6M-02

# 遠隔音声認識のためのブラインド音源分離に基づくビームフォーマ

島田 一希　坂東 宜昭　板倉 光佑　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 知能情報学専攻

## 1. はじめに

マイクロホンと話者が離れた遠隔音声認識 (DSR) の実現には，入力信号に目的音声だけでなく雑音も含まれるので，前処理として音声強調処理が不可欠である．DSR のための音声強調では，できるだけ少ない環境の事前情報で動作し，強調処理による歪みが少ないことが求められている．このようなシステムを実現するため，国際技術評価会 CHiME Challenge [1] が開催され，マイクロホンアレイ付きタブレット端末を用いた雑踏中での DSR の性能評価が行われている．

目的音声の伝達関数と雑音信号の空間情報が既知であれば，マイクロホンアレイを用いたビームフォーミングによる音声強調が可能である [2]．この手法は音の空間中の伝達過程のモデルに基づき動作するため，歪みの少ない音声強調が可能である．一般に目的音声の伝達関数や雑音信号の空間情報は未知であるので，これらはブラインド音源分離手法を用いて推定される．例えば，音声と背景雑音の音色の違いなどから，観測スペクトログラムの各時間周波数点に含まれる目的音声の度合いを表すソフトマスクを推定する分離手法 [3,4] が CHiME Challenge で高い性能を達成している．

本稿では，ブラインド音源分離の一種である多チャンネル非負値行列因子分解 (MNMF) [5] とビームフォーマを用いた音声強調について述べる (図 1)．従来のソフトマスクを用いた分離手法では音の空間的な伝達過程が適切に考慮されておらず，DSR の性能劣化の一因と考えられる．そこで，音の伝達過程として線形時不変性を仮定した MNMF を用いることで性能向上を実現する．

## 2. 関連研究

関連研究としてマスクに基づくビームフォーマを示す．

### 2.1 ビームフォーマ

マイクロホン観測信号 $y$ について次のように定義する．

$$y_{m,ft} = h_{m,f} s_{ft} + u_{m,ft} \tag{1}$$

$y_{m,ft}, u_{m,ft}$ は，$m$ 番目のマイクロホンにおける時間 $t$ と周波数 $f$ の観測音と雑音であり，$s_{ft}$ は単一の目的音源である．$h_{m,f}$ は有限インパルス応答であり，$\boldsymbol{h}_f = [h_{1,f}, ..., h_{M,f}]$ をステアリングベクトルと呼ぶ．ビームフォーマによる音声強調は線形フィルタ $\boldsymbol{w}$ を用いる．

$$z_{ft} = \boldsymbol{w}_f^H \boldsymbol{y}_{ft} \tag{2}$$

$\boldsymbol{y}_{ft} = [y_{1,ft}, ..., y_{M,ft}]$ は観測音であり，$z_{ft}$ は強調音声である．線形フィルタ $\boldsymbol{w}$ は時不変な場合 $\boldsymbol{w}_f$ であり，時変な場合 $\boldsymbol{w}_{ft}$ である．

**MV(Minimum Variance) ビームフォーマ**　目的音源の全域通過特性を保証する拘束条件をかけながら，ビームフォーマの平均出力パワー (分散) を最小化する．

$$\boldsymbol{w}_f^{(\mathrm{MV})} = \frac{\boldsymbol{R}_f^{\mathrm{all}^{-1}} \boldsymbol{h}_f^{\mathrm{speech}}}{\boldsymbol{h}_f^{\mathrm{speech}^H} \boldsymbol{R}_f^{\mathrm{all}^{-1}} \boldsymbol{h}_f^{\mathrm{speech}}} \tag{3}$$

Beamformer based on Blind Source Separation for Distant Speech Recognition : *Kazuki Shimada, Yoshiaki Bando, Kousuke Itakura, Masato Mimura, Katsutoshi Itoyama, Kazuyoshi Yoshii, Tatsuya Kawahara (Kyoto Univ.)*

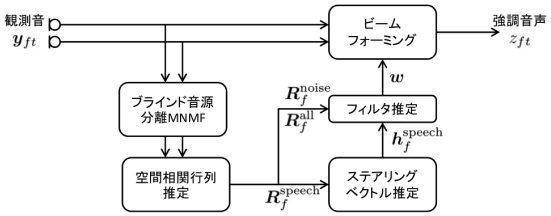


図 1: MNMF に基づくビームフォーミング処理の流れ

**ML(Maximum Likelihood) ビームフォーマ**　観測音 $\boldsymbol{y}_{ft}$ を単一の目的音源とガウス雑音の和と仮定した場合の尤度関数を最大化する．

$$\boldsymbol{w}_f^{(\mathrm{ML})} = \frac{\boldsymbol{R}_f^{\mathrm{noise}^{-1}} \boldsymbol{h}_f^{\mathrm{speech}}}{\boldsymbol{h}_f^{\mathrm{speech}^H} \boldsymbol{R}_f^{\mathrm{noise}^{-1}} \boldsymbol{h}_f^{\mathrm{speech}}} \tag{4}$$

式 (3)(4) では，空間相関行列 $\boldsymbol{R}_f^{\mathrm{noise}}, \boldsymbol{R}_f^{\mathrm{all}}$ とステアリングベクトル $\boldsymbol{h}_f^{\mathrm{speech}}$ が必要になる．ステアリングベクトルは目的音声の空間相関行列 $\boldsymbol{R}_f^{\mathrm{speech}}$ の最大固有値に対する固有ベクトルを計算することで近似できる．

### 2.2 時間周波数マスクに基づくビームフォーマ

観測音を雑音と目的音声に分ける時間周波数マスク $M_{ft}$ $(0 \leq M_{ft} \leq 1)$ で分離し空間相関行列を推定する．

$$\boldsymbol{R}_f^{\mathrm{all}} = \frac{1}{T} \sum_{t=0}^{T} \boldsymbol{y}_{ft} \boldsymbol{y}_{ft}^H \tag{5}$$

$$\boldsymbol{R}_f^{\mathrm{noise}} = \frac{1}{\sum_{t=0}^{T} M_{ft}} \sum_{t=0}^{T} M_{ft} \boldsymbol{y}_{ft} \boldsymbol{y}_{ft}^H \tag{6}$$

$$\boldsymbol{R}_f^{\mathrm{speech}} = \boldsymbol{R}_f^{\mathrm{all}} - \boldsymbol{R}_f^{\mathrm{noise}} \tag{7}$$

マスク $M_{ft}$ は複素ガウス混合モデルによるクラスタリング [3] や DNN [4] を用いて推定される．

## 3. 提案法

本稿では、音の混合過程を適切に表現したブラインド音源分離の一つである MNMF [5] に基づくビームフォーマを提案する (図 1)．提案法は，まず 1) MNMF を用いて観測音から空間相関行列を推定し，2) 得られた空間相関行列を用いてビームフォーミングすることで，音声強調を行う．MNMF は，音の空間的な混合を表す空間モデルとして，線形時不変モデルを持つ．さらに，音源の音色構造を表す音源モデルとして，音源のスペクトログラムに低ランク性を仮定する．

ビームフォーマには関連研究の章で述べた式 (3)(4) を用いる．ここでは，MNMF を用いて観測音から目的音声や雑音の空間相関行列を推定する手法について述べる．

### 3.1 MNMF

観測音をエルミート半正定値行列 $\mathsf{Y}_{ft}$ として扱う．

$$\mathsf{Y}_{ft} = \boldsymbol{y}_{ft} \boldsymbol{y}_{ft}^H \tag{8}$$

行列の対角成分は各チャネルのパワーであり，非対角成分はチャネル間の相関である．観測音のモデル $\hat{\mathsf{Y}}_{ft}$ は

$$\mathsf{Y}_{ft} \approx \hat{\mathsf{Y}}_{ft} = \sum_{k=1}^{K} \left( \sum_{n=1}^{N} \mathsf{H}_{n,f} z_{nk} \right) t_{fk} v_{kt} \tag{9}$$

である．$t_{fk}, v_{kt}$ は NMF 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す．$\mathsf{H}_{n,f}$ は空間における混合過程を表す．$z_{nk}$ は $k$ 番目の基底がどの音源 $n$ を指し示しているかを表す．今回は Sawada らにより導出された更新式 [5] を用いて $\mathsf{H}_{n,f}, z_{nk}, t_{fk}, v_{kt}$ を推定している．

### 3.2 MNMF を用いた空間相関行列推定

MNMF を用いて空間相関行列の推定を行う場合，(9) で計算された音源のモデル $\hat{\mathsf{Y}}_{ft}$ を直接用いる手法と，分離した音を各音源の観測とみなして求める手法がある．

**モデルベース**　混合音の空間相関行列は

$$\boldsymbol{R}_{ft}^{\mathrm{all}} = \sum_{k=1}^{K} \left( \sum_{n=1}^{N} \mathsf{H}_{n,f} z_{nk} \right) t_{fk} v_{kt} \tag{10}$$

目的音声が $n=1$ にあるとすれば

$$\boldsymbol{R}_{ft}^{\mathrm{speech}} = \sum_{k=1}^{K} \mathsf{H}_{1,f} z_{1k} t_{fk} v_{kt} \tag{11}$$

$$\boldsymbol{R}_{ft}^{\mathrm{noise}} = \sum_{k=1}^{K} \left( \sum_{n=2}^{N} \mathsf{H}_{n,f} z_{nk} \right) t_{fk} v_{kt} \tag{12}$$

**観測ベース**　$\hat{\boldsymbol{s}}_{ft}^{(n)}$ は分離音の $n$ 音源目を表している．

$$\hat{\boldsymbol{s}}_{ft}^{(n)} = \left( \sum_{k=1}^{K} z_{nk} t_{fk} v_{kt} \right) \mathsf{H}_{n,f} \hat{\mathsf{Y}}_{ft}^{-1} \boldsymbol{y}_{ft} \tag{13}$$

すなわち，混合音の空間相関行列は，

$$\boldsymbol{R}_{ft}^{\mathrm{all}} = \left( \sum_{n=1}^{N} \hat{\boldsymbol{s}}_{ft}^{(n)} \right) \left( \sum_{n=1}^{N} \hat{\boldsymbol{s}}_{ft}^{(n)} \right)^{H} \tag{14}$$

$$= \boldsymbol{y}_{ft} \boldsymbol{y}_{ft}^{H} \tag{15}$$

目的音声が $n=1$ にあるとすれば

$$\boldsymbol{R}_{ft}^{\mathrm{speech}} = \hat{\boldsymbol{s}}_{ft}^{(1)} \hat{\boldsymbol{s}}_{ft}^{(1)H} \tag{16}$$

$$\boldsymbol{R}_{ft}^{\mathrm{noise}} = \left( \sum_{n=2}^{N} \hat{\boldsymbol{s}}_{ft}^{(n)} \right) \left( \sum_{n=2}^{N} \hat{\boldsymbol{s}}_{ft}^{(n)} \right)^{H} \tag{17}$$

(10) から (17) で求めた空間相関行列は時変だが，時間方向に平均をとれば時不変な空間相関行列を推定できる．

$$\boldsymbol{R}_{f} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{R}_{ft} \tag{18}$$

こうして MNMF を用いて求めた空間相関行列に基づきビームフォーマ (3)(4) を計算できる．

## 4. 評価実験

提案法の有効性について，実録音データの音声認識における単語誤り率 (WER) を用いて評価する．

**実験設定**　CHiME-4 [1] の実録音評価データ 1320 発話 (“et05_real_noisy”) を用いて，音声認識性能を WER で評価した．バックエンドには，マルチコンディションデータを用いて学習した DNN-HMM 音響モデルおよび WSJ 標準 5k トライフォン言語モデルを用いた．音響特徴量は対数メルフィルタバンク出力 40 次元とその Δ 及び ΔΔ である．デコーダは Kaldi WFST デコーダを用いた．マルチチャネル処理に用いるチャネル数は 5 とした．MNMF はクロススペクトル法による初期値を与えた rank-1 MNMF [6] で初期化した．

比較手法として，Beamformit [7] と DNN を用いたマスクに基づく手法 (DNNm-BF) [4] を用いた．Beamformit は CHiME-4 においてベースライン・フロントエンドの手法である．DNNm-BF は DNN で推定したマスクに基づくビームフォーマを行う手法である．また MNMF の分離結果を直接バックエンドに与える場合も評価した．

表 1: CHiME-4 実録音データでの単語誤り率 (WER)

| 音声強調手法 | | | WER[%] |
|---|---|---|---|
| 強調なし | | | 22.39 |
| Beamformit (baseline) | | | 15.60 |
| DNNm-BF | | | 11.51 |
| MNMF | | | 12.89 |
| MNMF-BF | | | |
| モデルベース | 時変 | MV | 12.89 |
| | | ML | 12.85 |
| | 時不変 | MV | 12.73 |
| | | ML | 12.73 |
| 観測ベース | 時変 | MV | - |
| | | ML | 12.94 |
| | 時不変 | MV | 13.80 |
| | | ML | 12.81 |

**実験結果**　表 1 に示すように，提案する組み合わせではモデルベースで時不変の MV/ML ビームフォーマが最も高い性能を示し，MNMF 単体よりも性能が向上した．モデルベースと観測ベースの結果の違いは観測音の直接使用が影響したと思われる．また時不変では時変よりも歪みが抑えられ性能が向上したと考えられる．全体としては ML が MV と比較して高い性能を示した．

提案法は Beamformit より 2.87pt 高い性能が得られた．DNNm-BF と比較すると 1.22pt 劣っているが，提案法は事前の学習を行わずに高い性能を達成している．

さらなる性能向上は，適切な目的音声のモデル化により期待できる．目的音源とマイクロホンの相対位置不変性を仮定できれば，rank-1 MNMF [6] が使える．時間周波数領域でスパース制約を置くことも考えられる．

## 5. おわりに

本稿では，遠隔音声認識で使用する音声強調手法として，音の混合過程を表現したブラインド音源分離の一つである MNMF に基づくビームフォーマを提案した．CHiME-4 においてベースラインである Beamformit よりも高い認識性能を示し，事前学習なしで DNNm-BF に近い性能を示した．今後は，より適切な目的音声のモデル化により性能向上を実現する．

## 参考文献


[1] J. Barker *et al.* The third CHiME speech separation and recognition challenge: Dataset, task and baselines. *IEEE ASRU*, 2015.

[2] B. D. Van Veen *et al.* Beamforming: A versatile approach to spatial filtering. *IEEE ASSP magazine*, 5(2):4–24, 1988.

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

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

[5] H. Sawada *et al.* Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE TASLP*, 21(5):971–982, 2013.

[6] D. Kitamura *et al.* Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model. *IEEE ICASSP*, 276–280, 2015.

[7] X. Anguera *et al.* Acoustic beamforming for speaker diarization of meetings. *IEEE TASLP*, 15(7):2011–2022, 2007.