非混合時間周波数ビン検出を用いた空間共分散推定と最小分散無歪ビームフォーマへの応用

Spatial correlation matrix estimation using detection of clean time-frequency bins and application to MVDR beamformer

1w173052-1　黒沢 琢登　　指導教員　及川 靖広　教授
KUROSAWA Takuto　　Prof. OIKAWA Yasuhiro

概要：音源分離は音理解のための処理の前段階として重要である．音源分離を行うための手法の一つとしてアレイ信号処理が広く用いられている．アレイ信号処理では位相差などのチャンネル間の情報を用いて空間情報を推定し，音源分離などの処理に利用することが多い．しかし，複数音が混合している場合，元のチャンネル間の関係とは無関係な情報の乱れが生じるためチャンネル間の情報の推定に悪影響を与えうる．そこで，位相微分情報を用いて，音源が混ざっていない非混合時間周波数ビンを検出する手法が提案されている [1]．本研究では，非混合な時間周波数ビンのみを用いて空間共分散を推定し，それを MVDR ビームフォーマに導入することで分離性能が向上することを確認した．

キーワード：最小分散法 (MVDR)，アレイ信号処理，空間相関行列，非混合時間周波数ビン検出
Keywords: MVDR, array signal processing, spatial correlation matrix, detection of clean time-frequency bins.

1. ま え が き

アレイ信号処理を用いた音源分離では，チャンネル間の情報を用いて推定した空間情報を利用することが多い．しかし，複数音が混合している時間周波数ビンでは，元のチャンネル間の関係とは無関係な情報の乱れが生じるためチャンネル間の情報の推定に悪影響を与えうる．本研究では，非混合な時間周波数ビンのみを用いて空間共分散を推定し，それを MVDR ビームフォーマに導入することで分離性能が向上することを確認した．

2. 非混合時間周波数ビン検出

音源の個数を N，チャンネル数を M とし，音源信号を $\boldsymbol{s} = (s_1, \cdots, s_N)$，観測信号を $\mathbf{x} = (x_1, \cdots, x_M)$ と表す．チャンネル m で音源 s_n のみが観測されたとき，時間周波数領域 ($[t, \omega]$ 領域) の観測信号 x_m は

$$x_m[t,\omega] = h_{m,n}[\omega] s_n[t,\omega] \tag{1}$$

と表せる．ただし，$h_{m,n}$ は音源 s_n の位置とチャンネル m 間の伝達関数である．式 (1) は複素数であり，これを極形式で表現すれば観測信号の位相 φ_{x_m} は

$$\varphi_{x_m}[t,\omega] = \varphi_{h_{m,n}}[\omega] + \varphi_{s_n}[t,\omega] \tag{2}$$

となる．ここで，位相を時間微分した瞬時周波数 IF は

$$\mathrm{IF}_{x_m}[t,\omega] = \partial_t \varphi_{h_{m,n}}[\omega] + \partial_t \varphi_{s_n}[t,\omega] \tag{3}$$

と表せる．位相微分値は，対数振幅スペクトログラムが滑らかであればチャンネル m によらず全チャンネルでほぼ同じ値をとることが知られており，式 (3) は

$$\mathrm{IF}_{x_m}[t,\omega] \approx \partial_t \varphi_{s_n}[t,\omega] + C \tag{4}$$

と表せる．ただし，C は m に依存しない定数である．ここで $\partial_t \varphi_{s_n}$ は m に依存しないため，音源が混ざっていない非混合な時間周波数ビンにおける IF は全チャンネルでほぼ同じ値をとると考えられる．一方で，混合なビンでは位相が乱れることによりチャンネル間で IF の値にばらつきが生じるため，チャンネル間の標準偏差

$$\sigma_M(\mathbf{x}[t,\omega]) = \left(\frac{1}{M} \sum_{m=1}^{M} \left(\mathrm{IF}_{x_m}[t,\omega] - \overline{\mathrm{IF}_{\mathbf{x}}}[t,\omega] \right)^2 \right)^{1/2}$$

の計算により各時間周波数ビンの混合具合を評価できる．ただし，$\overline{\mathrm{IF}_{\mathbf{x}}}[t,\omega]$ は IF のチャンネル間の平均値であり，σ_M は混合であるほど大きい値をとる．

3. 提 案 手 法

σ_M について，各周波数における k 番目に小さい値を $\delta_k[\omega]$ とする．ここで以下の式を定義する．

$$\mathcal{P}_k[t,\omega] = \begin{cases} 1 & (\sigma_M[t,\omega] \leqq \delta_k[\omega]) \\ 0 & (\text{otherwise}) \end{cases} \tag{5}$$

σ_M は非混合であるほど小さい値をとるため，$\mathcal{P}_k$ は非混合と考えられる順に取った k 個のビンでのみ 1 を，その他のビンでは 0 をとる．よって，観測信号 $\mathbf{x}$ に対し

$$\check{\mathbf{x}}[t,\omega] = \mathcal{P}_k[t,\omega] \mathbf{x}[t,\omega] \tag{6}$$

として，観測信号 $\mathbf{x}$ から非混合と考えられる順に k 個のビンを取り出した，推定信号 $\check{\mathbf{x}}$ を得る．ここで，目的信号を除去する時間周波数マスクを $\boldsymbol{\mathcal{M}}$ として

$$\hat{\mathbf{x}}[t,\omega] = \boldsymbol{\mathcal{M}}[t,\omega] \odot \check{\mathbf{x}}[t,\omega] \tag{7}$$

としたうえで $\hat{\boldsymbol{R}} = E[\hat{\mathbf{x}}\hat{\mathbf{x}}^H]$ とする．ただし $\hat{\boldsymbol{R}}$ は空間共分散であり，$E[\cdot]$ は期待値を表し，実際の計算ではサン

プル平均で代用する．これにより，非混合と考えられる順に取った k 個のビンのみを用いたうえで，目的音除去マスクを適用した空間共分散推定を行う．$\hat{\boldsymbol{R}}$ を用いて

$$\hat{\boldsymbol{w}} = \frac{\hat{\boldsymbol{R}}^{-1}\boldsymbol{a}}{\boldsymbol{a}^H \hat{\boldsymbol{R}}^{-1}\boldsymbol{a}} \tag{8}$$

とすることで MVDR ビームフォーマを設計し，これを

$$y[t,\omega] = \hat{\boldsymbol{w}}^H \mathbf{x}[t,\omega] \tag{9}$$

として用いることで，観測信号 $\mathbf{x}$ から目的信号を分離した結果 y を得る．ここで，マスクは $\hat{\boldsymbol{w}}$ の推定にのみ用いられており，y 自体には直接マスクをかけないため線形なフィルタリングである点に注意されたい．

4. 数 値 実 験

本稿では目的音源除去マスク $\boldsymbol{\mathcal{M}}$ について，以下に示す 3 種類の理想マスクおよび 3 種類の推定マスクを使用した．理想マスクにおいて，チャンネル m での観測信号について目的音源のみが観測された場合を $v_m[t,\omega]$，目的音源以外が観測された場合を $n_m[t,\omega]$，全音源が観測された場合を $x_m[t,\omega]$ とする．また，推定マスクにおいて，遅延和（DS）ビームフォーマを用いて作成した目的音源強調信号を $\mathcal{V}$，ブロッキング行列を用いて作成した目的音源除去信号を $\mathcal{N}$ とする．

- 理想比マスク (IRM)

$$\mathrm{IRM}_m[t,\omega] = \left(\frac{|n_m[t,\omega]|^2}{|v_m[t,\omega]|^2 + |n_m[t,\omega]|^2} \right)^{1/2} \tag{10}$$

- スペクトル振幅マスク (SMM)

$$\mathrm{SMM}_m[t,\omega] = \frac{|x_m[t,\omega]| - |v_m[t,\omega]|}{|x_m[t,\omega]|} \tag{11}$$

- 目的信号完全除去マスク (FRM)

$$\mathrm{FRM}_m[t,\omega] = \frac{n_m[t,\omega]}{x_m[t,\omega]} \tag{12}$$

- 推定比マスク (ERM)

$$\mathrm{RM}_m[t,\omega] = \left(\frac{|\mathcal{N}[t,\omega]|^2}{|\mathcal{V}[t,\omega]|^2 + |\mathcal{N}[t,\omega]|^2} \right)^{1/2} \tag{13}$$

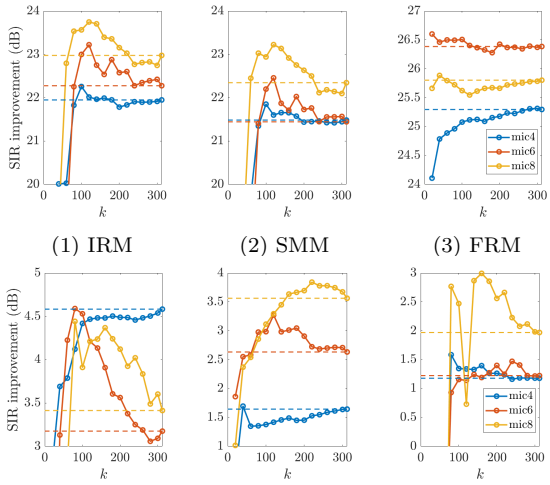
- 推定スペクトル振幅マスク (ESMM)

$$\mathrm{ESMM}_m[t,\omega] = \frac{|x_m[t,\omega]| - |\mathcal{V}[t,\omega]|}{|x_m[t,\omega]|} \tag{14}$$

- 推定バイナリマスク (EBM)

$$\mathrm{EBM}_m[t,\omega] = \begin{cases} 1 & (|\mathcal{V}[t,\omega]| > |x_m[t,\omega][\omega]|) \\ 0 & (\text{otherwise}) \end{cases} \tag{15}$$

これらのうちいずれかのマスクを $\boldsymbol{\mathcal{M}}$ として式 (7)，式 (8)，式 (9) を計算し，提案手法を用いた音源分離を行う．なお，本実験では 2 音源を $\{-30,45\}$ 度から流し $\{8,6,4\}$ チャンネルの線形マイクロフォンアレイで観測した．サンプリング周波数は 16 kHz，窓長は 128 ms，ずらし幅は 32 ms とした．インパルス応答は残響時間 160 ms，マイク間隔 8 cm，マイクと音源の距離 2 m のデータ [1] を使用し，音源は日本語音声データ [2] を使用した．また，観測スペクトログラムの時間ビンは 312 個であり，その



(1) IRM　(2) SMM　(3) FRM

(4) ERM　(5) ESMM　(6) EBM

図–1　数値実験結果（実線：提案手法，点線：従来手法）

うち上位 $k = \{300, 280, \ldots, 20\}$ 個のビンを取り出すマスクを作成した．$k = 312$ の場合はすべてのビンを取り出しているため，式 (7) において $\mathcal{P}_k$ をかけない従来手法である．提案手法の有効性を評価するため，従来手法と提案手法の観測音源からの SIR 改善量を比較した．

実験結果を図–1 に示す．図中の点線は従来手法の SIR 改善量を表している．k の値を減らし提案手法を適用した場合，理想マスクでは全てのマイク数において従来手法を上回る SIR 改善量を示した．しかし推定マスクでは，マイク数が 4 のとき SIR 改善量の向上が見られない場合があった．これは，目的音除去マスクの作成に用いた DS ビームフォーマやブロッキング行列を用いた手法が空間情報を用いない手法であり，実験データやマイク数などの条件によっては空間特性による悪影響がもたらされたためであると考えられる．今後は空間情報を用いたより安定性のある手法を用いて目的音除去マスクを推定し，同様の実験を行う必要があるという課題が挙げられる．また，最も効果的である k の値がマスクやマイク数等の条件ごとに異なるといった課題も挙げられる．

5. む　す　び

本研究では，非混合時間周波数ビンのみを利用し，空間共分散を推定する手法を提案した．さらに，非混合ビンのみを利用することで，音源分離の性能が向上しうることを確認した．今後は時間周波数マスクを用いて空間共分散を推定する手法に，非混合時間周波数ビン検出を利用することで分離性能の改善が見られるかを検討していく所存である．

参 考 文 献

[1] A. Hiruma, K. Yatabe and Y. Oikawa, “Detection of clean time-frequency bins based on phase derivative of multichannel signals, ” Int. Congr. Acoust. (ICA), 2019.

[1] https://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/multi-channel-impulse-response-database/

[2] https://voice.mozilla.org/ja/datasets