βダイバージェンスに基づく一般化振幅補間によるヴァーチャル多素子化を用いた目的音源強調*

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

1 はじめに

近年、音声通信や音声認識などの需要の高まりにより、ビームフォーマをはじめとした、目的音強調技術が広く研究されている。目的音強調の代表的なアプローチとして、マイクロホンアレーなど複数マイクロホンによる録音から音の空間情報などを利用することが挙げられる。このようなマイクロホンアレーを用いた手法の多くは、多チャネル録音を対象としたものであり、その性能は録音のチャネル数 (マイクロホン数) 依存する。つまり、少ないチャネル数の録音に対してこのような手法を適用しても、十分な性能が得られない場合が多い。対して、携帯電話の内蔵マイクや IC レコーダーといった小型録音機器は近年広く普及しており、限られたチャネル数に対応した目的音強調の枠組みの開発が求められている。

このような少ないチャネル数での目的音強調を高性能化する枠組みとして、我々はこれまでに「ヴァーチャルマイクロホン」を提案してきた [1, 2]。これは、実際にはマイクロホンの置かれていない位置での録音を推定する形でマイクロホンアレーを擬似的に多素子化する枠組みである。本研究では、2 本の実マイクロホンによる録音を元に、任意のチャネル数のヴァーチャルマイク信号を合成する手法を提案する。

「ヴァーチャルマイクロホン」の語は、高次元統計量を導入した擬似的多素子化 [3] や空間音響収録 [4, 5] などの分野で用いられる。ただし、高次統計量による擬似的多素子化においては、音声信号とは異なる性質を持つ高次元特徴量を用いて信号処理を行うことから、目的音強調に適用した場合、出力音声のひずみが大きくなるといった問題がある。また、空間音響収録におけるヴァーチャルマイク信号は実マイク信号の線形結合で構成されるため、非線形ひずみは発生しないものの、アレー信号処理の入力として有用な情報とはならず、信号処理の多素子化のような用途には適さない。これらに対して、本研究におけるヴァーチャルマイクロホンは、特徴量などではなく、録音自体のチャネル数を擬似的に増加させることで、アレー信号処理の入力信号の多チャネル化による性能改善を目的とする。

ヴァーチャルマイク信号の導出として、我々はこれまでに複素スペクトルの対数をとって補間する手法を提案した [1, 2]。また、ヴァーチャルマイクロホンアレーを SN 比最大化ビームフォーマによる目的音強調に適用し、性能の向上を確認した。本稿では、この複素対数補間によるヴァーチャル多素子アレーの拡張として、β ダイバージェンスに基づく補間を提案する。この拡張により、新たなパラメータ β が導入され、補間の非線形性の程度の調整が可能となる。なお、 β の値により、β ダイバージェンスを用いた補間はこれまでの複素対数補間を内包する。また、本稿では、導入されたパラメータ β の様々な値を用いたヴァーチャルマイクロホンアレーを構成し、β と目的音強調性能の関係を検証する。

2 補間によるヴァーチャルマイク信号

我々の提案するヴァーチャルマイクロホンでは、2 チャネルの実マイクロホン信号から任意のチャネル数のヴァーチャルマイク信号を生成し、実マイク信号、ヴァーチャルマイク信号双方からなる多素子化録音信号に信号処理を施す (図 1)。ヴァーチャルマイク信号は実際にはマイクロホンの置かれていない位置での録音信号の推定として生成され、ヴァーチャルマイク信号 $v = v(\omega, t, \alpha)$ を実マイクロホン位置を $\alpha : (1-\alpha)$ に内分する点での録音信号として定義する。なお、信号は短時間フーリエ変換による時間周波数領域で表され、$v(\omega, t, \alpha)$ は、周波数ビン ω、時間フレーム t での複素振幅を表す。最もシンプルなアプローチとして、次式のような線形補間が考えられる。

$$v = (1-\alpha)\, x_1 + \alpha x_2 \tag{1}$$

$x_i = x_i(\omega, t)$ は i 番目の実マイクロホンによる録音信号である。ここで、このような線形補間から生成されるヴァーチャルマイク信号と実マイクロホン信号が線形従属となってしまうため、信号処理に用いる際に有用な情報とはなりえない。そのため、ヴァーチャルマイク信号の合成には非線形関係を取り入れる必要があり、これまでに我々は、信号の複素対数ドメインでの補間 [2] を提案した。この複素対数補間は、次式のように表される。

$$v = \exp\left((1-\alpha)\log x_1 + \alpha \log x_2\right) \tag{2}$$

ここで、複素対数の実部と虚部には信号の対数振幅と位相がそれぞれ次式のように現れる。

$$\log x_i = \log |x_i| + j\angle x_i \tag{3}$$

このため、式 (2) の複素対数補間は $A_i = |x_i(\omega, t)|$ と $\phi_i = \angle x_i(\omega, t)$ をそれぞれ i チャネル目の信号の振幅と位相として次のように表すことができる。

$$A_{\mathrm{v}} = \exp\left((1-\alpha)\log A_1 + \alpha \log A_2\right) \tag{4}$$

$$\phi_{\mathrm{v}} = (1-\alpha)\,\phi_1 + \alpha\phi_2 \tag{5}$$

$$v = A_{\mathrm{v}} \exp(j\phi_{\mathrm{v}}) \tag{6}$$

ここで、式 (5) より位相については線形補間していることがわかる。平面波の位相はマイク位置 α に対して線形に変化するため、このような線形補間は適切であると考えられる。

---

*Speech enhancement with virtual microphone array by generalized amplitude interpolation based on beta-divergence.
by Hiroki KATAHIRA (University of Tsukuba), Nobutaka ONO (National Institute of Informatics / The Graduate University for Advanced Studies (Sokendai)), Shigeki MIYABE, Takeshi YAMADA, Shoji MAKINO (University of Tsukuba)

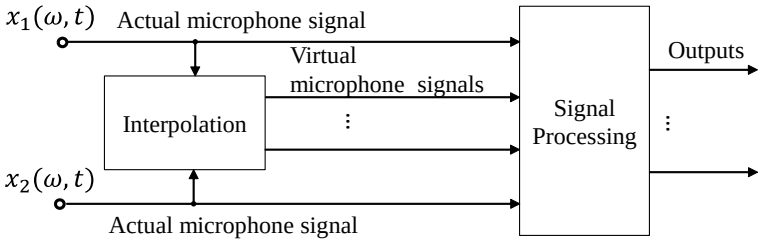


Fig. 1: Block diagram of signal processing with virtual microphone array technique

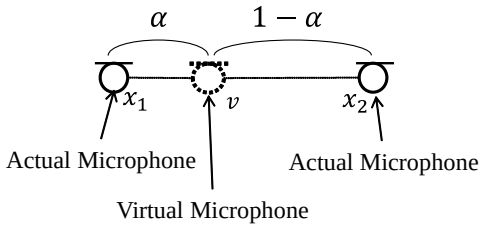


Fig. 2: Arrangement of actual and virtual microphones

3 β ダイバージェンス導入による振幅補間の一般化

前節で述べたように、位相の線形補間は、平面波の伝搬の性質に合致する。対して、式 (4) での対数振幅補間は特定のモデルを仮定したものではなく、演算、数式の単純性という観点で導入されたものである。そのため、この振幅補間の部分には拡張、改善の余地があると考えられる。そこで本節では、振幅補間に β ダイバージェンスを導入した拡張、一般化を考えるものとする。

β ダイバージェンスは非負値同士の間に定義される距離関数であり、非負値行列因子分解 (NMF) のコスト関数などとしてしばしば用いられる [6, 7]。ヴァーチャルマイク信号の振幅 A_{v} と i チャネル目の実マイク信号の振幅 A_i の間の β ダイバージェンス $D_\beta(A_{\mathrm{v}}, A_i)$ は次のように定義される。

$$D_\beta(A_{\mathrm{v}}, A_i) = \begin{cases} A_{\mathrm{v}}(\log A_{\mathrm{v}} - \log A_i) + (A_i - A_{\mathrm{v}}) & (\beta = 1) \\ \dfrac{A_{\mathrm{v}}}{A_i} - \log\dfrac{A_{\mathrm{v}}}{A_i} - 1 & (\beta = 0) \\ \dfrac{A_{\mathrm{v}}^\beta}{\beta(\beta-1)} + \dfrac{A_i^\beta}{\beta} - \dfrac{A_{\mathrm{v}} A_i^{\beta-1}}{\beta-1} & (otherwise) \end{cases} \quad (7)$$

ここで、D_β は $\beta = 0$ および $\beta = 1$ で β について連続である。β ダイバージェンス基準による補間は、次式で定義するように、ヴァーチャルマイク信号と各実マイク信号との間の β ダイバージェンスの重み付け合計 σ_{D_β} を最小化することで行われる。

$$\sigma_{D_\beta} = (1-\alpha) D_\beta(A_{\mathrm{v}}, A_1) + \alpha D_\beta(A_{\mathrm{v}}, A_2) \quad (8)$$

$$A_{\mathrm{v}\beta} = \mathrm{argmin}_{A_{\mathrm{v}}} \sigma_{D_\beta} \quad (9)$$

よって、σ_{D_β} を A_{v} で微分し 0 と置くことで、次式のような一般化振幅補間が導出される。

$$A_{\mathrm{v}\beta} = \begin{cases} \exp((1-\alpha)\log A_1 + \alpha \log A_2) & (\beta = 1) \\ \left((1-\alpha) A_1^{\beta-1} + \alpha A_2^{\beta-1}\right)^{\frac{1}{\beta-1}} & (otherwise) \end{cases} \quad (10)$$

ここで次式より β ダイバージェンス D_β と同様に、$A_{\mathrm{v}\beta}$ は $\beta = 1$ で連続である。

$$\begin{aligned} A_{\mathrm{v}1} &= \lim_{\beta\to 1} \left((1-\alpha) A_1^{\beta-1} + \alpha A_2^{\beta-1}\right)^{\frac{1}{\beta-1}} \\ &= \exp((1-\alpha)\log A_1 + \alpha \log A_2) \end{aligned} \quad (11)$$

ここで、$\beta = 1$ において補間は式 (2) に示す従来の複素対数補間と等価になる。また、この振幅補間は、α で重み付けられた振幅を要素とするベクトル $\left[(1-\alpha)x_1, \alpha x_2\right]^{\mathrm{T}}$ の $\beta - 1$ 乗ノルムに相当する。そのため、$\beta \to +\infty$、$\beta \to -\infty$ の極限を取ることで、それぞれ次式のような最大値選択、最小値選択を表すことになる。

$$A_{\mathrm{v}\beta} = \max(A_1, A_2) \quad (\beta \to +\infty) \quad (12)$$

$$A_{\mathrm{v}\beta} = \min(A_1, A_2) \quad (\beta \to -\infty) \quad (13)$$

なお、位相については従来と同様に線形補間を用い、最終的なヴァーチャルマイク信号は次のように表される。

$$v = A_{\mathrm{v}\beta} \exp(j\phi_{\mathrm{v}}) \quad (14)$$

4 SN 比最大化ビームフォーマによる音声強調

本稿では、SN 比最大化ビームフォーマ [8] にヴァーチャルマイクロホンによる多素子化を導入し、目的音強調性能を検証する。SN 比最大化ビームフォーマは、録音信号中の目的音声区間と非目的音声区間それぞれの空間相関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。

4.1 SN 比最大化ビームフォーマの構成

SN 比最大化ビームフォーマによる音声強調では、次のようなマルチチャネルフィルタを構成する。

$$\mathbf{w}(\omega) = \left[w_1(\omega), \cdots, w_M(\omega)\right]^{\mathrm{T}} \quad (15)$$

なお、$\{*\}^{\mathrm{T}}$ は行列、ベクトルの転置を表す。また、フィルタ $\mathbf{w}(\omega)$ を次のようにして時間周波数領域の録音信号 $\mathbf{x}(\omega,t) = \left[x_1(\omega,t), \cdots, x_M(\omega,t)\right]^{\mathrm{T}}$ に掛けることで、強調音声 $y(\omega,t)$ が実現される。

$$y(\omega,t) = \mathbf{w}^{\mathrm{H}}(\omega)\, \mathbf{x}(\omega,t) \quad (16)$$

ただし、$\{*\}^{\mathrm{H}}$ は複素共役転置を表す。SN 比最大化ビームフォーマでは、出力信号中の目的信号と非目

Table 1: Experimental conditions

実マイク数	2
ヴァーチャルマイク数 N	0–9
実マイク間隔	4 cm
残響時間	640 ms
サンプリング周波数	8 kHz
FFT フレーム長	1024 samples
FFT フレームシフト幅	256 samples
テスト区間長	20 sec
目的音区間長 $\|\theta_{\mathrm{T}}\|$	10 sec
非目的音区間長 $\|\theta_{\mathrm{I}}\|$	10 sec

的信号のパワー比を最大化するようにフィルタ $\mathbf{w}(\omega)$ を構成する。ここでパワー比 $\lambda(\omega)$ は

$$\lambda(\omega) = \frac{\mathbf{w}^{\mathrm{H}}(\omega)\mathbf{R}_{\mathrm{T}}(\omega)\mathbf{w}(\omega)}{\mathbf{w}^{\mathrm{H}}(\omega)\mathbf{R}_{\mathrm{I}}(\omega)\mathbf{w}(\omega)} \tag{17}$$

のように表される。なお、$\mathbf{R}_{\mathrm{T}}$ は目的音区間、$\mathbf{R}_{\mathrm{I}}$ は非目的音区間それぞれの信号の空間相関行列であり、

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

$$\mathbf{R}_{\mathrm{I}}(\omega) = \frac{1}{|\Theta_{\mathrm{I}}|}\sum_{t\in\Theta_{\mathrm{I}}}\mathbf{x}_{\mathrm{I}}(\omega,t)\mathbf{x}_{\mathrm{I}}^{\mathrm{H}}(\omega,t) \tag{19}$$

と表される。ここで、$|\Theta_{\mathrm{T}}|$ は目的音区間、$|\Theta_{\mathrm{I}}|$ は非目的音区間それぞれ時の間フレーム数である。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。

$$\mathbf{R}_{\mathrm{T}}(\omega)\mathbf{w}(\omega) = \lambda(\omega)\mathbf{R}_{\mathrm{I}}(\omega)\mathbf{w}(\omega) \tag{20}$$

4.2 ビームフォーマのスケーリング補正

一般化固有値問題にはノルムに関する不定性があり、音声信号に適用した際のひずみの原因となる。そのため、任意の k チャネル目の録音信号を基準とし、次のようにビームフォーマのスケーリングを補正する [9]。

$$\mathbf{w}(\omega) \leftarrow b_k(\omega)\mathbf{w}(\omega) \tag{21}$$

ただし、$b_k(\omega)$ は録音信号全区間の空間相関行列

$$\mathbf{R}(\omega) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}(\omega,t)\mathbf{x}^{\mathrm{H}}(\omega,t) \tag{22}$$

を用いて以下のように定義される $\mathbf{b}(\omega)$ 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。

$$\mathbf{b}(\omega) = \frac{\mathbf{R}(\omega)\mathbf{w}(\omega)}{\mathbf{w}^{\mathrm{H}}(\omega)\mathbf{R}(\omega)\mathbf{w}(\omega)} \tag{23}$$

5 目的音源強調性能の検証実験

ヴァーチャルマイクロホンアレーによる多素子化の性能を検証するため、SN 比最大化ビームフォーマによる目的音強調の実験を行った。

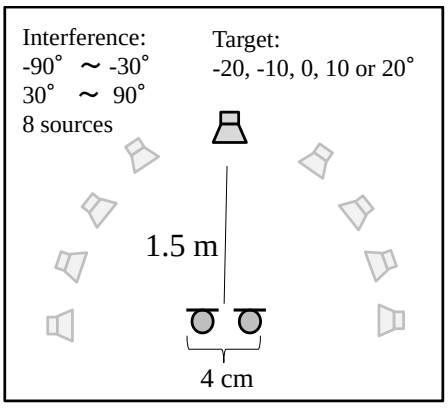


Fig. 3: Source and microphone layout in experiment

5.1 実験条件

音源と実マイクロホンの配置を図 3 に示す。また、その他の実験条件は表 1 に示す。目的音には、計 3 種類の日本語、英語の発話音声サンプルを使用した。また、目的音であるサンプルごとに図 3 中に記した 5 種類の到来方向を用意し、音声サンプル・到来方向の組み合わせ合計 15 通りについて実験を行った。また、妨害音としては、8 方向から 1 音声ずつが到来する、合計 8 音声からなる混合信号 1 種類を用いた。目的音、妨害音は音声サンプルと別途測定したインパルス応答の畳み込み混合により生成した。ヴァーチャルマイクロホンは 2 本の実マイクの間に等間隔に配置され、ヴァーチャルマイクロホン数を N とした時の i 番目のヴァーチャルマイクロホンに対する位置パラメータ α は次式で表される。

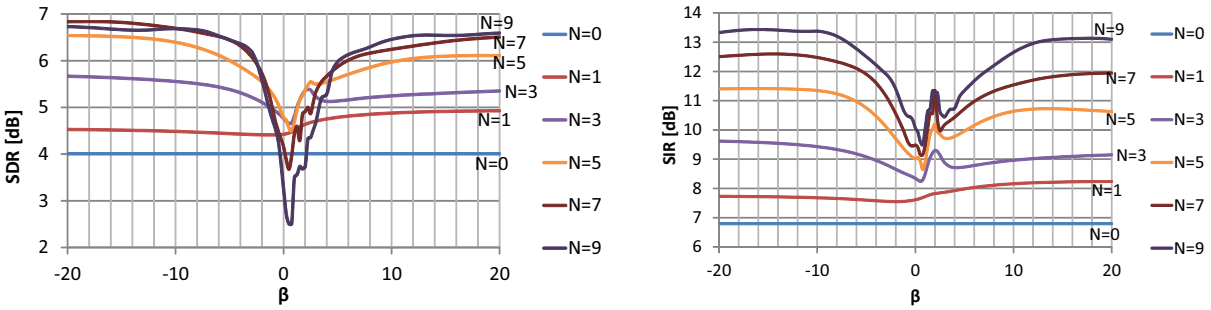
$$\alpha = \frac{i}{N+1} \tag{24}$$

ここで、音声強調処理は 2 個の実マイクと N 個のヴァーチャルマイクロホン、合計 $(N+2)$ チャネルからなるマイクロホンアレーで行われる。また、本実験では 1 番目のマイクロホンを 4.2 節で述べたスケーリング補正基準に用いる ($k=1$)。また、共分散行列の正則化処理は本稿では行っていない。

ビームフォーマによる目的音強調性能を評価するため、客観評価値として信号対ひずみ比 (SDR: Signal-to-distortion ratio)、および信号対妨害音成分比 (SIR: Signal-to-interference ratio) [10] を用いた。実験結果として、先述の 15 通りのサンプルについて平均した SDR, SIR を示す。

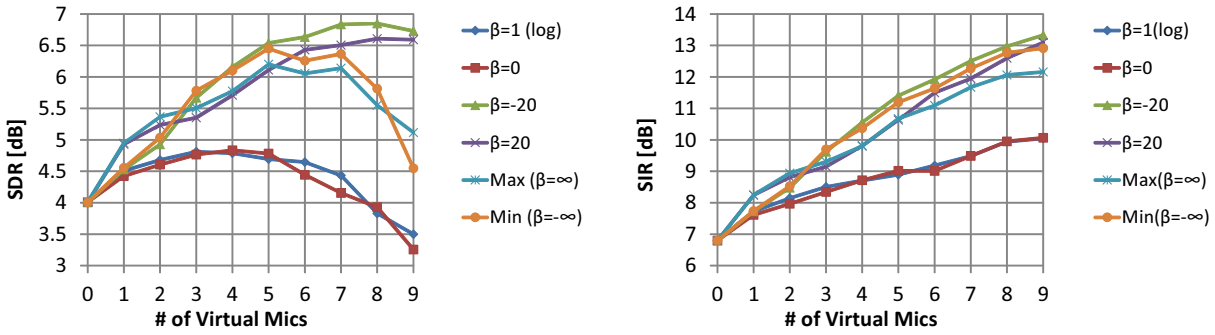
5.2 結果と考察

図 4 に、パラメータ β と音声強調性能の関係をヴァーチャルマイクロホン数ごとに示す。また、図 5 に音声強調性能とヴァーチャルマイクロホン数の関係を β の値ごとに示す。なお、ヴァーチャルマイク数 0 の条件は、実マイクのみからなる従来の SN 比最大化ビームフォーマを適用した条件に相当する。β の値によらず、ヴァーチャルマイクロホンを少数導入することにより SDR は向上する。また、SIR は多数のヴァーチャルマイクロホンを導入した際にも継続的な向上が見られることもわかる。対して、SDR については、β を 0 付近の値に設定し、多数のヴァーチャルマイクロホンを導入した場合、性能の大きな低下が見られる。これまでの複素対数補間 ($\beta=1$) でも同様に、ヴァーチャルマイクロホン数を増加させた際に



(a) SDR (b) SIR

Fig. 4: The relationship between β and separation performance



(a) SDR (b) SIR

Fig. 5: Performance shift by number of virtual microphones

大きく性能が低下する。しかし、β の値を 0 から離れた値に設定することにより、ヴァーチャルマイクロホンを多数導入した際にも SDR の向上が見られる。β を 0 周辺の値に設定した際の性能低下の原因として、空間相関行列のランク落ちが考えられる。また、β の値を変化させることで、補間の非線形性が調整され、ランク落ちが低減されたことが評価値の向上につながっていると考えられ、空間相関行列のランクと β の値の関係性について更に調査を進める予定である。

6 まとめ

本稿では、振幅補間への β ダイバージェンスの導入によるヴァーチャルマイクロホンアレーの一般化を提案した。また、新たに導入されたパラメータである β の値を変化させ、ヴァーチャル多素子化による目的音強調性能の向上について検証した。これまでに提案してきた複素対数補間では、ヴァーチャルマイクロホン数を増加させた際に大きな性能の低下が見られる。しかし、β を調整することにより、ヴァーチャルマイクロホン数を増加させても性能の向上が見られた。このことから、β ダイバージェンスによる振幅補間の拡張が、ヴァーチャルマイクロホンアレーによる目的音強調に有用であることが確認された。

参考文献

[1] 片平 他., “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,” 音講論 (春), pp. 741–744, 2013.

[2] Katahira *et al.*, “Virtually increasing microphone array elements by interpolation in complex-logarithmic domain,” Proc. EUSIPCO, TH-L 5.3, 5 pages, 2013.

[3] Chevalier *et al.*, “On the virtual array concept for higher order array processing,” IEEE Trans. Signal Processing, vol. 53, no. 4, pp. 1254–1271, 2005.

[4] Del Galdo *et al.*, “Generating virtual microphone signals using geometrical information gathered by distributed arrays,” Proc. HSCMA, pp. 185–190, 2011.

[5] Kowalczyk *et al.*, “Generating virtual microphone signals in noisy environments,” Proc. EUSIPCO, FR-L 3.6, 5 pages, 2013.

[6] Nakano *et al.*, “Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence,” Proc. IEEE MLSP, pp. 283–288, 2010.

[7] Févotte and Idier, “Algorithms for nonnegative matrix factorization with the β-divergence,” Neural Computation, 2011.

[8] Van Trees, “Optinum array processing,” John Wiley & Sons, 2002.

[9] 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), pp. 571–572, 2007.

[10] Vincent *et al.*, “Performance measurement in blind audio source separation,” IEEE Trans. on Audio, Speech & Language Processing, vol. 14, no. 4, pp. 1462–1469, 2006.