マルチチャネル非負値行列因子分解の初期値設定における空間相関行列推定法 *

☆三浦伊織 (大分大), 太刀岡勇気, 成田知宏 (三菱電機), 上ノ原進吾, 古家賢一 (大分大)

1 はじめに

非負値行列因子分解 (Nonnegative Matrix Factorization: NMF)[1] とは非負値の行列を分解し、解析を行う手法である。行列表現できるデータならば分解可能であるため、音や画像、文書など多種多様なものに利用できる。音響分野ではマルチチャネル拡張によって空間情報を活用することで音源分離を行う手法が提案されている [2, 3]。しかし、従来のマルチチャネル NMF (MNMF) は自由度の高いモデルであるため、多くの局所最適解が存在し、分離性能に対する初期値依存性が課題となっている [4]。

本稿は、初期値依存性の大きい空間相関行列に着目し、あらかじめバイナリマスクで分離したデータから計算した値を、空間相関行列の初期値に設定することで、分離性能を向上させることを目的とする。騒音環境下音声認識実験 (CHiME4) により、提案法の有効性を示す。

2 MNMF

2.1 概要

MNMF[2, 3] とは、NMF をマルチチャネル拡張したものであり、観測行列 $\mathbf{X}$ を 4 つの行列 $\mathbf{H}$、$\mathbf{Z}$、$\mathbf{T}$、$\mathbf{V}$ に分解する。MNMF では空間情報を用いてスペクトル基底を L 個の音源にクラスタリングすることで事前の学習なしで音源分離を実現する。位相情報を扱うために複素数を用いるので、複素数における非負性に対応するものとして、エルミート半正定値行列を用いる [2]。

2.2 定式化

M をマイクロホン数として入力ベクトルを $\tilde{\mathbf{x}} = [\tilde{x}_1, \cdots, \tilde{x}_M]^\top$ とする。ただし、$^\top$ は転置を表す。$\tilde{x}_m$ は m 番目のマイクロホンでの Short Time Fourier Transform (STFT) の複素係数であり、スペクトログラムを指す。周波数 i $(1 \leq i \leq I)$, 時間 j $(1 \leq j \leq J)$ のとき $\tilde{\mathbf{x}}_{ij}$ で表すと行列 $\mathbf{X}$ の i、j 成分を X_{ij} とし、$\mathsf{X}_{ij} = \tilde{\mathbf{x}}_{ij}\tilde{\mathbf{x}}_{ij}^H$ もしくは i、j それぞれについて

$$\mathsf{X}_{ij} = \tilde{\mathbf{x}}_{ij}\tilde{\mathbf{x}}_{ij}^H = \begin{bmatrix} |\tilde{x}_1|^2 & \cdots & \tilde{x}_1\tilde{x}_M^* \\ \vdots & \ddots & \vdots \\ \tilde{x}_M\tilde{x}_1^* & \cdots & |\tilde{x}_M|^2 \end{bmatrix} \quad (1)$$

で表される。ただし、^{H} はエルミート転置を表す。すなわち、I 行 J 列の行列 $\mathbf{X}$ はそれぞれの要素が $M \times M$ の複素行列を持つ階層的なエルミート半正定値行列となる。この行列 $\mathbf{X}$ を MNMF で分解すると、式 (2) で表されるように、K 個の基底から成る基底行列

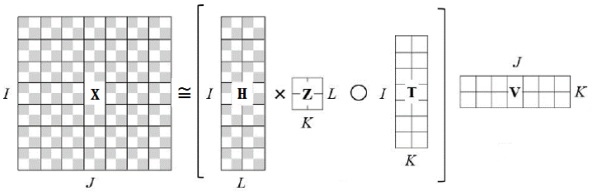

Fig. 1 MNMF で分解された行列の例

$\mathbf{T}(\in \mathbb{R}^{I\times K})$、アクティベーション行列 $\mathbf{V}(\in \mathbb{R}^{K\times J})$、音源の空間情報を示す空間相関行列 $\mathbf{H}$ と音源の空間情報と各基底を関連付ける潜在変数行列 $\mathbf{Z}(\in \mathbb{R}^{L\times K})$ という 4 つの行列に分解できる。

$$\mathbf{X} = (\mathbf{HZ} \circ \mathbf{T})\mathbf{V} \quad (2)$$

ただし、$\circ$ はアダマール積を表す。行列 $\mathbf{H}$ は行列 $\mathbf{X}$ と同様にそれぞれの要素が $M \times M$ の複素行列を持つ I 行 L 列の階層的なエルミート半正定値行列である。図 1 は式 (2) を図式化したものである。このとき、右辺は

$$\hat{\mathsf{X}}_{ij} = \sum_{k=1}^{K}\left(\sum_{l=1}^{L}\mathsf{H}_{il}z_{lk}\right)t_{ik}v_{kj} \quad (3)$$

と表すことができ、理想的には行列 X と $\hat{\mathsf{X}}_{ij}$ を要素に持つ行列 $\hat{\mathsf{X}}$ は等しくなる。しかし、一般的には誤差が生じるため、MNMF では行列 X と行列 $\hat{\mathsf{X}}$ との距離 $D_*(\mathsf{X}, \hat{\mathsf{X}})$ を定義し、この距離を最小化する行列 $\mathbf{H}$、$\mathbf{Z}$、$\mathbf{T}$、$\mathbf{V}$ を求める。今回はダイナミックレンジが大きい音楽や音声に適している Itakura-Saito (IS) divergence[5] を用いて以下のように定義する。

$$D_{IS}(\mathsf{X}_{ij}, \hat{\mathsf{X}}_{ij}) = tr(\mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1}) - \log\det\mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1} - M \quad (4)$$

ただし、$tr\,(\cdot)$ は対角要素の和を表している。

2.3 行列分解アルゴリズム

Multiplicative update rule[6] と呼ばれる反復アリゴリズムを、ランダムな非負の値で初期化した行列 $\mathbf{T}$、$\mathbf{V}$、$\mathbf{Z}$ ならびに各要素へ単位行列を持たせた行列 $\mathbf{H}$ に繰り返し適用する。IS divergence を用いた更新式は以下のようになる。

$$t_{ik} \leftarrow t_{ik}\sqrt{\frac{\sum_l z_{lk}\sum_j v_{kj}tr(\hat{\mathsf{X}}_{ij}^{-1}\mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1}\mathsf{H}_{il})}{\sum_l z_{lk}\sum_j v_{kj}tr(\hat{\mathsf{X}}_{ij}^{-1}\mathsf{H}_{il})}} \quad (5)$$

$$v_{kj} \leftarrow v_{kj}\sqrt{\frac{\sum_l z_{lk}\sum_i t_{ik}tr(\hat{\mathsf{X}}_{ij}^{-1}\mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1}\mathsf{H}_{il})}{\sum_l z_{lk}\sum_i t_{ik}tr(\hat{\mathsf{X}}_{ij}^{-1}\mathsf{H}_{il})}} \quad (6)$$

*Spatial Correlation Matrix Estimation Method in Initial Value Setting of Multi-channel Non-negative Matrix Factorization. by Iori Miura (Oita University), Yuuki Tachioka, Tomohiro Narita (Mitsubishi Electric), Shingo Uenohara, and Ken'ichi Furuya (Oita University)

$$z_{lk} \leftarrow z_{lk}\sqrt{\frac{\sum_{i,j} t_{ik}v_{kj}tr(\hat{\mathsf{X}}_{ij}^{-1}\mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1}\mathsf{H}_{il})}{\sum_{i,j} t_{ik}v_{kj}tr(\hat{\mathsf{X}}_{ij}^{-1}\mathsf{H}_{il})}} \tag{7}$$

H_{il} については次式の

$$A = \sum_k z_{lk}t_{ik}\sum_j v_{kj}\hat{\mathsf{X}}_{ij}^{-1} \tag{8}$$

$$B = \mathsf{H}'_{il}\left(\sum_k z_{lk}t_{ik}\sum_j v_{kj}\hat{\mathsf{X}}_{ij}^{-1}\mathsf{X}_{ij}\mathsf{X}_{ij}^{-1}\right)\mathsf{H}'_{il} \tag{9}$$

A、B を係数に持つ代数リッカチ方程式で求めることができる。

$$\mathsf{H}_{il}A\mathsf{H}_{il} = B \tag{10}$$

ただし、H'_{il} は更新前の行列 H_{il} を表しており、解き方は文献 [2] に示されている。

2.4 正規化

行列 $\mathbf{H}$ は式 (2) の一意性を保つため、行列 $\mathbf{Z}$ は確率の定義からの要請によるため、正規化を行わなければならない。正規化は以下の式で行った。

$$\mathsf{H}_{il} = \frac{\mathsf{H}_{il}}{tr(\mathsf{H}_{il})}, \quad z_{lk} = \frac{z_{lk}}{\sum_l z_{lk}} \tag{11}$$

2.5 音源分離

音源分離を行うために次式で表されるウィナーフィルタを用いる。

$$Y = \frac{\hat{S}}{\hat{S}+N}X \tag{12}$$

ただし、Y は目的信号、$\hat{S}$ は目的信号の推定値、N は雑音信号、X は雑音信号を含んだ目的信号を示す。$\bar{y}_{ij}^{(l)}$ を分離後の音源としたとき、$Y = \tilde{\mathbf{y}}_{ij}^{(l)}$、$\hat{S} = (\sum_{k=1}^{K} z_{lk}t_{ik}v_{kj})\ \mathsf{H}_{il}$、$\hat{S}+N = \hat{\mathsf{X}}_{ij}$、$X = \tilde{\mathbf{x}}_{ij}$ を代入すると、次式のマルチチャネルウィナーフィルタとなり、各音源に対応した分離信号を得られる [2]。

$$\tilde{\mathbf{y}}_{ij}^{(l)} = \left(\sum_{k=1}^{K} z_{lk}t_{ik}v_{kj}\right)\mathsf{H}_{il}\hat{\mathsf{X}}_{ij}^{-1}\tilde{\mathbf{x}}_{ij} \tag{13}$$

2.6 MNMF の課題

MNMF は自由度の高いモデルであるため、局所最適解が増え、初期値依存による分離性能のばらつきが問題となる。図 2 は MNMF アルゴリズムにランダムな初期値を 10 回与えて音源分離を行った際の分離性能 (SDR[3]) を示している [4]。この図から、分離性能は初期値ごとに大きく異なっていることがわかる。

3 空間相関行列推定法

クロススペクトル法 [7] を用いて、インパルス応答から空間相関行列 $\mathbf{H}$ の初期値を計算することで、分離性能が向上することが分かっている [4]。しかし、多くの応用において、事前にインパルス応答を取得することは困難である。そこで、バイナリマスク [8] を用いて取得したデータから、空間相関行列 $\mathbf{H}$ の初期値を求めることで、MNMF の分離性能を向上させる。

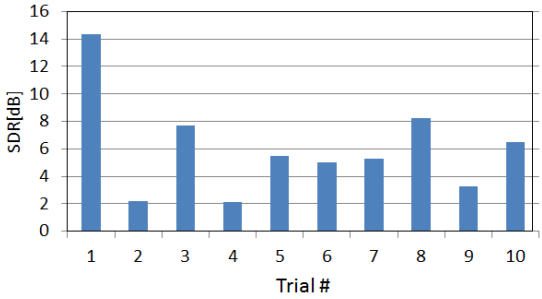

Fig. 2 音源分離性能の初期値依存性

本稿では、2 種類の空間相関行列推定法を提案する。第 1 の方法で、バイナリマスクで求めた分離信号から、クロススペクトル法を用いて空間相関行列 $\mathbf{H}$ の初期値を計算する。第 2 の方法で、バイナリマスクで生成した目的音と雑音のマスクを初期値として、中谷らの EM アルゴリズムに基づいたマスク強調 [9] により、マスクの精度を向上させる。精度が向上したマスクから目的音のステアリングベクトル (SV) を推定し、空間相関行列 $\mathbf{H}$ を計算する。

3.1 バイナリマスク

バイナリマスク [8] とは、各音源の到来時間差に基づいて時間周波数上でマスキングを行い、音源分離を行う手法である。例えば、目的音源が正面方向である場合、マイク間の位相差は 0 である。雑音が 0 度方向から到来する場合、位相差は大きくなるので、マイク間の位相差がゼロから離れた時間周波数ビンのパワーをマスキングすれば、目的音源を強調することができる。マスク M は以下のように閾値を用いて設定される。

$$M_{i,j} = \begin{cases} \epsilon & \text{if } |\theta_{i,j}| > \theta_{\mathrm{c}}, \\ 1 & \text{if } |\theta_{i,j}| \leq \theta_{\mathrm{c}}, \end{cases}$$

ϵ は十分小さい定数、$\theta_{i,j}$ は時間周波数ビンの位相差、θ_c は事前に定めておく閾値である。事前に音源方向が分かっていれば、それぞれの音源が強調されるようにマスキングすることができる。

3.2 クロススペクトル法

音源データのスペクトルをフーリエ変換することで

$$A_i = \begin{bmatrix} a_{i,1} & \dots & a_{i,M} \end{bmatrix}^{\top} \tag{14}$$

M 行 1 列の A_i が与えられる。A_i と、そのエルミート転置 (1 行 M 列) の積

$$\mathsf{H}_i = A_iA_i^H = \begin{bmatrix} |a_{1,1}|^2 & \cdots & a_{i,1}a_{i,M}^* \\ \vdots & \ddots & \vdots \\ a_{i,M}a_{i,1}^* & \cdots & |a_{i,M}|^2 \end{bmatrix} \tag{15}$$

は周波数ビン i における空間相関を表す。L 個の各音源から H_{i} を作成することで、MNMF における I 行 L 列の空間相関行列 $\mathbf{H}$ として設定出来る [7]。本稿では、各マイクロホンのスペクトル成分を要素に持つ M 行 1 列の行列とそのエルミート転置の積から空間相関行列 $\mathbf{H}$ を算出する手法をクロススペクトル法と呼ぶ。ここでは、データの全区間から空間相関行列

$\mathbf{H}$ を計算できるように、フレームサイズおよびシフトサイズを 1024 として、STFT を行う。各フレームからクロススペクトル法で空間相関行列 $\mathbf{H}$ を計算し、全フレームの空間相関行列 $\mathbf{H}$ の平均の値を MNMF の初期値とした。

3.3 EM アルゴリズムに基づくマスク強調

マスク強調には複素 GMM (CGMM)[10] によるクラスタリングを用いる。CGMM において、観測信号ベクトル $\tilde{\mathbf{x}}_{ij}$ は複素ガウス分布を要素に持つ以下の混合分布でモデル化される。

$$P(\tilde{\mathbf{x}}_{ij};\theta) = \sum_n w_i^{(n)} N_c(\tilde{\mathbf{x}}_{ij}; 0, \sigma_{i,j}^{(n)} \mathbf{B}_i^{(n)}) \quad (16)$$

ここで、n は雑音クラス ($n = v$) と音声＋雑音クラス ($n = x + v$) を区別するインデックス、$N_c(\tilde{\mathbf{x}}; \mu, \Sigma)$ は平均 μ、共分散行列 Σ の複素ガウス分布、$w_j^{(n)}$ は混合比を表す。各クラスの共分散行列は時変のスカラー値 $\sigma_{i,j}^{(n)}$ と時変の行列 $\mathbf{B}_i^{(n)}$ の積に分解できると仮定される。θ はモデルパラメータ全体の集合を表す。

周波数 i ごとに、全時刻 ($1 \leq j \leq J$) の観測信号ベクトル $\tilde{\mathbf{x}}_{ij}$ を用いて、EM アルゴリズムに基づき CGMM のモデルパラメータを最尤推定する。以下の E-step と M-step を繰り返し適用し、強調を行う。

1. E-step: M-step で得られた CGMM のモデルパラメータの推定値 (初期値はバイナリマスクによりクラスタリングしたクラスタから計算) に基づき、以下のように、各時間周波数点が各クラス n に属する事後確率を計算し、マスクの推定値 $\hat{M}_{i,j}^{(n)}$ として更新する。

$$\hat{M}_{i,j}^{(n)} = \frac{\hat{w}_i^{(n)} N_c(\tilde{\mathbf{x}}_{ij}; 0, \hat{\sigma}_{i,j}^{(n)} \hat{\mathbf{B}}_i^{(n)})}{\sum_{n'} \hat{w}_i^{(n')} N_c(\tilde{\mathbf{x}}_{ij}; 0, \hat{\sigma}_{i,j}^{(n')} \hat{\mathbf{B}}_i^{(n')})} \quad (17)$$

2. M-step: E-step で得られたマスクの推定値 $\hat{M}_{i,j}^{(n)}$ に基づき、CGMM のモデルパラメータの推定値を以下のように更新する。

$$\hat{\mathbf{B}}_i^{(n)} = \frac{1}{\sum_j \hat{M}_{i,j}^{(n)}} \sum_j \hat{M}_{i,j} \frac{\tilde{\mathbf{x}}_{ij} \tilde{\mathbf{x}}_{ij}^H}{\hat{\sigma}_{i,j}^{(n)}} \quad (18)$$

$$\hat{\sigma}_{i,j}^{(n)} = \frac{1}{M} \tilde{\mathbf{x}}_{ij}^H (\hat{\mathbf{B}}_i^{(n)})^{-1} \tilde{\mathbf{x}}_{ij} \quad (19)$$

$$\hat{w}_i^{(n)} = \frac{\sum_j \hat{M}_{i,j}^{(n)}}{J} \quad (20)$$

提案法では E-step で得られた $M_{i,j}$ を利用する。

3.4 マスクに基づく SV 推定

観測信号の空間相関行列 $\mathbf{R}_i^{(x+v)}$ と雑音の空間相関行列 $\mathbf{R}_i^{(v)}$ が既知の時、目的音声と雑音は無相関との仮定の下、目的音声の空間相関行列 $\mathbf{R}_i^{(x)}$ は以下のように求めることができる。

$$\mathbf{R}_i^{(x)} = \mathbf{R}_i^{(x+v)} - \mathbf{R}_i^{(v)} \quad (21)$$

また、マスクを用いて各空間相関行列は以下のように求められる。

Table 1　本実験における CHiME4 の実験条件

音声認識システム	Kaldi
語彙数	5000
目的音声の言語	英語
音響モデル	GMM

Table 2　本実験における MNMF の実験条件

サンプリング周波数	16kHz
フレームサイズ	1024
シフトサイズ	256
基底数	30
音源数	2
更新回数	500

$$\mathbf{R}_i^{(x+v)} = \frac{1}{J} \sum_{j=1}^{J} \tilde{\mathbf{x}}_{ij} \tilde{\mathbf{x}}_{ij}^H \quad (22)$$

$$\mathbf{R}_i^{(v)} = \frac{1}{\sum_{j=1}^{J} M_{i,j}^{(v)}} \sum_{j=1}^{J} M_{i,j}^{(v)} \tilde{\mathbf{x}}_{ij} \tilde{\mathbf{x}}_{ij}^H \quad (23)$$

ここで、$M_{i,j}^{(v)}$ は各時間周波数点が雑音に属するかどうかを示すマスクである。音声信号の空間相関行列 $\mathbf{R}_i^{(x)}$ を求めることで、SV はその第一固有ベクトルとして近似的に求めることができる。本稿では、求めた SV から空間相関行列を計算することで、MNMF の初期値として設定する。

4 CHiME4 による音声認識実験

騒音環境下音声認識のタスクである第 4 回 CHiME Challenge (CHiME 4) の 2ch トラック [11] を対象として、提案法により音声認識性能が改善するか評価を行う。

4.1 実験条件

CHiME 4 は 4 つの環境 (バス (BUS)、カフェ (CAF)、市街地 (PED)、交差点 (STR)) において、タブレットに取り付けられた最大 6 つのマイクを用いて録音された音声を認識するタスクである。性能は音声認識性能を計る指標 (単語誤り率 (WER)) で評価する。2ch トラックでは、6 マイクの内から無作為に選択された 2 つのマイクで録音されたデータを使用する。また、目的音は学習セット、開発セット (dt)、評価セット (et) の 3 種類が用意されており、実環境での録音データ (REAL) と仮想環境で疑似的に作成されたデータ (SIMU) が存在する。ここでは、以下の手法を比較する。

1. 未処理のまま音声認識 (Noisy)
2. 重み付き遅延和アレーにより強調 (Baseline)
3. ランダムな初期値による MNMF(Random)
4. バイナリマスクで分離したデータから空間相関行列を計算し、MNMF の初期値に設定 (BinMask)
5. EM アルゴリズムで強調したマスクを用いて推定した SV から、空間相関行列を計算し、MNMF の初期値に設定 (EM)

Table 3　CHiME4 の dt の REAL における WER[%]

-	BUS	CAF	PED	STR	AVE
Noisy	27.3	23.1	16.3	22.0	22.2
Baseline	20.1	16.3	12.4	16.9	16.2
Random	33.5	30.7	24.6	27.4	29.0
BinMask	**17.4**	**16.0**	12.0	16.1	**15.4**
EM	19.1	16.8	**11.8**	**15.5**	15.8

Table 4　CHiME4 の dt の SIMU における WER[%]

-	BUS	CAF	PED	STR	AVE
Noisy	20.4	29.8	20.5	27.3	24.5
Baseline	16.1	23.6	15.5	21.4	19.2
Random	23.0	32.1	25.2	29.9	27.5
BinMask	**13.0**	**19.0**	**14.1**	**18.6**	**16.1**
EM	13.9	19.8	14.4	18.7	16.7

ただし、Random に関しては、得られた 2 つの分離音のうち、どちらに音源が含まれるかは自明ではないため、双方を音声認識させて、WER の良い方を選択した値である。行列 **H** 以外の行列の初期値は各手法で共通である。また、表 1 に本実験における CHiME4 の条件を、表 2 に本実験における MNMF の条件を示す。

4.2　実験結果

表 3 に dt の REAL における実験結果を、表 4 に dt の SIMU における実験結果を、表 5 に et の REAL における実験結果を、表 6 に et の SIMU における実験結果を示す。また、AVE は 4 環境の WER を平均した値を表す。この結果から、Noisy や Baseline、Random と比べて、BinMask や EM による空間相関行列推定法の音声認識性能が高い (WER が低い) ことが分かる。しかし、BinMask と EM を比較すると、WER の差が小さいことが分かる。

4.3　考察

実験結果から、雑音下の音声認識に対して、MNMF の初期値設定における空間相関行列推定法が有効であることを確認した。ただし、本実験で設定した基底数や音源数などが最適な値ではない可能性が考えられるので、環境に対応したパラメータの選択方法などを検討する必要がある。

EM アルゴリズムによるマスク強調の効果が小さかった原因としては、文献 [9] では 6 チャネルの音声データに適用していたのに対し、本稿での実験では、2 チャネルの音声データを対象としている点が考えられる。EM アルゴリズムによるマスク強調の効果はチャネル数が少ないデータに対して効果が低いと考えられるので、3 チャネル以上のデータへの適用を今後行う必要がある。

5　まとめ

本稿では、バイナリマスクと EM アルゴリズムを用いて推定した SV から、空間相関行列の初期値を計算する手法を提案し、音声認識における実験を行った。実験結果から、バイナリマスクだけから計算した場合と比べて、EM アルゴリズムを組み合わせた手法

Table 5　CHiME4 の et の REAL における WER[%]

-	BUS	CAF	PED	STR	AVE
Noisy	51.9	39.7	34.0	24.5	37.5
Baseline	39.4	28.4	27.6	20.8	29.0
Random	56.8	44.6	38.6	31.1	42.8
BinMask	**37.7**	**26.0**	**21.2**	19.2	**26.0**
EM	39.8	27.0	23.5	**18.9**	27.3

Table 6　CHiME4 の et の SIMU における WER[%]

-	BUS	CAF	PED	STR	AVE
Noisy	26.7	38.4	34.7	33.5	33.3
Baseline	20.2	31.8	30.0	28.4	27.6
Random	23.4	30.9	31.8	32.7	29.7
BinMask	**15.3**	**23.6**	**22.8**	**23.9**	**21.4**
EM	16.9	28.5	24.6	24.4	23.6

の性能が変わらなかったことが分かった。チャネル数が少ないデータに対して効果が低いと考えられるので、チャネル数が多いデータから空間相関行列を計算することを今後検討する。

参考文献

[1] D.D. Lee *et al.*, “Learning the parts of objects with nonnegative matrix factorization,” Nature, vol. 401, pp. 788-791,1999.

[2] H. Sawada *et al.*, “Multichannel Extensions of Non-Negative Matrix Factorization With Complex-Valued Data,” IEEE Trans. ASLP, vol.21, no.5, pp. 971-982, 2013.

[3] E. Vincent *et al.*, “First stereo audio source separation evaluation campaigh: Data algprithm and results,” Independent Component Analysis and Signal Separation(Springer, Bearlin, 2007), pp. 552-559.

[4] 三浦 伊織, 他: “ マルチチャネル非負値行列因子分解における初期値依存性の挙動解析” 日本音響学会講演論文集, pp. 669-672, 2016 春.

[5] C. Fvotte *et al.*, “Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis,” Neural Comput., vol. 21, no. 3, pp. 793-830, 2009.

[6] M. Nakano *et al.*, “Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence,” In Proc. MLSP 2010, pp. 283-288, 2010.

[7] 北村 大地, 他: “ランク 1 空間モデルを用いた効率的な多チャネル非負値行列因子分解” 日本音響学会講演論文集, pp. 579-582, 2014 秋.

[8] H Sawada *et al.*, “Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment” IEEE Trans. Audio, Speech, Language Process., vol. 19, pp. 516-527, Mar. 2011.

[9] 中谷 智広, 他: “NTT CHiME-3 音声認識システム：耐雑音フロントエンド” 日本音響学会講演論文集, pp. 57-60, 2016 春.

[10] N. Ito, *et al.*, “Relaxed disjointness based clustering for joint blind source separation and dereverberation,” in Proc. IWAENC, pp. 268-272, 2014.

[11] J. Barker *et al.*, “An Analysis of Environment, Microphone and Data Simulation Mismatches in Robust Speech Recognition,” in Proc. of the 4th Intl. Workshop on Speech Processing in Everyday Envionments (CHiME 2016), 2016