一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
EA2017-107,SIP2017-116,SP2017-90 (2018-03)

マルチチャネル非負値行列因子分解に基づく ビームフォーミングを用いた雑音環境下音声認識

島田　一希† 　坂東　宜昭† 　三村　正人†

糸山　克寿† 　吉井　和佳† 　河原　達也†

† 京都大学　大学院情報学研究科

E-mail: †{shimada,bando,mimura,itoyama,yoshii,kawahara}@sap.ist.i.kyoto-u.ac.jp

あらまし　本稿では，雑音に頑健な音声認識のためのマルチチャネル音声強調について述べる．音声認識のための音声強調では，マイクロホンアレイにより観測したマルチチャネル信号から，目的音声方向の信号を強調し雑音方向の信号を除去するビームフォーミングが効果的である．ビームフォーミングを行うために必要な目的音声の方向を表すステアリングベクトルや雑音の空間相関行列の推定については，時間周波数マスクに基づく手法が活発に研究されており，時間周波数ビンを目的音声と雑音に分類するディープニューラルネットワーク (DNN) を用いてマスクを推定する手法が高い性能を示すことが知られている．このような事前の教師あり学習による手法は，未知環境において性能が低下するおそれがある．そこで本研究では，教師なしで空間相関行列を正確に推定するために，マルチチャネル非負値行列因子分解 (MNMF) に基づくブラインド音源分離を用いて，観測の各時間周波数ビンを目的音声とそれ以外の音源 (雑音) に分解する手法を提案する．本研究では MNMF をオンライン処理に拡張し，音声強調に適した初期化を行う．MNMF に適したビームフォーミングを明らかにするために，最小分散無歪 (MVDR) ビームフォーミング及びマルチチャネルウィーナーフィルタリング (MWF) において，時変及び時不変フィルタの両方を比較した．実録音データに対する音声認識実験を行い，提案法が未知環境において DNN マスクに基づくビームフォーミングと比べて頑健に動作することを示した．

キーワード　雑音環境下音声認識，音声強調，ビームフォーミング，マルチチャネル非負値行列因子分解

1. はじめに

雑音に頑健な音声認識のために，ビームフォーミングを用いたマルチチャネル音声強調が活発に研究されている．ビームフォーミングは，目的音声方向の信号を強調し雑音方向の信号を除去するものである [1]．CHiME Challenge など近年の国際技術評価会において，ビームフォーミングが雑音環境下における音声認識の前処理として効果的であることが示されている [2]．最小分散無歪 (MVDR) ビームフォーミング [1] をはじめ，一般化サイドローブキャンセリング (GSC) [3]，マルチチャネルウィーナーフィルタリング (MWF) [4,5]，SN 比最大化ビームフォーミング (MaxSNR) [6] といったビームフォーミングが提案されている．時間周波数領域においてこれらのビームフォーミングを用いるには，ステアリングベクトル及び空間相関行列から線形フィルタを計算することが必要である [7–12]．

ステアリングベクトル及び空間相関行列の推定に関して多くの研究が行われてきた．従来の信号の相互相関を用いる GCC-PHAT 法に基づいてステアリングベクトルを推定したビームフォーミング [13] は，実環境の音声認識において十分な性能を得ることができない [2]．近年，時間周波数マスクに基づく手法が注目を集めている [7–12]．この手法は観測信号スペクトログラムの各時間周波数ビンが目的音声と雑音に排他的に分類されるという仮定に基づく [7–12]．目的音声及び雑音の空間相関行列は，分類した時間周波数ビンから計算される [7–12]．目的音声のステアリングベクトルは，空間相関行列の第一固有ベクトルとして近似される [7–9]．この分類手法として，例えば複素混合ガウス分布 (CGMM) に基づく教師なし手法がある [7]．一方で近年，最もよく使われている分類手法として，ディープニューラルネットワーク (DNN) で時間周波数マスクを推定する手法がある [8–12]．DNN を教師ありで学習する際には，入力となるスペクトログラム及び出力となる理想的なバイナリマスク (IBM) のペアが大量に必要となる．

様々なマルチチャネル音響信号処理において位相情報は重要な役割を担っているが，これらのマスク推定では位相を扱うことが難しい．また事前学習を行うマスク推定では，分類器が訓練データに対して過学習を起こし，そのデータでカバーできない未知環境において音声認識性能が低下するおそれがある．この問題は，種々の雑音を含むマルチコンディションデータを使った DNN の学習により緩和される [14]．しかし，マイクロホンの種類や収音環境が異なる場合にも頑健であるか議論の余地がある．

これに対して我々は，マルチチャネル非負値行列因子分解 (MNMF) によるブラインド音源分離 [15,16] に基づいて推定した空間相関行列を用いてビームフォーミングを行う音声強

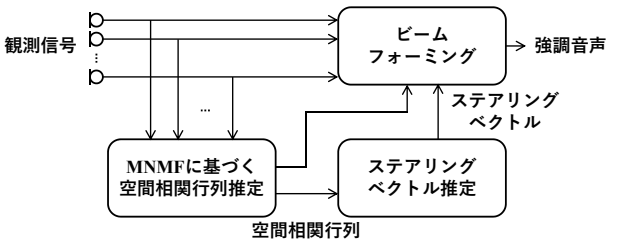

図 1　MNMF に基づく空間相関行列推定とビームフォーミングを統合した提案する音声強調の枠組み

調を提案している [17,18]．図 1 のように，提案する音声強調ではまず観測信号から MNMF に基づいて空間相関行列を推定する．次に空間相関行列からステアリングベクトルを推定する．そして推定したステアリングベクトル及び空間相関行列から線形フィルタを計算し，ビームフォーミングにより観測信号から強調音声を出力する．MNMF により，マイクロホンアレイで観測した混合音の複素スペクトログラムから，各音源の空間相関行列を推定する．ここで各音源のパワー成分については基底行列 (スペクトルテンプレートの集合) とアクティベーション行列 (各テンプレートの時間的な強度の集合) の積で近似して表される．MNMF はそれぞれの時間周波数ビンにおいて位相情報を扱い，観測信号を各音源 (目的音声と雑音) の信号へと分解している．DNN や混合ガウス分布を用いたマスク推定に基づく手法は因子分解を行わずに空間相関行列を推定しているのに対して [7–12]，提案法ではより正確な空間相関行列推定が可能になると考えられる．提案する音声強調は事前学習を必要とせず，適切な訓練データがない環境でも頑健に機能することが期待される．マスクに基づく手法以外のブラインド音源分離に基づく音声強調手法の研究は未だ少ない [19,20]．本稿では，MNMF に基づく空間相関行列推定とビームフォーミングを統合し，その強調性能を異なる 2 つの音声認識タスクで評価した．バッチ処理の MNMF に加えて，オンライン処理の MNMF を導入し，適切な初期化を行う．ビームフォーミングとして，MVDR ビームフォーミング [1]，ランク 1 MWF ビームフォーミング及びフルランク MWF [4,5] の時変フィルタと時不変フィルタを比較した．

2. ビームフォーミング

本研究では, 3 つのビームフォーミング手法, すなわち MVDR ビームフォーミング，ランク 1 MWF ビームフォーミング及びフルランク MWF を用いる．これらのビームフォーミング手法を，各音源の伝搬過程とフィルタの推定法という 2 つの観点から整理する．まず各音源からマイクロホンアレイへの 2 種類の伝搬過程を定義する．一つ目は，各音源の伝搬に関して単一の音源とステアリングベクトルでモデル化するものであり，ここではランク 1 伝搬過程と呼び，単一の直接波を中心として表現する．二つ目は，フルランクの空間相関行列でモデル化するもので，より複雑な伝搬過程を表現できる．ここではフルランク伝搬過程と呼ぶ．次にビームフォーミングフィルタを得るための 2 つの推定法，MAP 推定と最尤推定について述べる．

表 1　各ビームフォーミング手法の関係

推定法 ＼ 伝搬過程	目的音声 | ランク 1 雑音 | フルランク	目的音声 | フルランク 雑音 | フルランク
最尤推定 目的音声のスケールを考慮しない	MVDR 式 (3) & (4)	- -
MAP 推定 目的音声のスケールを考慮する	ランク 1 MWF 式 (5) & (6)	フルランク MWF 式 (8) & (9)

表 2　目的音声と雑音の信号，ステアリングベクトル及び空間相関行列に関する表記法

種類	目的音声	雑音
信号	$\boldsymbol{s}$	$\boldsymbol{n}$
ステアリングベクトル	$\boldsymbol{p}$	-
空間相関行列	$\boldsymbol{P}$	$\boldsymbol{Q}$

MAP 推定では目的音声のスケールをガウス分布で表現した上で推定に用いる．一方，最尤推定ではそのスケールに関して考慮しない．表 1 にこれらのビームフォーミング手法の関係をまとめる．

ビームフォーミングは短時間フーリエ変換領域で行われる．$\boldsymbol{x}_{ft} \in \mathbb{C}^M$ を時間フレーム t，周波数ビン f における M ch マイクロホンアレイ観測信号とする．このマルチチャネル観測信号に線形フィルタ $\boldsymbol{w}_{ft} \in \mathbb{C}^M$ をかけることで強調音声 $y_{ft} \in \mathbb{C}$ を出力する．

$$y_{ft} = \boldsymbol{w}_{ft}^H \boldsymbol{x}_{ft} \tag{1}$$

各ビームフォーミング手法を説明する前に信号，ステアリングベクトル及び空間相関行列の表記法を表 2 に示す．

2.1 MVDR ビームフォーミング

MVDR ビームフォーミング [1] では，M ch マイクロホン観測信号 $\boldsymbol{x}_{ft}$ について，次のように仮定する．

$$\boldsymbol{x}_{ft} = \boldsymbol{p}_f s_{ft} + \boldsymbol{n}_{ft} \tag{2}$$

$s_{ft} \in \mathbb{C}$ は時間フレーム t，周波数ビン f での単一の目的音声であり，$\boldsymbol{p}_f \in \mathbb{C}^M$ は目的音声のステアリングベクトルである．$\boldsymbol{n}_{ft} \in \mathbb{C}^M$ は雑音を表している．目的音声はランク 1 伝搬過程でモデル化し，M ch マイクロホンアレイで観測する目的音声は $\boldsymbol{s}_{ft} = \boldsymbol{p}_f s_{ft} \in \mathbb{C}^M$ となる．雑音はフルランク伝搬過程でモデル化し，平均 $\boldsymbol{0} \in \mathbb{C}^M$，分散共分散行列 $\boldsymbol{Q}_f \in \mathbb{C}^{M \times M}$ のガウス分布に従う．観測信号は $\boldsymbol{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{p}_f s_{ft}, \boldsymbol{Q}_f)$ と表現できる．MVDR ビームフォーミングは目的音声方向から来る信号に歪み無し制約を置き，残存する雑音を最小化する．これに従い，MVDR ビームフォーミングフィルタ $\boldsymbol{w}_f^{\mathrm{MVDR}} \in \mathbb{C}^M$ を次のように得る．

$$\boldsymbol{w}_f^{\mathrm{MVDR}} = \frac{\boldsymbol{Q}_f^{-1} \boldsymbol{p}_f}{\boldsymbol{p}_f^H \boldsymbol{Q}_f^{-1} \boldsymbol{p}_f} \tag{3}$$

MVDR ビームフォーミングは尤度関数 $p(\boldsymbol{x}_{ft}|s_{ft})$ を最大化することでも定式化できる [21]．このとき目的音声のスケールは

考慮しない．さらに，空間相関行列が時変である $\boldsymbol{Q}_f \to \boldsymbol{Q}_{ft}$ とした場合は時変な MVDR ビームフォーミングフィルタを得ることができる．

$$\boldsymbol{w}_{ft}^{\mathrm{MVDR}} = \frac{\boldsymbol{Q}_{ft}^{-1}\boldsymbol{p}_f}{\boldsymbol{p}_f^H \boldsymbol{Q}_{ft}^{-1}\boldsymbol{p}_f} \tag{4}$$

2.2 ランク 1 MWF ビームフォーミング

ランク 1 MWF ビームフォーミングは，M ch マイクロホン観測信号 $\boldsymbol{x}_{ft}$ について，MVDR ビームフォーミングと同様の伝搬過程を仮定する．加えて単一の目的音声がガウス分布 $s_{ft} \sim \mathcal{N}_{\mathbb{C}}(0, \phi_f)$ に従うと仮定する．$\phi_f \in \mathbb{C}$ は目的音声の分散である．ランク 1 MWF ビームフォーミングフィルタは $p(s_{ft}|\boldsymbol{x}_{ft})$ を最大化する MAP 推定によって得られる．

$$\boldsymbol{w}_{f}^{\mathrm{r1MWF}} = \frac{\boldsymbol{Q}_{f}^{-1}\boldsymbol{p}_f}{\boldsymbol{p}_f^H \boldsymbol{Q}_{f}^{-1}\boldsymbol{p}_f + \phi_f^{-1}} \tag{5}$$

$$\boldsymbol{w}_{ft}^{\mathrm{r1MWF}} = \frac{\boldsymbol{Q}_{ft}^{-1}\boldsymbol{p}_f}{\boldsymbol{p}_f^H \boldsymbol{Q}_{ft}^{-1}\boldsymbol{p}_f + \phi_f^{-1}} \tag{6}$$

もし目的音声のスケールに仮定をおかない，すなわち式 (5) 及び (6) で目的音声の分散を無限大 $\phi_f \to \infty$ とするのであればランク 1 MWF ビームフォーミングは MVDR ビームフォーミングと一致する [4]．

2.3 フルランク MWF

ビームフォーミングと同様にフルランク MWF を空間フィルタリングに用いる．M ch マイクロホン観測信号 $\boldsymbol{x}_{ft}$ を次のように仮定する．

$$\boldsymbol{x}_{ft} = \boldsymbol{s}_{ft} + \boldsymbol{n}_{ft} \tag{7}$$

目的音声 $\boldsymbol{s}_{ft}$ は分散共分散行列 $\boldsymbol{P}_f \in \mathbb{C}^{M\times M}$ のガウス分布に従い，$\boldsymbol{s}_{ft} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{P}_f)$ である．雑音 $\boldsymbol{n}_{ft}$ は分散共分散行列 $\boldsymbol{Q}_f$ のガウス分布に従う．目的音声，雑音ともにフルランク伝搬過程である．観測信号 $\boldsymbol{x}_{ft}$ は同様にガウス分布に従い，$\boldsymbol{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{s}_{ft}, \boldsymbol{Q}_f)$ である．フルランク MWF はランク 1 MWF ビームフォーミングフィルタと同様に $p(\boldsymbol{s}_{ft}|\boldsymbol{x}_{ft})$ を最大化する MAP 推定によって得られる．フルランク MWF はスケールの情報を保持した空間相関行列のみから構成される [4]．

$$\boldsymbol{w}_{f}^{\mathrm{frMWF}} = (\boldsymbol{P}_f + \boldsymbol{Q}_f)^{-1}\boldsymbol{P}_f\boldsymbol{u} \tag{8}$$

$$\boldsymbol{w}_{ft}^{\mathrm{frMWF}} = (\boldsymbol{P}_{ft} + \boldsymbol{Q}_{ft})^{-1}\boldsymbol{P}_{ft}\boldsymbol{u} \tag{9}$$

M 次元単位ベクトル $\boldsymbol{u} \in \mathbb{C}^M$ はリファレンスチャネルとして選択したマイクロホンチャネルに対応する要素以外は 0 である．

本研究では，リファレンスチャネルは平均事後 SNR に基づき選択する [10]．

$$\mathrm{SNR}_{\mathrm{post},m} = \frac{\sum_{t=0}^{T}\sum_{f=0}^{F} \boldsymbol{w}_{ft,m}^H \boldsymbol{P}_f \boldsymbol{w}_{ft,m}}{\sum_{t=0}^{T}\sum_{f=0}^{F} \boldsymbol{w}_{ft,m}^H \boldsymbol{Q}_f \boldsymbol{w}_{ft,m}} \tag{10}$$

ここで $\boldsymbol{w}_{ft,m} \in \mathbb{C}^M$ はリファレンスチャネルとして m 番目のマイクロホンチャネルが選ばれたときの時間フレーム t，周波数ビン f での M 次元のフィルタである．

$$\text{リファレンスチャネル} = \operatorname*{argmax}_m \mathrm{SNR}_{\mathrm{post},m} \tag{11}$$

各発話ではそれぞれ独立にリファレンスチャネルが選択される．リファレンスチャネルを固定するよりも話者の移動に柔軟に対応できる．

3. 提 案 法

ビームフォーミングを効果的に行うには，空間相関行列を正確に推定することが重要である．近年，DNN に基づいて推定した時間周波数マスクが、空間相関行列推定に広く用いられている [8–12]．様々な環境で頑健な音声認識システムのために，DNN に基づく手法では大量の訓練データを必要とする．本研究では，従来のマスクに基づく空間相関行列推定とは異なり，各音源の位相情報を保持するマルチチャネル音源分離の枠組みに基づく教師なしの空間相関行列推定手法を提案する．

3.1 マルチチャネル非負値行列因子分解 (MNMF)

MNMF は因子分解モデルに基づく推定を利用する音源分離手法である [15, 16]．これは NMF の多チャネル拡張である．NMF では，与えられた非負値行列 $\boldsymbol{x}$ をより小さい非負値行列のペア $\boldsymbol{b}$ 及び $\boldsymbol{c}$ に分解する．音響信号処理において，頻出するスペクトル群は基底行列 $\boldsymbol{b}$ で表現され，各スペクトル群がどのタイミングでどれだけの強度を持つかはアクティベーション行列 $\boldsymbol{c}$ で表現される．

マルチチャネル音源分離に際して空間的な伝搬を考慮する必要がある．そこで観測信号をエルミート半正定値行列 $\boldsymbol{X}_{ft} = \boldsymbol{x}_{ft}\boldsymbol{x}_{ft}^H \in \mathbb{C}^{M\times M}$ として扱う．行列の対角成分は各チャネルのパワーであり，非対角成分は各チャネル間の相関を表す．MNMF では音源 l の周波数ビン f での正規化された空間相関行列 $\boldsymbol{H}_{fl} \in \mathbb{C}^{M\times M}$ を導入している．また，音源割当変数 z_{lk} で k 番目の行列が音源 l に割り当てられるかどうかを $z_{lk} = 1$ あるいは $z_{lk} = 0$ で示す．これにより次のように MNMF の因子分解モデルを定式化する．

$$\hat{\boldsymbol{X}}_{ft} = \sum_{k=1}^{K}\left(\sum_{l=1}^{L}\boldsymbol{H}_{fl}z_{lk}\right)b_{fk}c_{kt} \tag{12}$$

$b_{fk} \in \mathbb{R}_+$ 及び $c_{kt} \in \mathbb{R}_+$ は基底とアクティベーションであり，音源の低ランク構造を表現する．MNMF は観測行列 $\boldsymbol{X}$ を $[(\boldsymbol{H}\boldsymbol{z}) \circ \boldsymbol{b}]$ 及び $\boldsymbol{c}$ へと階層的に分解する．ここで $\circ$ はアダマール積を表し，$[(\boldsymbol{H}\boldsymbol{z}) \circ \boldsymbol{b}]_{fk} = \sum_{l=1}^{L}\boldsymbol{H}_{fl}z_{lk}b_{fk}$ である．

この MNMF のモデル (12) を用いて $\boldsymbol{H}_{fl}$，z_{lk}，b_{fk} 及び c_{kt} を推定する．観測した信号の空間相関行列 $\boldsymbol{X}_{ft}$ とその MNMF の因子分解モデルとの IS ダイバージェンス

$$D_{IS}(\boldsymbol{X}, \{\boldsymbol{H}, \boldsymbol{z}, \boldsymbol{b}, \boldsymbol{c}\}) = \sum_{f=1}^{F}\sum_{t=1}^{T} d_{IS}(\boldsymbol{X}_{ft}, \hat{\boldsymbol{X}}_{ft}) \tag{13}$$

を最小化する MNMF のバッチ処理アルゴリズムは，Sawada らによって乗法更新式の形で導出されている [15]．

3.2 MNMF 初期化

MNMF の性能は，正規化された空間相関行列 $\boldsymbol{H}_{fl}$ の初期値に大きく依存する [20]．効果的な $\boldsymbol{H}_{fl}$ の初期値を与えるため

に，独立低ランク行列分析 (ILRMA) [16] と観測を用いた初期化を行う．$\boldsymbol{H}_{fl}$ 以外のパラメータは無作為に初期値を与える．

ILRMA は MNMF と同様の構造で $\boldsymbol{H}_{fl}$ をランク 1 行列と仮定しており，初期値に頑健であることが知られている [16]．各音源のステアリングベクトル $\boldsymbol{r}_{fl} \in \mathbb{C}^M$ の外積で，各音源の $\boldsymbol{H}_{fl}$ を表現する．

$$\boldsymbol{H}_{fl} = \boldsymbol{r}_{fl}\boldsymbol{r}_{fl}^H \tag{14}$$

ILRMA は教師なしで各音源のステアリングベクトルを推定する．本研究では，$\boldsymbol{H}_{fl}$ を ILRMA で推定したステアリングベクトルを用いて初期化している．

さらに目的音声のステアリングベクトルとなることを期待して，$\boldsymbol{r}_{f1}$ について観測信号の空間相関行列 $\boldsymbol{X}_f$ における最大固有値に対する固有ベクトルを初期値として与える．

$$\boldsymbol{r}_{f1}^{\text{init}} = \mathcal{PE}(\boldsymbol{X}_f) \tag{15}$$

ただし，$\mathcal{PE}(\cdot)$ は第一固有ベクトルを表す．

3.3 MNMF のオンライン拡張

本研究ではバッチ処理の MNMF に加えて，オンライン処理の MNMF を導入する．時間 t に関する総和を含む統計量に着目して [22]，MNMF をオンライン拡張できる．$j(1 \leqq j \leqq J)$ 番目のブロックバッチを受け取った時の行列 $\boldsymbol{H}_{fl}$ の更新式を次に示す．上付添字 (j) は j 番目のブロックバッチ内での値であることを表す．過去の統計量は重み ρ をかけて参照している．ρ を 1 以下とすれば，現在の観測がより重み付けられる．

$$\boldsymbol{H}_{fl}^{(j)}\boldsymbol{\alpha}_{fl}^{(j)}\boldsymbol{H}_{fl}^{(j)} = \boldsymbol{H'}_{fl}^{(j)}\beta_{fl}^{(j)}\boldsymbol{H'}_{fl}^{(j)} \tag{16}$$

$$\boldsymbol{\alpha}_{fl}^{(j)} = \sum_k z_{lk}^{(j)} b_{fk}^{(j)} \sum_{t\in t^{(j)}} c_{kt}\hat{\boldsymbol{X}}_{ft}^{-1} + \rho\boldsymbol{\alpha}_{fl}^{(j-1)} \tag{17}$$

$$\beta_{fl}^{(j)} = \sum_k z_{lk}^{(j)} b_{fk}^{(j)} \sum_{t\in t^{(j)}} c_{kt}\hat{\boldsymbol{X}}_{ft}^{-1}\boldsymbol{X}_{ft}\hat{\boldsymbol{X}}_{ft}^{-1} + \rho\beta_{fl}^{(j-1)} \tag{18}$$

この代数リカッチ方程式 (16) を解き，$\boldsymbol{H}_{fl}^{(j)}$ を得る．ただし，$\boldsymbol{H'}_{fl}^{(j)}$ は更新前の行列である．z_{lk}，b_{fk} についても同様の更新式を考えることができる．c_{kt} の更新式はバッチ処理と共通である．このオンライン処理の MNMF を用いて，オンライン音声強調を実現できる．オンライン処理の MNMF の場合，3.2 節の初期化は最初のブロックバッチのみで行う．

3.4 空間相関行列及びステアリングベクトルの推定

ビームフォーミングフィルタを計算するために，MNMF で推定した値に基づいて目的音声及び雑音の空間相関行列 $\boldsymbol{P}$ 及び $\boldsymbol{Q}$ を定める．3.2 節で述べた特別な初期値を与えた音源を目的音声とみなし $l = 1$ を割り当てれば，それぞれの空間相関行列を次のように定めることができる．

$$\boldsymbol{P}_{ft} = \sum_{k=1}^{K} \boldsymbol{H}_{f1} z_{1k} b_{fk} c_{kt} \tag{19}$$

$$\boldsymbol{Q}_{ft} = \sum_{k=1}^{K} \left(\sum_{l=2}^{L} \boldsymbol{H}_{fl} z_{lk} \right) b_{fk} c_{kt} \tag{20}$$

$$\boldsymbol{P}_f = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{P}_{ft} \tag{21}$$

$$\boldsymbol{Q}_f = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{Q}_{ft} \tag{22}$$

ステアリングベクトル $\boldsymbol{p}_f$ は，目的音声の空間相関行列 $\boldsymbol{P}_f$ における最大固有値に対する固有ベクトルとして推定される．

$$\boldsymbol{p}_f = \mathcal{PE}(\boldsymbol{P}_f) \tag{23}$$

ランク 1 MWF ビームフォーミングは目的音声のスケールの分散 ϕ_f を必要とする．目的音声の空間相関行列がステアリングベクトルの二乗によって近似されるという仮定 $\boldsymbol{P}_f \simeq \phi_f \boldsymbol{p}_f \boldsymbol{p}_f^H$ に基づき計算する．

$$\phi_f \simeq \frac{\|\boldsymbol{P}_f\|}{\|\boldsymbol{p}_f\boldsymbol{p}_f^H\|} \tag{24}$$

ただし，$\|\cdot\|$ は行列ノルムを表現する．

4. 評価実験

実際の雑音環境下における音声認識実験により提案する音声強調手法を評価した．本研究では，2 つの異なる音声認識タスクを用いた．一つ目は CHiME-3 Challenge [2] であり，比較的大規模の訓練データが利用できる．二つ目は独自に収録した室内のデータを用いたタスクである．CHiME-3 Challenge のデータセットで訓練されたモデルにとっては未知環境となる．

4.1 実験設定

CHiME-3 Challenge [2] では，雑音環境に対応する訓練データとして実録音 1600 発話及び模擬録音 7138 発話があり，バス，カフェ，歩行者エリア，車道の 4 種類の雑音環境が用意されている．実録音 1320 発話で構成された評価セット (“et05_real_noisy”) における音声認識性能を単語誤り率 (WER) を用いて評価した．各発話は 6 ch で構成されており，マイクロホンの向きが異なるチャネル 2 を除いた 5 ch 分でマイクロホンアレイ処理を行った．音響モデルとして DNN-HMM [23] を前述の訓練データから構築し，その際ドロップアウト [24] 及びバッチ正規化 [25] を用いて学習した．入力は 1320 次元の特徴ベクトルであり，40 チャネルの対数メルスケールフィルタバンク (lmfb) 及びその 1 次差分と 2 次差分を 11 フレーム分用意したものである．言語モデルは標準的な WSJ コーパスで学習した単語トライグラムであり，Kaldi デコーダを用いた．

二つ目の音声認識タスクとして室内雑音環境下における日本語のテストセット Noisy JNAS を用意した．日本語新聞記事コーパス (JNAS) から音素バランス文を抽出し，男性 5 人が混雑した食堂内で計 200 文を読み上げたものである．遠隔音声認識システムとして現実的なシナリオとするために，携帯電話などの商用機器に搭載されている MEMS マイクロホンによる 5ch アレイで録音した．話者とアレイの距離は CHiME-3 よりも長い約 1 m とした．DNN-HMM 音響モデルは JNAS のクリーン音声 49678 発話に対して各々ランダムに CHiME-3 の雑音を付加したマルチコンディションデータを用いて構築した．トライグラム言語モデルは JNAS コーパスで学習し，Julius デコーダを用いた．この Noisy JNAS テストセットは CHiME-3

表 3　MNMF 実験パラメータ

サンプリング周波数	16 kHz
フレーム長	64 ms
フレームオーバーラップ	10 ms
窓関数	Hamming
マイクロホン個数 M	5
想定音源数 L	5
基底数 K	25
更新回数	10

表 4　CHiME-3 Challenge 及び Noisy JNAS の音声認識タスクにおける比較手法と提案手法 (バッチ処理) の単語誤り率

強調手法	フィルタ	式	CHiME-3	Noisy JNAS
強調なし	-	-	22.19	41.34
重み付き遅延和	時不変	-	15.54	35.28
DNNm-MVDR	時不変	-	**11.32**	16.59
MNMF-MVDR	時不変	(3)	12.05	11.39
	時変	(4)	12.06	11.39
MNMF-r1MWF	時不変	(5)	11.93	11.04
	時変	(6)	12.08	**10.94**
MNMF-frMWF	時不変	(8)	11.91	11.60
	時変	(9)	11.87	11.56

テストセットとは多くの異なる特徴を持つ．雑音環境が異なり，またマイクロホンの種類や配置も異なる．

実験にあたり，表 3 のように MNMF のパラメータを設定した．初期値に対するランダム性が実験結果に影響を与えないようにするため，各ビームフォーミング手法は同一の MNMF で推定した同じ空間相関行列でフィルタを作成した．オンライン音声強調では基本のブロックバッチサイズは 0.5 s で固定し，最初のバッチブロックバッチサイズを変えて実験を行った．過去の統計量にかける重み ρ は 0.9 とした．

比較手法として重み付き遅延和ビームフォーミングである Beamformit [13] をベースラインとした．また，マスク推定を行うフィードフォワード型 DNN を構築して，このマスクに基づくビームフォーミングとも比較を行った．DNN の構造は CHiME-3 のタスクで使用する音響モデルと同型である．入力は 1100 次元の特徴ベクトルであり，100 次元の lmfb を 11 フレーム分用意した．出力は F (= 201) 次元のマスクである．DNN は CHiME-3 データセットを使用して訓練され，MVDR ビームフォーミングのための時間周波数マスクを生成する (DNNm-MVDR)．これは CHiME-3 と Noisy JNAS の評価セットで共通して使用する．

4.2　実 験 結 果

図 2，3，4 及び 5 は Noisy JNAS の一発話における観測信号及び重み付き遅延和ビームフォーミング，DNN マスクに基づく MVDR ビームフォーミング，そして MNMF に基づく MVDR ビームフォーミング (時不変) で強調した音声のスペクトログラムである．提案法による強調音声では調波構造が最も明瞭に確認でき，背景雑音が抑圧されている．

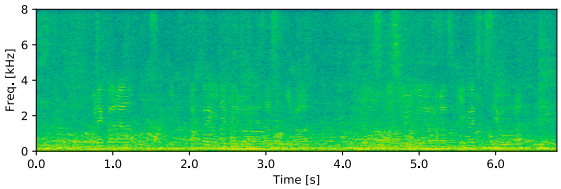

図 2　観 測 信 号

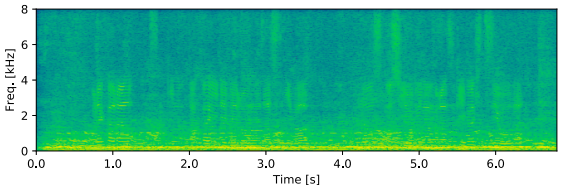

図 3　重み付き遅延和ビームフォーミング強調結果

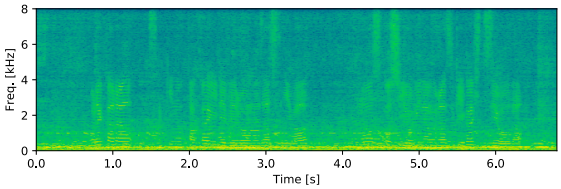

図 4　DNN マスクに基づく MVDR ビームフォーミング強調結果

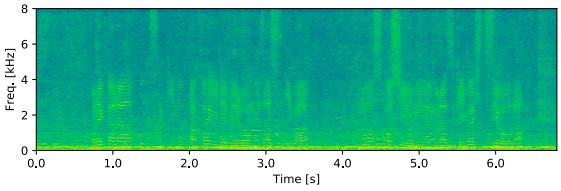

図 5　MNMF に基づく MVDR ビームフォーミング強調結果

比較手法と提案手法 (バッチ処理) の音声認識結果を表 4 に示す．CHiME-3 タスクにおいて，提案法で最も良い手法，MNMF-frMWF(時変) は重み付き遅延和ビームフォーミングと比較して WER を 3.67 ポイント改善した．提案法は，同一の適合データセットで学習した DNNm-MVDR の WER には及ばないが，事前学習なしで一貫して高い性能を示している．

Noisy JNAS の音声認識は，再訓練のためのデータを使わず行った．MNMF-r1MWF(時変) は 10.94%の WER を達成しており，これは DNNm-MVDR に対して 34.06%の改善を実現している．Noisy JNAS タスクは，マイクロホンの設定と雑音環境について CHiME-3 のタスクとは大きく異なる．提案法と比較して，DNN に基づくビームフォーミングの性能は未知環境において大きく低下している．これは CHiME-3 データに対して過学習を起こしているためと考えられる．対照的に，MNMF を空間相関行列推定に用いる提案法は双方のタスクにおいて高い性能を維持している．

どのビームフォーミング手法を用いるかについては大きな差はなかった．CHiME-3 評価データでは，フルランク MWF がわずかに高い性能を示した．Noisy JNAS タスクにおいては，ランク 1 MWF ビームフォーミングが提案する空間相関行列推定手法との組み合わせで最も有効であった．MVDR ビームフォーミングと比較して，ランク 1 MWF ビームフォーミングやフルランク MWF は目的音声のスケールに関する分布を導入している．また雑音の時変な空間共分散行列の使用は，有意な改善または劣化をもたらさなかった．

表 5　CHiME-3 Challenge 及び Noisy JNAS の音声認識タスクにおける提案手法の単語誤り率，オンライン処理におけるサイズは最初のブロックバッチサイズを表す

強調手法	処理	サイズ	CHiME-3	Noisy JNAS
MNMF-frMWF(時変)	バッチ	-	11.87	11.56
	オンライン	1.0 s	16.09	29.57
	オンライン	2.5 s	13.20	15.77
	オンライン	5.0 s	12.23	13.03

オンライン処理を含む提案手法の音声認識結果を表 5 に示す．ここでは MNMF-frMWF(時変) の結果のみを示すが，他のビームフォーミング手法も同じ傾向であった．最初のブロックバッチサイズを大きくすることで，バッチ処理には及ばないがそれに近い性能を実現した．オンライン処理で性能が低下する要因として，初期化がバッチブロックサイズに依存する問題がある．長いブロックバッチを使った方がどの周波数ビンにおいても正確に空間相関行列の初期値を与えることができる．実際に同じ初期値を使った場合は，オンライン処理とバッチ処理の性能は同程度になった．最初のブロックバッチサイズ 1.0 s の場合に性能が特に低いのは，最初のブロックバッチには目的音声があまり入っていないためだと考えられる．

5. おわりに

ビームフォーミングと MNMF の統合に基づく教師なし音声強調手法を提案した．提案法は MNMF を用いて空間相関行列を推定し，ビームフォーミングによって強調音声を出力する．実録音データに対する音声認識実験結果により，提案するバッチ音声強調処理・オンライン音声強調処理が，未知環境において DNN マスクに基づくビームフォーミングよりも頑健に動作することを確認した．

文　献

[1] B.D. Van Veen and K.M. Buckley, “Beamforming: a versatile approach to spatial filtering,” IEEE ASSP Magazine, vol.5, no.2, pp.4–24, 1988.

[2] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third CHiME speech separation and recognition challenge: Dataset, task and baselines,” Proc. of IEEE ASRU, pp.504–511, 2015.

[3] S. Gannot and I. Cohen, “Speech enhancement based on the general transfer function GSC and postfiltering,” IEEE Trans. ASLP, vol.12, no.6, pp.561–571, 2004.

[4] M. Souden, J. Benesty, and S. Affes, “On optimal frequency-domain multichannel linear filtering for noise reduction,” IEEE Trans. ASLP, vol.18, no.2, pp.260–276, 2010.

[5] Z. Wang, E. Vincent, R. Serizel, and Y. Yan, “Rank-1 constrained multichannel Wiener filter for speech recognition in noisy environments,” arXiv :1707.00201 [cs], 2017.

[6] E. Warsitz and R. Haeb-Umbach, “Blind acoustic beamforming based on generalized eigenvalue decomposition,” IEEE Trans. ASLP, vol.15, no.5, pp.1529–1539, 2007.

[7] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, “Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR,” IEEE/ACM Trans. ASLP, vol.25, no.4, pp.780–793, 2017.

[8] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, “Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming,” Proc. of IEEE ICASSP, pp.286–290, 2017.

[9] J. Heymann, L. Drude, and R. Haeb-Umbach, “Neural network based spectral mask estimation for acoustic beamforming,” Proc. of IEEE ICASSP, pp.196–200, 2016.

[10] H. Erdogan, J.R. Hershey, S. Watanabe, M.I. Mandel, and J. Le Roux, “Improved MVDR beamforming using single-channel mask prediction networks.,” Proc. of Interspeech, pp.1981–1985, 2016.

[11] X. Xiao, S. Zhao, D.L. Jones, E.S. Chng, and H. Li, “On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition,” Proc. of IEEE ICASSP, pp.3246–3250, 2017.

[12] T. Ochiai, S. Watanabe, T. Hori, and J.R. Hershey, “Multichannel end-to-end speech recognition,” Proc. of ICML, vol.70, pp.2632–2641, 2017.

[13] X. Anguera, C. Wooters, and J. Hernando, “Acoustic beamforming for speaker diarization of meetings,” IEEE Trans. ASLP, vol.15, no.7, pp.2011–2022, 2007.

[14] E. Vincent, S. Watanabe, A.A. Nugraha, J. Barker, and R. Marxer, “An analysis of environment, microphone and data simulation mismatches in robust speech recognition,” Computer Speech & Language, vol.46, pp.535–557, 2017.

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

[16] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,” IEEE Trans. ASLP, vol.24, no.9, pp.1626–1641, 2016.

[17] 島田一希，坂東宜昭，三村正人，糸山克寿，吉井和佳，河原達也，“雑音環境下音声認識のための多チャネル非負値行列因子分解に基づく教師なしビームフォーマ，” 電子情報通信学会 2017 年 8 月度音声研究会，第 117 巻，pp.19–24，2017.

[18] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, “Unsupervised beamforming based on multichannel nonnegative matrix factorization for noisy speech recognition,” Proc. of IEEE ICASSP, accepted, 2018.

[19] M. Mimura, Y. Bando, K. Shimada, S. Sakai, K. Yoshii, and T. Kawahara, “Combined multi-channel NMF-based robust beamforming for noisy speech recognition,” Proc. of Interspeech, pp.2451–2455, 2017.

[20] Y. Tachioka, T. Narita, I. Miura, T. Uramoto, N. Monta, S. Uenohara, K. Furuya, S. Watanabe, and J. Le Roux, “Coupled initialization of multi-channel non-negative matrix factorization based on spatial and spectral information,” Proc. of Interspeech, pp.2461–2465, 2017.

[21] V.A. Barroso and J.M. Moura, “Maximum likelihood beamforming in the presence of outliers,” Proc. of IEEE ICASSP, pp.1409–1412, 1991.

[22] A. Lefevre, F. Bach, and C. Févotte, “Online algorithms for nonnegative matrix factorization with the Itakura-Saito divergence,” Proc. of IEEE WASPAA, pp.313–316, 2011.

[23] A.R. Mohamed, G.E. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” IEEE Trans. ASLP, vol.20, no.1, pp.14–22, 2012.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting.,” JMLR, vol.15, no.1, pp.1929–1958, 2014.

[25] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” Proc. of ICML, pp.448–456, 2015.