複素対数補間によるヴァーチャル観測に基づく劣決定条件での音声強調*

☆片平拓希 (筑波大), 小野順貴 (NII),
宮部滋樹, 山田武志, 牧野昭二 (筑波大)

1 はじめに

マイクロホンアレーを用いた音声のアレー信号処理は、複数のセンサ (マイクロホン) を用いて観測された音響信号から、様々な処理や推定を行うものである [1]。マイクロホンアレー信号処理は、音源の位置を推定する音源定位、複数の音源が混合された観測信号を分離する音源分離などのように、目的に応じて様々な方法が検討されており、特に線形アレー信号処理は確立された枠組である。線形アレー信号処理の例として、ブラインド音源分離の代表的手法として知られる独立成分分析 (ICA: Independent Component Analysis) [2]、音源定位の手法である MUSIC(MUltiple SIgnal Classification) [3] などがある。しかし、これらの線形アレー信号処理は、対象とする音源の数がマイクロホン素子数 (観測チャネル数) に対して小さい条件では高い性能を発揮するものの、素子数に比べて音源数が大きい、いわゆる劣決定条件では、性能が大幅に劣化してしまう。そのため、多数の音源を扱う場合、大規模なアレーや多チャネル A/D 変換器が必要となってしまう。劣決定条件のための線形アレー信号処理の拡張として、クロネッカー積 [4] やカーネル関数 [5] を用いて観測信号を非線形に高次元写像することにより、観測信号を擬似的に多素子化する手法が提案されている。しかし、これらの手法における写像は信号の性質を大幅に変えてしまうため、音声強調に用いると出力の歪が大きくなるという問題がある。

本研究では、実際にはマイクロホンが存在しない位置における観測信号を、実際の観測信号の非線形補間により擬似的に生成し、ヴァーチャルな観測信号として扱うことにより、処理出力に大きな歪を生じない擬似的多素子化を試みる。この擬似的多素子化を SN 比最大化ビームフォーマ [6, 7] に導入し、その性能を検証することにより、提案するヴァーチャル観測の有用性を検討する。

2 問題設定

本稿では、信号を時間周波数領域で表すものとし、観測信号を以下のようにモデル化する。

$$\begin{aligned}\mathbf{x}(\omega,t) &= \Big[x_1(\omega,t),\cdots,x_M(\omega,t)\Big]^{\mathrm{T}} \\ &\approx \sum_{i=1}^{N}\mathbf{a}_i(\omega)s_i(\omega,t) \qquad (1)\\ \mathbf{a}_i(\omega) &= \Big[a_{1,i}(\omega),\cdots,a_{M,i}(\omega)\Big]^{\mathrm{T}} \qquad (2)\end{aligned}$$

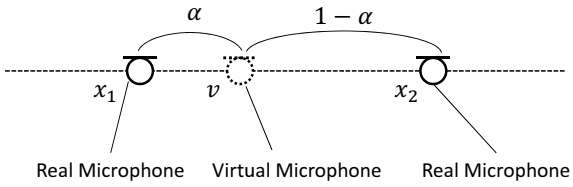

Fig. 1 Arrangement of virtual observations

ここで、ω は周波数、t は時間フレーム番号、M は観測チャネル数、N は音源数である。また、s_i は i 番目の音源信号、x_j は j 番目のマイクロホン素子における観測信号、$a_{j,i}$ は i 番目の音源から j 番目の素子への伝達関数を表す。

適応ビームフォーマを用いた音声強調では、観測信号 $\mathbf{x}(\omega,t)$ 中の非目的音を抑圧し、目的信号のみを強調した信号 $y(\omega,t)$ を得るための時間周波数領域マルチチャネルフィルタ $\mathbf{w}(\omega)$ を設計する。

$$y(\omega,t)=\mathbf{w}(\omega)\,\mathbf{x}(\omega,t) \qquad (3)$$

理想的には、目的音源のインデックスを i_{T} としたときに、すべての $\mathbf{a}_i(i\neq i_{\mathrm{T}})$ と直交して

$$\mathbf{w}(\omega)\,\mathbf{a}_i(\omega)=0,\ \ \forall i\neq i_{\mathrm{T}} \qquad (4)$$

となる $\mathbf{w}(\omega)$ を構成することができれば、

$$\begin{aligned}y(\omega,t) &= \mathbf{w}(\omega)\sum_{i=1}^{N}\mathbf{a}_i(\omega)\,s_i(\omega,t)\\ &= \mathbf{w}(\omega)\,\mathbf{a}_{i_{\mathrm{T}}}(\omega)\,s_{i_{\mathrm{T}}}(\omega,t) \qquad (5)\end{aligned}$$

となり目的信号のみを取り出すことができる。しかし、観測信号ベクトルの次元数が音源数より小さい場合、このような $\mathbf{w}(\omega)$ は一般には存在しえず、ビームフォーマの性能は大幅に劣化する。そのため、多数の音源の混合した音声をビームフォーマで扱う場合、多チャネルの観測信号が必要となる。

3 複素対数補間によるヴァーチャル多素子化

3.1 ヴァーチャル観測値の定式化

本節ではヴァーチャル観測値の具体的な定式化について述べる。ここでは、実際にはマイクを配置していない場所に新たにマイクを配置した場合に、どのような信号が観測されるかを実際に観測された信号から擬似的に生成し、いわばヴァーチャルな観測値を求

*Speech Enhancement in Underdetermined Condition Based on Virtual Observations Derived from Complex Logarithmic Interpolation. by Hiroki KATAHIRA (University of Tsukuba), Nobutaka ONO (National Institute of Informatics), Shigeki MIYABE, Takeshi YAMADA, Shoji MAKINO (University of Tsukuba)

める。ここではまず、最も簡単な場合として、空間上に 2 つのマイクロホン素子が配置された条件を考える。ここで、2 素子を結んだ直線上の任意の位置での観測値を、実際の 2 素子の観測信号から、関数 f を用いた以下の様な補間でヴァーチャルに合成することを考える (Fig. 1)。

$$v(x_1, x_2, \alpha) = f^{-1}\{(1-\alpha) f(x_1) + \alpha f(x_2)\} \quad (6)$$

ここで、f は補間を行うドメインを決定する関数、α はヴァーチャル観測の位置を表すパラメータである。本稿では、$0 < \alpha < 1$ に限定し、内挿の場合のみを考える。ここでの目的は、後段で適用する線形アレー信号処理に対して、線形独立な観測信号を擬似的に増やすことであるため、f は非線形関数である必要がある。(線形関数では線形従属な観測信号しか得られない。) この非線形関数 f の選択が、本提案法で最も重要となる。

そこで、適切な補間関数の選択について考える。ここでは、最も単純な場合として、単一の平面波のみが到来する場合を考える。この場合、x_1, x_2 の間には到来時間差に由来する位相差のみが生じ、これはマイクロホン間の距離に線形であるから、2 素子間を $\alpha : (1-\alpha)$ で内分する点での位相は次式で表せる。

$$\angle v = (1-\alpha)\angle x_1 + \alpha \angle x_2 \quad (7)$$

これを式 (6) で表したような非線形な補間として表すためには、位相が複素数 z に対する複素対数

$$\log z = \log|z| + j\angle z \quad (8)$$

の虚部として定義されることに注意すれば、

$$f(x) = \log x \quad (9)$$
$$f^{-1}(x) = \exp(x) \quad (10)$$

のように、$f(x)$ に複素対数関数を選べばよいことがわかる。このとき、複素対数の実部にあたる振幅についても、対数領域で補間することになるが、この妥当性について考える。平面波の場合には、距離減衰がなく振幅は一定であるため、対数領域の補間を考えても問題は生じない。一方、球面波の場合には、正確な振幅の補間は音源方向に依存し、ある 2 点の観測信号のみから、その間における振幅を単一の関数で正確に補間することはできないことがわかる。しかしここでは、$f(x)$ が連続で微分可能となる性質のよさや式の単純さから、$f(x)$ として複素対数関数を選び、振幅に対しても位相と同じく、対数領域で統一的に補間を行うことを考え、モデル化誤差等の影響等は実験で検証することとした。このとき、全体の補間式は以下のようになる。

$$v(x_1, x_2, \alpha) = \exp\{(1-\alpha)\log|x_1| + \alpha \log|x_2| + j((1-\alpha)\angle x_1 + \alpha \angle x_2)\} \quad (11)$$

また実際には、複数の波面が混合して観測される場合においても、音声のスパース性により多くの時間周波数において 1 つの音源信号のみが支配的であると考えることができることから、これらの補間は妥当であると考えることができる。つまり本手法は、スパース性が成り立っていれば近似的に有効な補間を導入して観測信号を多素子化するアプローチと位置づけることができると考えられる。

3.2 位相の線形補間

本手法では、式 (11) のように、観測信号の複素対数について線形補間を行なっている。しかし、その虚部である位相には $\pm 2n\pi$ の任意性があるため、その線形補間についても同様に任意性が発生することに注意する必要がある。ここでは、空間的エイリアシングが発生せず、2 つの観測信号の位相差が π を超えない、つまり

$$|\angle x_1 - \angle x_2| \leq \pi \quad (12)$$

のように仮定して、位相の線形補間を行う。

3.3 SN 比最大化ビームフォーマ

本研究ではヴァーチャル多素子化の有効性を検証するため、従来の線形アレー信号処理の一つである SN 比最大化ビームフォーマにこれを適用した。そこでまず、従来の SN 比最大化ビームフォーマについてまとめる。

SN 比最大化ビームフォーマは、観測信号中の目的音声区間と非目的音声区間それぞれの共分散行列を事前情報として与え、目的音声を強調する手法である。事前情報としてステアリングベクトルなどの方向情報を明示的に与える必要がなく、音源位置が未知の場合においても適応できるといった利点がある。

SN 比最大化ビームフォーマは、出力信号中の目的信号と非目的信号のパワー比を最大化するように指向特性を形成する。ここでパワー比 $\lambda(\omega)$ は

$$\lambda(\omega) = \frac{\mathbf{w}(\omega)\mathbf{R}_{\mathrm{T}}(\omega)\mathbf{w}^{\mathrm{H}}(\omega)}{\mathbf{w}(\omega)\mathbf{R}_{\mathrm{I}}(\omega)\mathbf{w}^{\mathrm{H}}(\omega)} \quad (13)$$

のように表される。なお、$\mathbf{R}_{\mathrm{T}}$ は目的信号区間、$\mathbf{R}_{\mathrm{I}}$ は非目的信号区間それぞれの共分散行列であり、

$$\mathbf{R}_{\mathrm{T}}(\omega) = \frac{1}{|\Theta_{\mathrm{T}}|}\sum_{t\in\Theta_{\mathrm{T}}} \mathbf{x}_{\mathrm{T}}(\omega, t)\mathbf{x}_{\mathrm{T}}^{\mathrm{H}}(\omega, t) \quad (14)$$
$$\mathbf{R}_{\mathrm{I}}(\omega) = \frac{1}{|\Theta_{\mathrm{I}}|}\sum_{t\in\Theta_{\mathrm{I}}} \mathbf{x}_{\mathrm{I}}(\omega, t)\mathbf{x}_{\mathrm{I}}^{\mathrm{H}}(\omega, t) \quad (15)$$

と表される。ここで、Θ_{T} は目的信号区間、Θ_{I} は非目的信号区間のそれぞれ時間フレームの集合である。このパワー比 $\lambda(\omega)$ を最大化する $\mathbf{w}(\omega)$ は、以下の一般化固有値問題の最大固有値に対応する固有ベクトルに相当する。

$$\mathbf{w}(\omega)\mathbf{R}_{\mathrm{T}}(\omega) = \lambda(\omega)\mathbf{w}(\omega)\mathbf{R}_{\mathrm{I}}(\omega) \quad (16)$$

しかし一般化固有値問題にはノルムに関する不定性があるので、スケールを決定するために次のように

ビームフォーマを補正する [7]。

$$\mathbf{w}(\omega) \leftarrow b_k(\omega)\mathbf{w}(\omega) \tag{17}$$

ただし、$b_k(\omega)$ は

$$\mathbf{R}_{\mathbf{x}}(\omega) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}(\omega,t)\mathbf{x}^{\mathrm{H}}(\omega,t) \tag{18}$$

として以下のように定義される $\mathbf{b}(\omega)$ の任意の k 次元目の要素である。なお、T は全観測区間の時間フレーム数である。

$$\mathbf{b}(\omega) = \frac{\mathbf{w}(\omega)\mathbf{R}_{\mathbf{x}}(\omega)}{\mathbf{w}(\omega)\mathbf{R}_{\mathbf{x}}(\omega)\mathbf{w}^{\mathrm{H}}(\omega)} \tag{19}$$

SN 比最大化ビームフォーマは線形アレー信号処理の枠組であり、扱う音源の数が観測チャネル数より大きい場合において、その性能は大幅に低下する。そこで本稿では、次節で述べるように、ヴァーチャル観測値を導入し、観測チャネル数を擬似的に増やすことにより、音源数の多い劣決定条件における SN 比最大化ビームフォーマの性能の改善を図る。

3.4 SN 比最大化ビームフォーマの拡張

前節で述べた SN 比最大化ビームフォーマの、ヴァーチャル観測値を用いた拡張について述べる。ここでは、N_{v} 個のヴァーチャル観測を素子間に等間隔に配置する条件を考え、n 番目のヴァーチャル観測値を次式のように表す。

$$v_n = v\left(x_1, x_2, \frac{n}{N_{\mathrm{v}}+1}\right) \tag{20}$$

また、ヴァーチャル観測を挿入して擬似的に多素子化した観測値を次のように表す。

$$\phi(\mathbf{x}) = \begin{bmatrix} x_1, & v_1, & \cdots, & v_{N_{\mathrm{v}}}, & x_2 \end{bmatrix}^{\mathrm{T}} \tag{21}$$

ヴァーチャル観測を含む観測値の目的信号区間、非目的信号区間それぞれの共分散行列は

$$\widehat{\mathbf{R}}_{\mathrm{T}}(\omega) = \frac{1}{|\Theta_{\mathrm{T}}|}\sum_{t\in\Theta_{\mathrm{T}}}\phi(\mathbf{x}_{\mathrm{T}}(\omega,t))\phi(\mathbf{x}_{\mathrm{T}}(\omega,t))^{\mathrm{H}} \tag{22}$$

$$\widehat{\mathbf{R}}_{\mathrm{I}}(\omega) = \frac{1}{|\Theta_{\mathrm{I}}|}\sum_{t\in\Theta_{\mathrm{I}}}\phi(\mathbf{x}_{\mathrm{I}}(\omega,t))\phi(\mathbf{x}_{\mathrm{I}}(\omega,t))^{\mathrm{H}} \tag{23}$$

のように表される。これらの共分散行列を用いて一般化固有値問題

$$\widehat{\mathbf{w}}(\omega)\widehat{\mathbf{R}}_{\mathrm{T}}(\omega) = \lambda(\omega)\widehat{\mathbf{w}}(\omega)\widehat{\mathbf{R}}_{\mathrm{I}}(\omega) \tag{24}$$

における最大固有値に対応する固有ベクトルを求めることで拡張 SN 比最大化ビームフォーマ $\widehat{\mathbf{w}}(\omega)$ を得る。また、式 (17)–(19) と同様に、次式で表される $\widehat{\mathbf{b}}(\omega)$ の任意の k 番目の要素 $\widehat{b}_k(\omega)$ を用いて一般化固有値問題のスケールを決定しビームフォーマを補正する。

$$\widehat{\mathbf{w}}(\omega) \leftarrow \widehat{b}_k(\omega)\widehat{\mathbf{w}}(\omega) \tag{25}$$

$$\widehat{\mathbf{b}}(\omega) = \frac{\widehat{\mathbf{R}}_{\mathbf{x}}(\omega)\widehat{\mathbf{w}}^{\mathrm{H}}(\omega)}{\widehat{\mathbf{w}}(\omega)\widehat{\mathbf{R}}_{\mathbf{x}}(\omega)\widehat{\mathbf{w}}^{\mathrm{H}}(\omega)} \tag{26}$$

$$\widehat{\mathbf{R}}_{\mathbf{x}}(\omega) = \frac{1}{T}\sum_{t=1}^{T}\phi(\mathbf{x}(\omega,t))\phi(\mathbf{x}(\omega,t))^{\mathrm{H}} \tag{27}$$

Table 1　Experimental conditions

素子数	2
音源数	3
混合音パターン数	10 パターン
音源到来方向	(正面: 0°, 時計回り) $-40°, 20°, 50°$
素子間隔	2.15 cm
残響時間	130 ms
サンプリング周波数	8 kHz
フレーム長	1024 samples
シフト長	1/4 フレーム
音声長	20 sec
目的音区間長	5 sec
非目的音区間長	5 sec
正則化項の係数	$\varepsilon = 10^{-14}, 10^{-12}$ $10^{-10}, 10^{-8}, 10^{-6}, 0$

これを

$$\widehat{y}(\omega,t) = \widehat{\mathbf{w}}(\omega)\phi(\mathbf{x}(\omega,t)) \tag{28}$$

のように、観測信号に適用し強調音声を得る。

なお、一般化固有値問題 (24) においては、ランク落ちによる数値誤差の発生が予想される。ここでは、正則化項として、非目的音共分散行列 $\widehat{\mathbf{R}}_{\mathrm{I}}$ に単位行列 $\mathbf{E}$ の正数 ε 倍を加算することで次式のように補正する。

$$\widehat{\mathbf{R}}_{\mathrm{I}} \leftarrow \widehat{\mathbf{R}}_{\mathrm{I}} + \varepsilon\mathbf{E} \tag{29}$$

4 実験

ヴァーチャル観測を導入した SN 比最大化ビームフォーマの性能を検証するため、混合音声に対する分離性能の比較実験を行った。比較手法として、従来のヴァーチャル観測を導入しない SN 比最大化ビームフォーマを用いた。

4.1 実験条件

Table 1 に実験条件を示す。各音源にインパルス応答を畳み込んだものを混合する形で、シミュレーションによって混合音声を作成した。音源には男女の発話音声を用いた。また、目的音声、非目的音声のみが存在する、それぞれ 5 秒間の時間区間を用いて共分散行列を求め、SN 比最大化ビームフォーマを作成した。分離性能の評価には客観評価値を用いた。評価値は音声の組み合わせについて平均を取ったものである。評価値としては、信号対歪比 (SDR: Signal-to-Distortion Ratio)、信号対干渉比 (SIR: Signal-to-Interference Ratio) [8] を用いた。これらの評価値は、数値が高い方が高い性能を表す。

4.2 結果と考察

Fig. 2 には、ヴァーチャル観測チャネル数を 0 – 9 で変化させたときの、SDR および SIR の変化を示す。

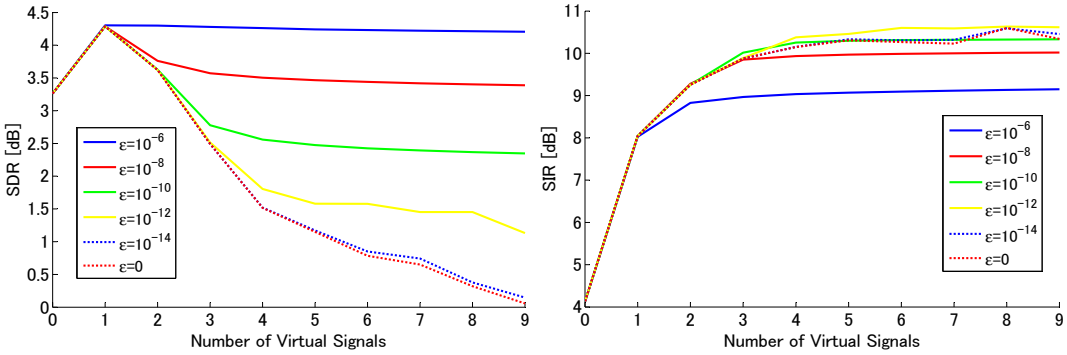

Fig. 2　Evaluation values for maximum SNR beamformers with virtual observation signals

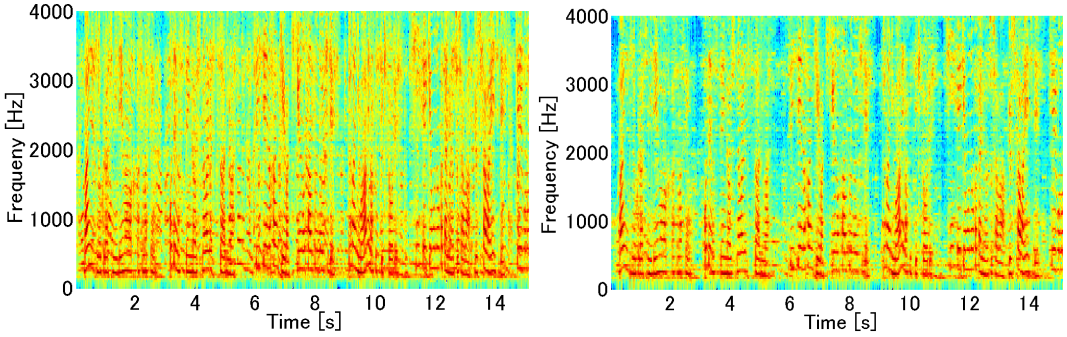

Fig. 3　Spectrograms of separated signal: without virtual signal (left), with virtual signals (right)

また、分離音声のスペクトログラムの例を Fig. 3 に示す。共分散行列の正則化をしない条件 ($\varepsilon = 0$) においては、多素子化するとともに SDR が低下しているのに対し、$\varepsilon = 10^{-6}$ の条件においては、全体的に改善が見られ、正則化により信号の歪を抑えることができると言える。SIR に関しては ε を小さく設定した方が高い値を示している。これは、正則化によりランク落ち誤差が抑えられる反面、ビームフォーマの学習に誤差が生じ、非目的音の抑圧性能に影響が出ているものと考えられる。ただ、ヴァーチャル観測チャネル数 4 の条件で、ヴァーチャル観測値なしの条件に比べて 5 dB 程度上昇しており、SIR についても改善が見られる。また、ヴァーチャル観測なしの状態から 1 チャネル加えた場合での SDR,SIR の上昇は顕著であり、特に SIR においては、約 4 dB と大幅な改善が見られる。このことから、提案手法による拡張 SN 比最大化ビームフォーマは、非目的音成分の抑圧という点で特に高い性能を示すと言える。

5　まとめ

本稿では、ヴァーチャル観測値により、観測信号を仮想的に多素子化し、SN 比最大化ビームフォーマにこれを導入して、劣決定条件における音声強調の性能改善を図った。ヴァーチャル観測値の導入により、分離性能の改善が見られ、SN 比最大化ビームフォーマへのヴァーチャル観測の導入の有用性が確認できた。今後は、他の補間関数を用いたヴァーチャル観測値を導入し、本稿の複素対数補間との性能比較を行う予定である。また、他の線形アレー信号処理にヴァーチャル観測を導入し、その効果を検証する予定である。

参考文献

[1] 浅野 太, “音のアレイ信号処理 —音源の定位・追跡と分離—,” コロナ社, 2011.
[2] 澤田他, “音源分離技術の最新動向,” 信学誌, 91(4), 292–296, 2008.
[3] R. O. Schmidt, “Multiple emitter location and signal parameter estimation,” IEEE Trans. Antennas and Propagation, 34(3), 1986.
[4] Chevarier *et al.*, IEEE Trans. Signal Process., 53 (4), 1254–1271, 2005.
[5] S. Miyabe, *et al.*, “Kernel-based nonlinear independent component analysis for underdetermined blind source separation,” Proc. ICASSP, 1641–1644, 2008.
[6] H. L. Van Trees, ed., “Optinum Array Processing,” Wiley, 2002.
[7] 荒木他, “話者分類と SN 比最大化ビームフォーマに基づく会議音声強調,” 音講論 (春), 571–572, 2007.
[8] E. Vincent, *et al.*, “Performance measurement in blind audio source separation,” IEEE Trans. ASLP, 14(6), 1462–1469, 2006.