用の基底数は 11 本に設定した．分離精度の客観評価指標として signal-to-distortion ratio（SDR）[10] の改善量を算出した．

4.2　実験結果

Figures 4 は，異なる初期値について 10 回試行を行った際の各手法の音声信号に対する平均 SDR 改善量を示している．本図をみると，IVA と PSMNF では柔軟索状ロボットによる録音信号に対して十分な音声強調が達成できていないことが確認できる．また，音声及びエゴノイズの音源モデルをブラインドに推定する従来の Rank-1 MNMF は，IVA 及び PSNMF よりも高精度であるが，それでも 4 dB 程度しか改善されていない．一方で提案手法は，事前学習時のエゴノイズが混合信号中のものと異なる場合（mismatched）においても，従来のどの手法よりも高精度な分離を達成している．参考として示した matched の精度には 2 dB 程度及ばなかったものの，確実に効果的な音声強調を実現している．

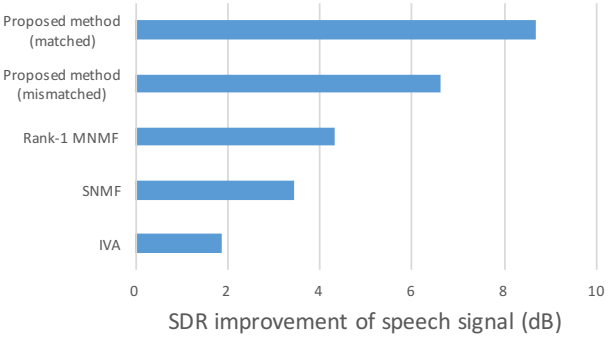

Fig. 4　Comparison of separation performance for each method.

5　おわりに

柔軟索上ロボットにおける音声強調として，事前学習及び教師基底の変形に基づく多チャネル音源分離手法を提案した．提案手法は，従来の音源分離手法よりも良い音源分離性能を示すことを実験的に確認した．

謝辞　本研究は，総合科学技術・イノベーション会議により制度設計された革新的研究開発推進プログラム（ImPACT）により，科学技術振興機構を通して委託されたものである．実験データを提供して頂いた早稲田大学奥乃博教授と京都大学坂東宜昭氏に感謝の意を表する．

References

[1] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” *Proc. Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[2] D. Kitamura, N. Ono, H. Sawada, H. Kameoka and H. Saruwatari, “Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model,” *Proc. ICASSP*, pp. 276–280, 2015.

[3] C. Breithaupt and R. Martin, “Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions,” *IEEE Trans. ASLP*, vol. 19, no. 2, pp.277–289, 2010.

[4] 中嶋広明, 北村大地, 高宗典玄, 小山翔一, 猿渡洋, 小野順貴, 高橋祐, 近藤多伸, “全極モデルを用いた基底変形型教師あり NMF による音楽信号分離,” 日本音響学会 音講論秋, pp. 573–576, 2015.

[5] T. Kim, H. T. Attias, S.-Y. Lee and T.-W. Lee, “Blind source separation exploiting higher-order frequency dependencies,” *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2007.

[6] H. Sawada, H. Kameoka, S. Araki and N. Ueda, “Multichannel extensions of non-negative matrix factorization with complex-valued data,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[7] Y. Murota, D. Kitamura, S. Nakai, H. Saruwatari, S. Nakamura, K. Shikano, Y. Takahashi, K. Kondo, “Music signal separation based on bayesian spectral amplitude estimator with automatic target prior adaptation,” *Proc. ICASSP*, pp. 7490–7494, 2014.

[8] A. R. López, N. Ono. U. Remes, K. Palomäki and M. Kurimo, “Designing multichannel source separation based on single-channel source separation,” *Proc. ICASSP*, pp. 469–473, 2015.

[9] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, K. Kondo, “Music signal separation based on supervised non-negative matrix factorization with orthogonality and maximum-divergence penalties,” *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol. E97-A, no. 5, pp. 1113–1118, 2014.

[10] E. Vincent, R. Gribonval, C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.