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
SP2017-23(2017-08)

［ポスター講演］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

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，雑音環境下音声認識のための教師なし多チャネル音声強調について述べる．音声認識における多チャネル音声強調ではビームフォーマが一般的であり，その構成要素であるステアリングベクトルや空間相関行列の推定は DNN を用いて作成したマスクに基づく手法が主流になっている．このような教師あり手法は訓練データに過学習し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，多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空間相関行列を推定する教師なしビームフォーマを提案する．MVDR ビームフォーマ及び目的音声のスケールを考慮する MAP 推定に基づくビームフォーマにおいて，時変及び時不変フィルタに加えて，発話内で変化しないステアリングベクトルと時間フレームごとに変わる雑音の空間相関行列によるフィルタを構築した．実録音データに対する音声認識実験を行った結果，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マより頑健に動作することを示した．また時不変な目的音声のステアリングベクトル及び時変な雑音の空間相関行列を MNMF により推定した MAP ビームフォーマが最も高い性能を示した．

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，多チャネル非負値行列因子分解

1. はじめに

雑音環境下での音声認識性能の改善に向けて，ビームフォーミングを用いた多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では，目的音声方向の信号を強調し，それ以外の方向から来る音を抑圧する [1]．時間的に不変なフィルタを用いる MVDR ビームフォーミング [1] は歪みを低減しながら音声強調ができ [2–6]，CHiME Challenge など近年の国際技術評価会 [7] において雑音環境下での音声認識性能を大きく改善することが示されている [8]．時間周波数領域において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 [2–6]．

ステアリングベクトル及び空間相関行列の推定について多くの研究がある．音声区間検出（VAD）を用いた手法は，観測信号において音声区間と非音声区間の検出を利用して空間相関行列を推定するが，これだけでは音声認識性能を十分に向上させることができない [7]．近年時間周波数マスクを用いた手法が注目を集めている [2–6]．この手法は観測信号の各時間周波数ビンが雑音を含む目的音声と雑音のみに排他的に分類されるという仮定に基づく [2]．マスクを推定する際には DNN が用いられることが多く [3–6]，その DNN の学習の際に入力となるスペクトログラムと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．目的音声のステアリングベクトルは，空間相関行列の最大固有値に対する第一固有ベクトルとして近似される [2, 3, 5]．

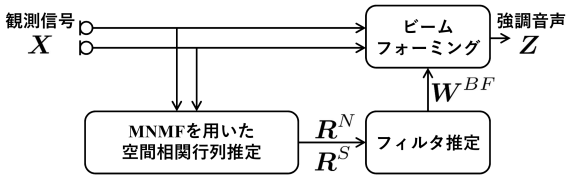

図 1　MNMF に基づくビームフォーマの処理過程

この DNN で推定したマスクに基づく MVDR ビームフォーミングにはいくつか問題点がある．MVDR ビームフォーミングでは，目的音声のステアリングベクトルが一般的に単位ベクトルとして扱われる．すなわち，空間相関行列が保持する目的音声の音量や残響に関する情報を考慮せずビームフォーミングを行っている．また話者方向が一つの発話内で変化しないとしても，雑音を常に同じ空間相関行列で表現することは不適切と考えられる．DNN に基づく教師ありマスク推定では，DNN 分類器が訓練データに対して過学習を起こしやすく，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大きく劣化するおそれがある．一方，雑音の種類やマイクロホンの配置など録音環境によらない大量の訓練データを集めるのは現実的には難しい．

本研究ではこれらの問題を解決するため，多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 [9, 10] を用いて空間相関行列を推定する教師なしビームフォーマを提案する（図 1 参照）．MNMF では多チャネルで観測した混合信号のイルミート半正定値行列から各音源の空間相関行列を推定する．これはモノラルの非負値行列因子分解（NMF）でパワー

スペクトログラムを基底行列（基底スペクトルの集合）とアクティベーション行列（時間的なアクティベーションの集合）の積で近似して表すことと同様に考えることができる．提案法の利点として教師なし手法であることに加えて，それぞれの時間周波数ビンにおいて観測音を各音源（目的音声と各雑音など）の信号へと分解していることが挙げられる．DNN や混合ガウス分布を利用したマスク推定に基づく手法は因子分解を行わずに直接空間相関行列を推定しているのに対して [2–6]，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．

本稿では目的音声のスケールや残響に関する情報を考慮するため，最大事後確率（MAP）推定に基づくビームフォーミング [11] 及び多チャネルウィナーフィルタ（MWF） [12] を用いた．MNMF で推定した空間相関行列及びステアリングベクトルからビームフォーミングフィルタを構築する際に，要素が全て時不変なフィルタ及び時変なフィルタに加えて，時不変なステアリングベクトルと雑音の時変な空間相関行列から構築される時混合フィルタを用いた．提案法は 2 つの雑音環境下音声認識タスクを用いて性能を評価した．

2. 関 連 研 究

音声認識におけるビームフォーミングの初期の試みとしては，大まかに推定した VAD の結果を使い空間相関行列を推定する手法や幾何情報を用いて推定した到達時間差（TDOA）に基づく手法がある．ただしこれらの手法は実際の雑音環境での音声認識において十分な性能を発揮できなかった．

近年時間周波数マスクに基づく手法が数多く提案されている [2–6]．この手法は観測信号を排他的に二値分類，すなわち雑音含む目的音声及び雑音のみへと分類する．特に DNN を用いて推定したスペクトルマスクに基づきビームフォーミングを行う手法が CHiME Challenge の音声認識タスクで広く使われており [3–6]，観測スペクトログラム及び IBM のペアを用いて DNN 分類器を学習している．学習した DNN により推定される時間周波数マスク $m_{ft}(0 \leqq m_{ft} \leqq 1)$ は観測信号を雑音含む目的音声及び雑音のみへ分類する．このマスクに基づく手法では空間相関行列は次のように推定される．

$$\mathsf{R}_f^{\mathrm{obs}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_{ft}\boldsymbol{x}_{ft}^H \tag{1}$$

$$\mathsf{R}_f^{N} = \frac{1}{\sum_{t=1}^{T} m_{ft}}\sum_{t=1}^{T} m_{ft}\boldsymbol{x}_{ft}\boldsymbol{x}_{ft}^H \tag{2}$$

$$\mathsf{R}_f^{S} = \mathsf{R}_f^{\mathrm{obs}} - \mathsf{R}_f^{N} \tag{3}$$

$\boldsymbol{x}_{ft} \in \mathbb{C}^M$ は時間フレーム t，周波数ビン f での M ch マイクロホンアレイにおける実観測音である．$\mathsf{R}_f^{\mathrm{obs}}$，$\mathsf{R}_f^{N}$ 及び $\mathsf{R}_f^{S} \in \mathbb{C}^{M\times M}$ はそれぞれ観測音，雑音及び目的音声の時不変な空間相関行列である．環境によらない頑健性を実現するには大量の教師ありデータが必要となる．一方，スペクトルマスクを推定する教師なしの手法として，複素混合ガウス分布に基づくブラインド音源分離を用いた手法が提案されている [2]．このマスク推定は複素混合ガウス分布を用いた 2 つのカテゴリへのクラスタリングに基づいており，空間相関行列は推定されたマスクを用いて同様に推定される．

表 1　ビームフォーミングフィルタ（添字 f 及び t は省略した）

		目的音声と雑音の混合に関する仮定	
		$\boldsymbol{x} = \boldsymbol{a}s + \boldsymbol{n}$	$\boldsymbol{x} = \sum_l \boldsymbol{x}_l$
フィルタ推定法	最尤推定	**MVDRBF** $\boldsymbol{w} = \frac{\mathsf{R}^{N^{-1}}\boldsymbol{a}}{\boldsymbol{a}^H\mathsf{R}^{N^{-1}}\boldsymbol{a}}$	ー
	事後確率最大 (MAP) 推定	**MAPBF** $\boldsymbol{w} = \frac{\mathsf{R}^{N^{-1}}\boldsymbol{a}}{\boldsymbol{a}^H\mathsf{R}^{N^{-1}}\boldsymbol{a} + \sigma^{-2}}$	**MWF** $\mathsf{W} = \mathsf{R}^S(\mathsf{R}^S + \mathsf{R}^N)^{-1}$

3. ビームフォーマ

本研究で用いる 3 種類のビームフォーマについて述べる．提案法は MNMF [9, 10] を用いた空間相関行列推定に基づくビームフォーマである．MNMF はマスクを用いない教師なし手法であり，各音源の音色構造及び音の空間的混合を考慮した因子分解モデルに基づき，観測信号を分離し空間相関行列を正確に推定する．MNMF の詳細は 4.1 節で述べる．3 種類のビームフォーマはいずれも空間相関行列 $\mathsf{R}_{f(t)}$ 及びステアリングベクトル $\boldsymbol{a}_{f(t)}$ から構築され，そのビームフォーミングフィルタは表 1 のようにそれぞれ定式化できる．表 1 は目的音声と雑音の混合に関する仮定，あるいはその仮定に基づくフィルタ推定法がそれぞれ異なることを表現している．今回ステアリングベクトルは目的音声の空間相関行列 $\mathsf{R}_{f(t)}^S$ における最大固有値に対する固有ベクトルとして計算している．

各ビームフォーマは短時間フーリエ変換領域において導出される．MVDR ビームフォーマの歴史は長く [1]，先に述べたような空間相関行列の推定手法と組み合わせて広く使われている [2–6]．MAP 推定に基づくビームフォーマは MVDR ビームフォーマと同様の混合に関する仮定を用いた上で，目的音声のスケールに関する事前分布を導入する [11]．MWF はステアリングベクトルを使わず空間相関行列のみから分離フィルタを作成するのに用いられている [12]．

3.1 MVDR ビームフォーマ

M ch マイクロホン観測信号について，$\boldsymbol{x}_{ft} = \boldsymbol{a}_{ft}s_{ft} + \boldsymbol{n}_{ft}$ と仮定する．$\boldsymbol{x}_{ft} \in \mathbb{C}^M$ は時間フレーム t, 周波数ビン f での M ch マイクロホンアレイで観測した音信号である．$\boldsymbol{n}_{ft} \in \mathbb{C}^M$ は M ch マイクロホンアレイの観測における雑音を表している．$s_{ft} \in \mathbb{C}$ 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る．$\boldsymbol{a}_{ft} = [h_{1ft}, ..., h_{Mft}] \in \mathbb{C}^M$ はステアリングベクトルを表しており，その各要素 $h_{mft} \in \mathbb{C}$ は目的音声から各マイクロホンへの伝達関数である．観測雑音は平均 $\mathbf{0} \in \mathbb{C}^M$，分散 $\mathsf{R}_{ft}^N \in \mathbb{C}^{M\times M}$ のガウス分布に従うとすると，観測音は $\boldsymbol{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{ft}s_{ft}, \mathsf{R}_{ft}^N)$ と表現できる．MVDR ビームフォーマは目的音声方向から来る信号に歪みなし制約（$\boldsymbol{w}_{ft}^H\boldsymbol{a}_{ft} = 1$）をかけた上で残存する雑音を最小化する（$\boldsymbol{w}_{ft} = \mathrm{argmin}_{\boldsymbol{w}_{ft}}\, \boldsymbol{w}_{ft}^H\mathsf{R}_{ft}^N\boldsymbol{w}_{ft}$）ように設計されており，次のように MVDR ビームフォーミングフィルタ $\boldsymbol{w}_{ft}^{\mathrm{MVDRBF}} \in \mathbb{C}^M$ が導出できる．

$$\boldsymbol{w}_{ft}^{\mathrm{MVDRBF}} = \frac{\mathsf{R}_{ft}^{N\,-1}\boldsymbol{a}_{ft}}{\boldsymbol{a}_{ft}^{H}\mathsf{R}_{ft}^{N\,-1}\boldsymbol{a}_{ft}} \tag{4}$$

MVDR ビームフォーマは最尤法に基づくビームフォーマとしても定式化でき [13]，尤度関数 $p(\boldsymbol{x}_{ft}|s_{ft})$ を最大化することでも導出できる．

$$\boldsymbol{w}_{ft}^{\mathrm{MLBF}} = \frac{\mathsf{R}_{ft}^{N\,-1}\boldsymbol{a}_{ft}}{\boldsymbol{a}_{ft}^{H}\mathsf{R}_{ft}^{N\,-1}\boldsymbol{a}_{ft}} \tag{5}$$

雑音の空間相関行列は各時間フレームごとに変化するが，話者方向は一つの発話内で変化しない，すなわちステアリングベクトルが発話単位で時不変という仮定（$\boldsymbol{a}_{ft} \to \boldsymbol{a}_f$）を置くことで，別の MVDR ビームフォーミングフィルタを導出できる．

$$\boldsymbol{w}_{ft}^{\mathrm{MVDRBF}} = \frac{\mathsf{R}_{ft}^{N\,-1}\boldsymbol{a}_{f}}{\boldsymbol{a}_{f}^{H}\mathsf{R}_{ft}^{N\,-1}\boldsymbol{a}_{f}} \tag{6}$$

さらにステアリングベクトル，空間相関行列ともに時不変であるという仮定（$\boldsymbol{a}_{ft} \to \boldsymbol{a}_f$ 及び $\mathsf{R}_{ft}^{N} \to \mathsf{R}_{f}^{N}$）を置くことで，時不変なフィルタを考えることができる．

$$\boldsymbol{w}_{f}^{\mathrm{MVDRBF}} = \frac{\mathsf{R}_{f}^{N\,-1}\boldsymbol{a}_{f}}{\boldsymbol{a}_{f}^{H}\mathsf{R}_{f}^{N\,-1}\boldsymbol{a}_{f}} \tag{7}$$

マスクに基づく MVDR ビームフォーマでは $\mathsf{R}_{f(t)}^{N}$ 及び $\mathsf{R}_{f(t)}^{S}(\to \boldsymbol{a}_{f(t)})$ は式 (2) 及び (3) によって推定される．本研究では 4.1 節に示すように，因子分解モデルに基づく MNMF を用いた新たな空間相関行列推定手法を提案する．

3.2 MAP 推定に基づくビームフォーマ

M ch マイクロホン観測信号について，MVDR ビームフォーマと同様の仮定を置いた上で，MAP 推定に基づくビームフォーマ [11] は単一の目的音声に対して事前分布を導入し，$s_{ft} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{ft}^2)$ とする．これに基づき，MAP ビームフォーミングフィルタは事後確率 $p(s_{ft}|\boldsymbol{x}_{ft})$ を最大化することで求められる．

$$\boldsymbol{w}_{ft}^{\mathrm{MAPBF}} = \frac{\mathsf{R}_{ft}^{N\,-1}\boldsymbol{a}_{ft}}{\boldsymbol{a}_{ft}^{H}\mathsf{R}_{ft}^{N\,-1}\boldsymbol{a}_{ft} + \sigma_{ft}^{-2}} \tag{8}$$

もし事前分布を置かない，すなわち $\sigma_{ft}^2 \to \infty$ であれば MAP 推定は最尤推定と一致する．この事前分布を置くことによって，MAP ビームフォーマは目的音声のスケールに関する情報を利用することができる．

3.1 節と同様にして，雑音の空間相関行列は各時間フレームごとに変化するが話者方向及び目的音声のスケールは一つの発話内で変化しない，すなわちステアリングベクトルと事前分布が発話単位で時不変であるという仮定（$\boldsymbol{a}_{ft} \to \boldsymbol{a}_f$ 及び $\sigma_{ft} \to \sigma_f$）を置くことができ，次のようなフィルタを考えられる．

$$\boldsymbol{w}_{ft}^{\mathrm{MAPBF}} = \frac{\mathsf{R}_{ft}^{N\,-1}\boldsymbol{a}_{f}}{\boldsymbol{a}_{f}^{H}\mathsf{R}_{ft}^{N\,-1}\boldsymbol{a}_{f} + \sigma_{f}^{-2}} \tag{9}$$

時不変な MAP ビームフォーミングフィルタについても同様に導出できる．

$$\boldsymbol{w}_{f}^{\mathrm{MAPBF}} = \frac{\mathsf{R}_{f}^{N\,-1}\boldsymbol{a}_{f}}{\boldsymbol{a}_{f}^{H}\mathsf{R}_{f}^{N\,-1}\boldsymbol{a}_{f} + \sigma_{f}^{-2}} \tag{10}$$

3.3 多チャネルウィーナーフィルタ

MWF は空間フィルタリングとして特定の目的音声を抽出する [12]．前述のビームフォーマと異なり，MWF は元の目的音声を出力するのではなく，観測された M ch の目的音声を出力する．ここで M ch マイクロホン観測信号について $\boldsymbol{x}_{ft} = \sum_l \boldsymbol{x}_{lft}$ と仮定する．$\boldsymbol{x}_{lft} \in \mathbb{C}^M$ は音源 l からの信号を M ch マイクロホンアレイで観測したものである．雑音環境下音声認識という目的のため，$l = 1$ を目的音声としその他の音源を雑音とする．各音源の信号は平均 $\boldsymbol{0} \in \mathbb{C}^M$，分散 $\mathsf{R}_{lft} \in \mathbb{C}^{M\times M}$ のガウス分布に従うものとする．観測信号 $\boldsymbol{x}_{ft} \in \mathbb{C}^M$ は同様にガウス分布に従い，$\boldsymbol{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{x}_{1ft}, \sum_{l\neq 1}\mathsf{R}_{lft})$ である．前述のビームフォーマで用いた表記に基づき，目的音源からの観測信号 $\boldsymbol{x}_{1ft}$ は $\boldsymbol{s}_{ft}$ とし，空間相関行列 $\sum_{l\neq 1}\mathsf{R}_{lft}$ 及び R_{1ft} は R_{ft}^{N} 及び R_{ft}^{S} に置き換える．フィルタ $\mathsf{W}_{ft}^{\mathrm{MWF}}$ は事後確率 $p(\boldsymbol{s}_{ft}|\boldsymbol{x}_{ft})$ を最大化するように計算され，以下のように導出される．

$$\mathsf{W}_{ft}^{\mathrm{MWF}} = \mathsf{R}_{ft}^{S}(\mathsf{R}_{ft}^{S} + \mathsf{R}_{ft}^{N})^{-1} \tag{11}$$

MWF は空間相関行列から直接構成され，残響及びスケールの情報を保持する．

空間相関行列を時不変のものとすると，各音源信号の観測音 $\boldsymbol{x}_{lft}$ は平均 $\boldsymbol{0} \in \mathbb{C}^M$，分散 $\mathsf{R}_{lf} \in \mathbb{C}^{M\times M}$ のガウス分布に従うものと仮定できる．これに基づき，時不変な MWF を考えることができる．

$$\mathsf{W}_{f}^{\mathrm{MWF}} = \mathsf{R}_{f}^{S}(\mathsf{R}_{f}^{S} + \mathsf{R}_{f}^{N})^{-1} \tag{12}$$

4. 提 案 法

本稿では MNMF [9,10] に基づくビームフォーマを提案する．提案法は 3 段階で構成される（図 1 参照）．

（1） MNMF を用いて入力となる観測信号 $\boldsymbol{X}$ から空間相関行列 $\boldsymbol{R}$ を推定する

（2） 推定された空間相関行列 $\boldsymbol{R}$ からビームフォーミングフィルタ $\boldsymbol{W}$ を推定する

（3） ビームフォーミングフィルタ $\boldsymbol{W}$ によって強調された音声 $\boldsymbol{Z}$ を出力する

4.1 MNMF を用いた空間相関行列推定

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．ここでは，MNMF [9] を用いて観測信号から空間相関行列を推定する方法について述べる．

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．そのモデルは NMF の多チャネル拡張である．NMF 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 $\boldsymbol{T}$ 及び $\boldsymbol{V}$ に分解し，$\hat{x}_{ft} = \sum_{k=1}^{K} t_{fk}v_{kt}$ のように観測信号を因子分解で近似する．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 $\boldsymbol{T}$ 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 $\boldsymbol{V}$ で表現される．

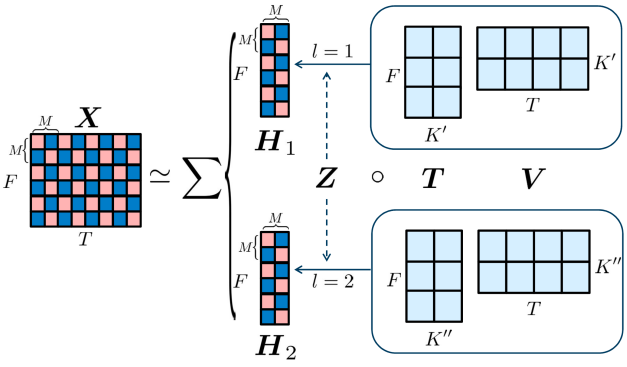

図 2　提案手法における MNMF の因子分解図例

MNMF はこの NMF を多チャネル音源分離に拡張したものである．各音源の音色構造は NMF 部分で表現できるが，多チャネル拡張には音の空間伝達過程を考慮する必要があり，観測信号をイルミート半正定値行列 $\mathsf{X}_{ft} = \boldsymbol{x}_{ft}\boldsymbol{x}_{ft}^H$ 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互相関である．観測された行列は各音源の元の音色構造を単純に足し合わせたものではなく，各音源からマイクアレイへの空間伝達過程を考慮する必要がある．MNMF では，NMF 部分の k 番目の基底における周波数ビン f での空間的な性質を行列 H_{fk} で表現する．行列の大きさは $M \times M$ であり，X_{ft} と一致する．$\hat{\mathsf{X}}_{ft}$ を要素積の形式でモデル化すると，$\hat{\mathsf{X}}_{ft} = \sum_{k=1}^{K} \mathsf{H}_{fk} t_{fk} v_{kt}$ であり，観測行列 $\boldsymbol{X}$ を $\boldsymbol{H} \circ \boldsymbol{T}$ 及び $\boldsymbol{V}$ へと階層的に分解するモデルを持つ．ここで $\circ$ はアダマール積を表現しており，$[\boldsymbol{H} \circ \boldsymbol{T}]_{fk} = \mathsf{H}_{fk} t_{fk}$ である．

MNMF を実際に複数音源の分離タスクに用いるためには，各 NMF 部分の基底を各音源に割り当てることが必要である．K 個の行列 $\mathsf{H}_{f1}, ..., \mathsf{H}_{fK}$ を音源 $l = 1, ..., L$ $(< K)$ に割り当てる．そして音源割当変数 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．これにより前述の H_{fk} は $\sum_{l=1}^{L} \mathsf{H}_{fl} z_{lk}$ に置き換えられ，次のような MNMF 因子分解モデルを考えることができる．

$$\hat{\mathsf{X}}_{ft} = \sum_{k=1}^{K} \left(\sum_{l=1}^{L} \mathsf{H}_{fl} z_{lk} \right) t_{fk} v_{kt} \tag{13}$$

t_{fk} 及び v_{kt} は NMF 部分の基底とアクティベーションであり，音源の音色構造を表現する．H_{fl} は空間の混合過程を表現する．z_{lk} は k 番目の基底が音源 l に所属するか否かを表している（図 2 参照）．

この MNMF 因子分解モデル（13）を用いて H_{fl}, z_{lk}, t_{fk} 及び v_{kt} を推定する．観測された行列 X_{ft} とその因子分解モデルの IS ダイバージェンス $D_{IS}(\boldsymbol{X}, \{\boldsymbol{T}, \boldsymbol{V}, \boldsymbol{H}, \boldsymbol{Z}\}) = \sum_{f=1}^{F} \sum_{t=1}^{T} d_{IS}(\mathsf{X}_{ft}, \hat{\mathsf{X}}_{ft})$ を最小化する MNMF のアルゴリズムは，Sawada らによって乗法更新式の形で導出されている [9]．この更新式を用いて観測信号 $\boldsymbol{x}_{ft}$ から空間伝達の性質を表す行列 H_{fl}，音源割当隠れ変数 z_{lk}，基底 t_{fk}，及びアクティベーション v_{kt} を推定する．

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き目的音声及び雑音の空間相関行列を定める．平均パワー最大の音源を目的音声とみなし $l = 1$ を割り当てれば，各行列を次のように定めることができる．

$$\mathsf{R}_{ft}^{S} = \sum_{k=1}^{K} \mathsf{H}_{f1} z_{1k} t_{fk} v_{kt} \tag{14}$$

$$\mathsf{R}_{ft}^{N} = \sum_{k=1}^{K} \left(\sum_{l=2}^{L} \mathsf{H}_{fl} z_{lk} \right) t_{fk} v_{kt} \tag{15}$$

4.2　ビームフォーミングフィルタ推定

推定した空間相関行列から 3. 節の定義に基づき，ビームフォーミングフィルタを計算する．式 (4)，(6)，(7)，(8)，(9)，(10)，(11) 及び (12) の計算にあたって，空間相関行列 $\mathsf{R}_{f(t)}$，ステアリングベクトル $\boldsymbol{a}_{f(t)}$ 及び目的音声の分散に関する事前分布 $\sigma_{f(t)}$ が必要である．各要素は時間フレーム単位で変化する場合と，一つの発話内では変化しない場合を考える．時変な空間相関行列 R_{ft}^{S} 及び R_{ft}^{N} は MNMF に基づき式 (14)(15) から計算する．時不変な空間相関行列 R_{f}^{S} 及び R_{f}^{N} は次のように定める．

$$\mathsf{R}_{f}^{S} = \frac{1}{T} \sum_{t=1}^{T} \mathsf{R}_{ft}^{S} \tag{16}$$

$$\mathsf{R}_{f}^{N} = \frac{1}{T} \sum_{t=1}^{T} \mathsf{R}_{ft}^{N} \tag{17}$$

ステアリングベクトル $\boldsymbol{a}_{f(t)}$ は目的音声の空間相関行列 $\mathsf{R}_{f(t)}^{S}$ における最大固有値に基づく第一固有ベクトルとして推定される．

$$\boldsymbol{a}_{f(t)} = \mathcal{P}(\mathsf{R}_{f(t)}^{S}) \tag{18}$$

MAP 推定に基づくビームフォーマの構築に必要な目的音声の分散に対する事前分布 $\sigma_{f(t)}$ は，目的音声の空間相関行列がステアリングベクトルの自乗によって近似されるという仮定 $\mathsf{R}_{f(t)}^{S} \simeq \sigma_{f(t)} \boldsymbol{a}_{f(t)} \boldsymbol{a}_{f(t)}^H$ に基づき計算する．

$$\sigma_{f(t)} \simeq \frac{\mathrm{norm}\{\mathsf{R}_{f(t)}^{S}\}}{\mathrm{norm}\{\boldsymbol{a}_{f(t)} \boldsymbol{a}_{f(t)}^H\}} \tag{19}$$

ただし $\mathrm{norm}\{\mathsf{M}\}$ は行列 M のノルムを表現する．

4.3　ビームフォーミングによる音声強調

ビームフォーミングフィルタ $\boldsymbol{w}_{f(t)}$ によって観測信号 $\boldsymbol{x}_{ft}$ から強調音声 $y_{ft} \in \mathbb{C}$ を取り出す．3. 節で示したように $\boldsymbol{w}_{f(t)}$ は線形フィルタである．

$$y_{ft} = \boldsymbol{w}_{f(t)}^H \boldsymbol{x}_{ft} \tag{20}$$

MWF では目的音源の M ch 観測信号を推定した $\boldsymbol{y}_{ft} \in \mathbb{C}^M$ を出力する．

$$\boldsymbol{y}_{ft} = \mathsf{W}_{f(t)} \boldsymbol{x}_{ft} \tag{21}$$

強調音声 y_{ft} として $\boldsymbol{y}_{ft}^{(m=1)}$ を用いる．

5.　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音声強調手法を評価した．

表 2　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
更新回数	200
マイクロホン個数 M	5
基底数 K	25
想定音源数 L	5

5.1　音声認識タスク

各手法の頑健性を評価するため，二つの音声認識タスクを用いた．一つ目は CHiME-3 Challenge [7] を用いた．雑音環境に対応する訓練データは実録音及び模擬録音があり，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．今回は実録音 1300 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，今回は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 [14] を前述の訓練データを用いて構築し，その際ドロップアウト [15] 及びバッチ正規化 [16] 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 階微分と 2 階微分を 11 フレーム分用意したものを用いた．言語モデルは標準的な WSJ トライグラムであり，デコーダは Kaldi デコーダ [17] を用いた．

二つ目の音声認識タスクは室内雑音環境下における日本語のテストセット（“Noisy_JNAS”）を用いた．日本語新聞記事コーパス（JNAS） [18] から音素バランス文を選び，男性 5 人が混雑した食堂内で計 200 文を読み上げた．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，携帯電話を含む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 よりも長く約 1 m とした．DNN-HMM 音響モデルは JNAS のクリーン音声を用いて構築した．トライグラム言語モデルは JNAS で学習し，Julius デコーダ [19] を用いた．

実験の準備にあたり，表 2 のように MNMF のパラメータを設定した．音源割当隠れ変数 z_{lk}，基底 t_{fk} 及びアクティベーション v_{kt} は初期値をランダムに与えている．空間伝達の性質を表す行列 H_{fl} に初期値を与える際にランク 1 MNMF [10] 及びクロススペクトル法を用いた．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，各提案法は同一の MNMF で推定した同じ空間相関行列でフィルタを作成した．

比較手法として Beamformit [20] 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，マスクに基づくビームフォーマと比較を行った．DNN の構造は CHiME-3 のタスクで使用した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 で予測したスペクトルマスクに基づく MVDR ビームフォーマ（DNNm-MVDRBF）に対して未知環境での性能を評価するため，“Noisy_JNAS” セットにおいても CHiME-3 コーパスで学習した DNN を用いてマスク推定を行った．

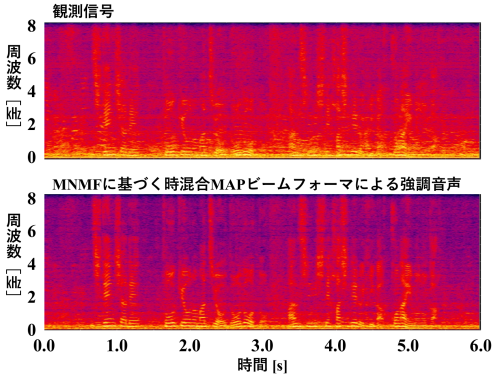

図 3　“Noisy_JNAS” における強調結果例

表 3　CHiME-3 Challenge 及び “Noisy_JNAS” の音声認識タスクにおける単語誤り率

強調手法	フィルタ	式	CHiME	JNAS
強調なし	-	-	22.39	69.97
Beamformit	時不変	-	15.60	59.43
DNNm-MVDRBF	時不変	-	**11.51**	28.47
MNMF-MVDRBF	時変	(4)	12.92	24.34
	時混合	(6)	12.61	21.28
	時不変	(7)	12.63	21.16
MNMF-MAPBF	時変	(8)	14.52	26.31
	時混合	(9)	**12.46**	**19.43**
	時不変	(10)	12.61	21.44
MNMF-MWF	時変	(11)	12.70	19.86
	時不変	(12)	12.89	21.57

5.2　雑音環境下音声認識結果

図 3 は観測音及び MNMF に基づく時混合 MAP ビームフォーマ （時混合 MNMF-MAPBF）で強調した音声のスペクトログラムである．時 “混合” とはビームフォーミングフィルタが一つの発話内で変化しない目的音声へのステアリングベクトルと時間フレームごとに変わる雑音の空間相関行列で構成されている場合である（式 (9)）．提案法により調波構造が復元され，背景雑音が抑圧されている．

音声認識結果を表 3 に示す．CHiME-3 Challenge において，提案法で最も良いビームフォーマ，時混合 MNMF-MAPBF は Beamformit と比較して WER を 3.14 ポイント改善した．提案法は，実験環境に適合したデータで学習した DNNm-MVDRBF と同等の WER を達成していないが，事前学習なしで一貫して高い性能を示している．“Noisy_JNAS” での音声認識タスクは，話者とマイクロホンアレイ間の距離が長く，また DNN-HMM 音響モデルがクリーン音声で学習されており，より難しいと考えられる．提案法と比較して，DNNm-MVDRBF の性能はこのタスクにおいて著しく低下している．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．時不変なステアリングベクトルと時変な雑音の空間相関行列から構成される時混合 MNMF-MAPBF は 19.43%の WER を達成しており，これは DNNm-MVDRBF に

対して 31.8%の改善を見せている．

MNMF に基づく MVDR 及び MAP ビームフォーマにおいて，時混合及び時不変ビームフォーマは時変ビームフォーマよりも明らかに良い性能を示している．この違いはステアリングベクトルが一つの発話内で不変かそうでないかである．毎時間フレームごとに正確なステアリングベクトルを推定することは難しい．一発話内で平均を取ることでより正確なステアリングベクトル推定につながり，時混合及び時不変ビームフォーマがより機能したと考えられる．時混合ビームフォーマはわずかながら時不変のものよりも良い性能を示した．時混合ビームフォーマは雑音の表現に時変な空間相関行列を用いている．MNMF を採用することで毎時間フレームごとに空間相関行列を正確に推定できており，よりよい音声強調につながったと考えられる．

MNMF に基づく MAP ビームフォーマにおいて，時混合及び時不変のものと時変ビームフォーマの間には大きな違いが見られる．MAP ビームフォーマは目的音声のスケールに対して事前分布 σ を導入しているが，時間フレームごとに変化する事前分布の推定はあまり機能していない．一発話の平均を取ることで目的音声に関する事前分布の正確な推定が可能になった．これにより時混合及び時不変フィルタにおいて事前分布の導入が効果を発揮し，MAP ビームフォーマは MVDR ビームフォーマよりも高い性能を示したと考えられる．

MNMF に基づく MWF はビームフォーマと同等の性能を示した．時不変フィルタはビームフォーマにおいて一貫して高い性能を発揮しているが，MWF においては時変フィルタの方がより良い性能を示した．これは MNMF を用いることで各時間フレームごとに空間相関行列が正確に推定されているためと考えられる．したがって時変ビームフォーマが機能していないのは，空間相関行列から時変なステアリングベクトルに適切に変換できていないためと考えられる．すなわち，空間相関行列を推定しその行列の第一固有ベクトルをステアリングベクトルとするという手法が効果的かどうかを検討する余地がある．

2 つの音声認識タスクにおける実験結果は，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マより頑健に機能することを示した．また提案法のうち最も効果的なものは，MNMF を用いて推定した一つの発話内で変化しないステアリングベクトルと時間フレームごとに変わる雑音の空間相関行列による時混合 MAP ビームフォーマであった．

6. おわりに

本稿では MNMF に基づくブラインド音源分離を用いた教師なしビームフォーマを提案した．提案法は MNMF を空間相関行列推定に用いて，教師なしでビームフォーマを構築する．MVDR 及び MAP ビームフォーマにおいて，時変及び時不変フィルタに加え，発話内で変化しないステアリングベクトルと時間フレームごとに変わる雑音の空間相関行列による時混合フィルタを構築した．実録音データに対する音声認識実験結果により，提案法が未知環境において最先端の DNN マスクに基づくビームフォーマよりも頑健に動作することを示した．また提案法のうち，MNMF により時不変な目的音声のステアリングベクトル及び時変な雑音の空間相関行列を推定した時混合 MAP ビームフォーマが最も効果的であった．今後他のブラインド音源分離手法，例えばランク 1 MNMF [10] をステアリングベクトル推定に用いることを検討する予定である．

文　献

[1] B.D. Van Veen, et al., “Beamforming: A versatile approach to spatial filtering,” IEEE ASSP Magazine, vol.5, no.2, pp.4–24, 1988.

[2] T. Higuchi, et al., “Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise,” IEEE ICASSP, pp.5210–5214, 2016.

[3] J. Heymann, et al., “Neural network based spectral mask estimation for acoustic beamforming,” IEEE ICASSP, pp.196–200, 2016.

[4] H. Erdogan, et al., “Improved MVDR beamforming using single-channel mask prediction networks,” INTERSPEECH, pp.1981–1985, 2016.

[5] T. Nakatani, et al., “Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming,” IEEE ICASSP, pp.286–290, 2017.

[6] X. Xiao, et al., “On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition,” IEEE ICASSP, pp.3246–3250, 2017.

[7] J. Barker, et al., “The third CHiME speech separation and recognition challenge: Dataset, task and baselines,” IEEE ASRU, pp.504–511, 2015.

[8] T. Yoshioka, et al., “The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices,” IEEE ASRU, pp.436–443, 2015.

[9] H. Sawada, et al., “Multichannel extensions of non-negative matrix factorization with complex-valued data,” IEEE TASLP, vol.21, no.5, pp.971–982, 2013.

[10] D. Kitamura, et al., “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,” IEEE TASLP, vol.24, no.9, pp.1626–1641, 2016.

[11] S. Malik, et al., “A bayesian framework for blind adaptive beamforming,” IEEE TSP, vol.62, no.9, pp.2370–2384, 2014.

[12] S. Doclo, et al., “GSVD-based optimal filtering for single and multimicrophone speech enhancement,” IEEE TSP, vol.50, no.9, pp.2230–2244, 2002.

[13] V.A. Barroso, et al., “Maximum likelihood beamforming in the presence of outliers,” IEEE ICASSP, pp.1409–1412, 1991.

[14] A.R. Mohamed, et al., “Acoustic modeling using deep belief networks,” IEEE TASLP, vol.20, no.1, pp.14–22, 2012.

[15] N. Srivastava, et al., “Dropout: a simple way to prevent neural networks from overfitting,” JMLR, vol.15, no.1, pp.1929–1958, 2014.

[16] S. Ioffe, et al., “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” ICML, pp.448–456, 2015.

[17] D. Povey, et al., “The Kaldi speech recognition toolkit,” IEEE ASRU, 2011.

[18] K. Itou, et al., “JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research,” JASJ (E), vol.20, no.3, pp.199–206, 1999.

[19] A. Lee, et al., “Julius — an open source real-time large vocabulary recognition engine,” EUROSPEECH, pp.1691–1694, 2001.

[20] X. Anguera, et al., “Acoustic beamforming for speaker diarization of meetings,” IEEE TASLP, vol.15, no.7, pp.2011–2022, 2007.