ヴァーチャル多素子化に基づくSN比最大化ビームフォーマの残響に対する性能変化*

☆山岡洸瑛（筑波大），小野順貴（NII/総研大），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，ビームフォーマを始めとした目的音源強調や，ブラインド音源分離などの様々な手法で利用されている．このようなマイクロホンアレーを用いた手法は，多チャネル録音を対象としており，その性能は録音チャネル数，すなわちマイクロホン数に依存する．実際には音源数と同数以上のチャネル数を必要とする手法が多く，音源数よりも少ない録音チャネル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．しかしながら，多くのスマートフォンやICレコーダーなどでマイク数は高々2個であり，こうした限られたチャネル数の小型機器でも利用できる手法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源強調を高性能化する手法として，我々はこれまでに「ヴァーチャル多素子化」を提案した[1, 2, 3]．本研究のヴァーチャルマイクロホンによる多素子化は，実マイクロホンによる観測信号から実際にはマイクロホンを置いていない位置の観測信号を推定するという形で，マイクロホンアレーを擬似的に多素子化する手法である．本手法では2本の実マイクロホンによる観測信号から，任意のチャネル数のヴァーチャルマイク信号を合成することが可能である．

ヴァーチャルマイク信号の合成方法としてこれまでに，実マイクによる観測信号の複素スペクトルの対数を用いてヴァーチャルマイク信号を補間する手法[1]と，それを拡張したβダイバージェンスに基づく補間[2, 3]を提案した．これらは時間周波数領域における音声のスパース性を仮定している．これにより，複数の音波が到来している場合も単一音波の補間とみなすことができる．しかしながら，残響時間が長い場合や，空間的エイリアシングが生じてしまう場合には，この仮定が成り立たなくなってしまい，そのときの性能は十分に調査されていない．マイク間距離を小さくすることで空間的エイリアシングは防止できるため，仮定が成り立たなくなる原因は主に残響であるといえる．そこで本稿では，残響がヴァーチャルマイクロホンによる多素子化に及ぼす影響を検証する．そのために，SN比最大化ビームフォーマ[4]にヴァーチャルマイクロホンによる多素子化を導入し，様々な残響時間の観測音に対して目的音源強調性能を比較する．

2 ヴァーチャルマイク信号の合成とSN比最大化ビームフォーマへの導入

2.1 βダイバージェンスに基づくヴァーチャルマイク信号の合成

我々の提案するヴァーチャル多素子化は，2チャネルの実マイクロホンから，任意のチャネルのヴァーチャルマイク信号を合成する手法である．実信号と合成したヴァーチャル信号の両方を用いることで，擬似的に多素子化されたアレー信号処理を行うことができる(Fig. 1)．ヴァーチャルマイク信号$v(\omega, t, \alpha)$は実マイクロホンの位置を$\alpha : (1-\alpha)$に内分する点での観測信号として定義される．信号は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．以降，特にω, t, αを区別する必要がないときは，単にvとする．また，i番目の実マイクロホン$(i = 1, 2)$による観測信号は$x_i(\omega, t)$とし，その振幅は$A_i = |x_i(\omega, t)|$, 位相は$\phi_i = \angle x_i(\omega, t)$と表す．

βダイバージェンスに基づくヴァーチャルマイク信号の合成において，ヴァーチャルマイク信号vの振幅$A_{\mathrm{v}\beta}$は次のように表される[2, 3]．

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

この振幅補間は、αで重み付けられた多チャネ

*Performance of maximum SNR beamformer based on virtual increase of channels in reverberant environments. by Kouei YAMAOKA (University of Tsukuba), Nobutaka ONO (National Institute of Informatics / SOKENDAI), Takeshi YAMADA, Shoji MAKINO (University of Tsukuba)

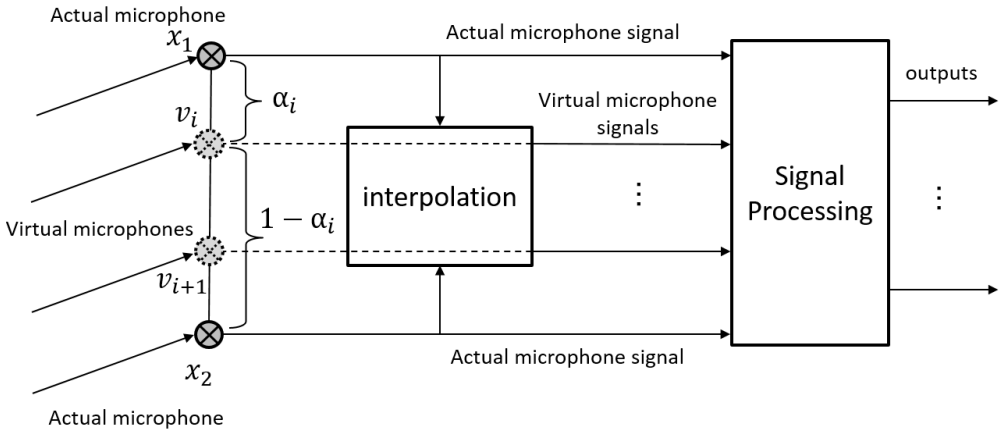


Fig. 1: Signal processing with virtual microphone array

ル観測ベクトル $\left[(1-\alpha)\,x_1, \alpha x_2\right]^{\mathrm{T}}$ の $\beta-1$ 乗ノルムに相当する。ここで

$$\begin{aligned} A_{\mathrm{v}1} &= \lim_{\beta\to 1}\left((1-\alpha)\,A_1^{\beta-1}+\alpha A_2^{\beta-1}\right)^{\frac{1}{\beta-1}} \\ &= \exp\left((1-\alpha)\log A_1+\alpha\log A_2\right) \quad (2) \end{aligned}$$

より $A_{\mathrm{v}\beta}$ は $\beta=1$ において連続であり，補間は複素対数補間 [1] と等価になる．また，$\beta\to+\infty$、$\beta\to-\infty$ のとき、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。

$$A_{\mathrm{v}\beta} = \max\left(A_1, A_2\right), \quad (\beta\to+\infty) \quad (3)$$

$$A_{\mathrm{v}\beta} = \min\left(A_1, A_2\right), \quad (\beta\to-\infty) \quad (4)$$

位相に関しては，次式のように線形補間される．

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

以上より，ヴァーチャルマイク信号は次のように表される．

$$v = A_{\mathrm{v}\beta}\exp\left(j\phi_{\mathrm{v}}\right) \quad (6)$$

2.2 SN 比最大化ビームフォーマ

SN 比最大化ビームフォーマ [4] 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．音源位置が未知の場合においても適用できるといった利点がある．本手法の先行研究 [1, 2, 3] においても，SN 比最大化ビームフォーマにヴァーチャルマイクロホンによる擬似的な多素子化を導入し，性能調査をしているため，本稿でもそれに従う．

3 実験方法

3.1 実験の概要

本研究では，残響下における本手法の性能評価のため，様々な残響時間のインパルス応答とドライソースの音声の畳み込みにより作成した観測信号を用いて，シミュレーション実験を行った．実マイクによる観測信号のみを用いて SN 比最大化ビームフォーマを適用した場合と，実マイクによる観測から合成したヴァーチャルマイク信号と実マイク信号の両方を用いて SN 比最大化ビームフォーマを適用した場合の結果を比較することで性能を評価する．評価には，信号対歪比 (SDR: Signal-to-Distortion Ratio) 及び信号対干渉比 (SIR: Signal-to-Interference Ratio) [5] を用いる．これらの評価値は数値が高い方が高い性能を表す．

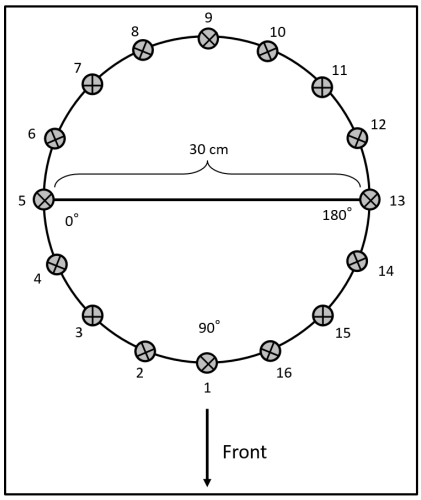


Fig. 2: Microphone array of RWCP-SSD

3.2 RWCP 実環境音声・音響データベース

RWCP 実環境音声・音響データベース [6] (RWCP Sound Scene Database in Real Acoustic Environment. 以下「RWCP-SSD」という.) とは，実環境における音声・音響信号処理の研究を対象とした共通の評価用データベースである．

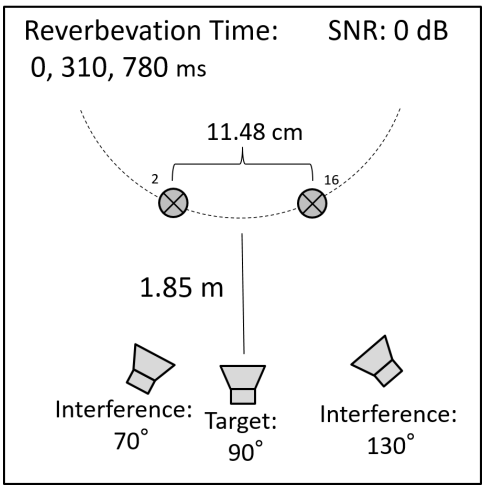


Fig. 3: Sound source and microphone layout in experiment

データベースには，実際に測定した様々な残響時間のインパルス応答が含まれており，その中から選択し使用する．Fig. 2 は測定時に使用された 16 ch 円形マイクロホンアレーである．音源は円形アレーの中心から 2 m の位置にある．また，Fig. 2 の各数字は素子番号で，角度は円形アレーの中心から 5 番目の素子の方向を 0° として半時計回りになっている．

本実験では残響時間が 0, 310, 780 ms のインパルス応答を使用する．これらは無響室と残響可変室で測定されたインパルス応答である．使用する実マイクロホンは Fig. 2 の 2, 16 番目の素子とした．この時，マイク間距離は約 11.48 cm と大きくなっているため，空間的エイリアシングが生じてしまう．本研究では，残響の影響を検証するため，空間的エイリアシングはない方が望ましい．従って，空間的エイリアシングの影響を抑えることができる方向に音源を配置した．

3.3 実験条件

音源と実マイクロホンの配置を Fig. 3 に示す．また，その他の実験条件を Table. 1 に示す．目的音には計 8 種類の日本語及び英語の発話音声サンプルを使用し，到来方向は 90° とした．これは Fig. 2 の Front 方向で，使用する 2, 16 番目の素子に対して垂直に到来する．妨害音には，70°, 130° 方向から 1 音声ずつ到来する，計 2 音声からなる混合信号を 2 種類使用した．目的音，妨害音は RWCP-SSD[6] のインパルス応答とドライソースの音声との畳込みにより生成し，入力 SN 比を 0 dB として混合した．インパルス応答は残響時間が 0, 310, 780 ms の 3 種類を使用した．ヴァーチャルマイクロホンは使用する実マイクロホンの位置の中点 ($\alpha = 0.5$) に 1 つ生成した．この時，目的音源強調は 2 つの実マイクロホンと 1 つのヴァーチャルマイクロホンの計 3 チャネルからなるマイクロホンアレーで行われる．実験結果として，先述の 8 種類の目的音と 2 種類の妨害音の組み合わせ，計 16 組に対して目的音源強調を行った結果の SDR, SIR を平均して示す．

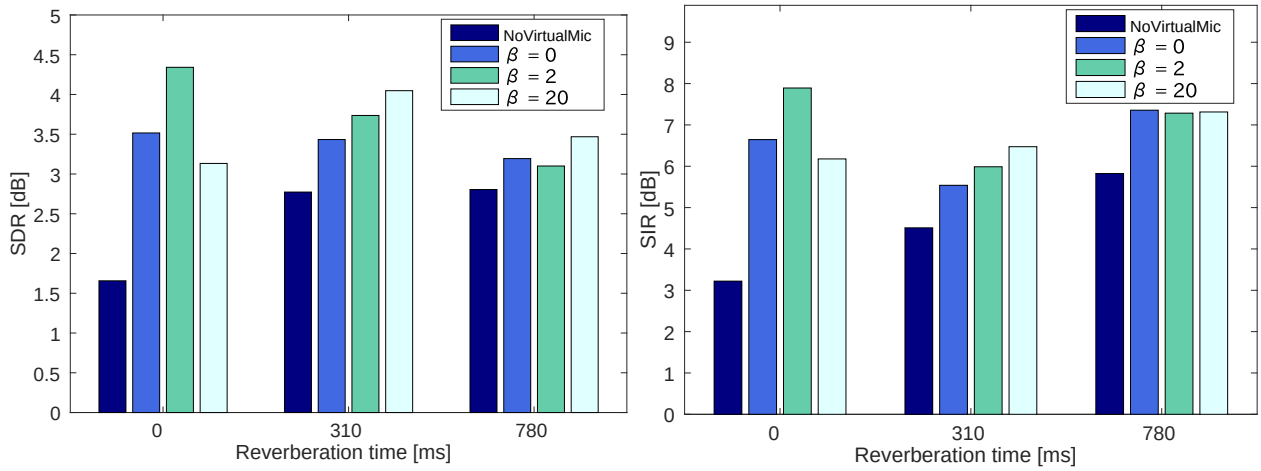
Table 1: Experimental conditions

実マイク数	2 (素子 2, 16)
ヴァーチャルマイク数	1 ($\alpha = 0.5$)
実マイク間隔	約 11.48 cm
残響時間	0, 310, 780 ms
入力 SN 比	0 dB
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 s
目的音区間長	10 s
非目的音区間長	10 s

4 結果と考察

Fig. 4 に，残響時間と目的音源強調性能の関係を，β の値毎に示す．なお，図中の NoVirtualMic はヴァーチャルマイクロホンを使用せず，実マイクのみからなる SN 比最大化ビームフォーマである．

無残響下においては，ヴァーチャル多素子化を導入しない場合に比べて，$\beta = 2$ のときに SDR が 2.7 dB, SIR が 4.7 dB 程度向上しており，本手法の有効性が確認できる．残響時間 310 ms における SDR, SIR は，ヴァーチャル多素子化を導入しない場合に比べて，$\beta = 20$ のときにそれぞれ 1.3 dB, 2.0 dB 程度の向上，780 ms においては $\beta = 20$ のときにそれぞれ 0.7 dB, 1.5 dB 程度の向上となっている．このように，いずれの場合も SDR, SIR が向上しているため，残響下においてもヴァーチャルマイクロホンによる擬似的な多素子化は性能の向上に寄与するといえる．一方で，残響が長くなるにつれて，ヴァーチャルマイクロホンを導入することによる改善量は少なくなっている．



(a) SDR (b) SIR

Fig. 4: The relationship between reverberation time and separation performance

補間の非線形性を調整するパラメータ β は，残響がある場合には正負どちらかの方向に大きくするほど性能は良くなり，$\beta = \pm 20$ 付近で，SDR の値が収束する傾向にあった．しかし残響時間 0 ms の場合は，$\beta = 2$ が最適値となっており，残響時間と β の値との関係性について更に調査を進める必要がある．

また，残響時間が長いほど，SN 比最大化ビームフォーマの性能が良くなっている理由として以下が考えられる．SN 比最大化ビームフォーマは指向性を制御している．しかし，2 個のマイクロホンのみで 2 つの妨害音の方向それぞれに指向性の零点を向けることはできない．従って，妨害音の方向へ指向性の急峻な零点を作ることができないため，妨害音の方向だけでなく，その周囲の方向も利得が小さいような指向性が作られる．その場合，残響時間が長いほど，妨害音除去能力が上がる可能性がある．残響時間と SN 比最大化ビームフォーマの性能との関係性についても，更に調査を進める必要がある．

5 まとめ

本稿では，ヴァーチャル多素子化で仮定している音のスパース性が成り立たない場合の性能を評価するため，残響下におけるヴァーチャル多素子化の性能を検証した．これまでは残響時間の影響が未知であったが，長い残響時間においても一定の性能が得られることが確認された．また，残響時間が短い場合には性能の大きな向上を得ることができる．このことから，ヴァーチャルマイクロホンによる擬似的な多素子化が，目的音源強調に有効であることが確認された．

参考文献

[1] 片平拓希, 小野順貴, 宮部滋樹, 山田武志, 牧野昭二, “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] 片平拓希, 小野順貴, 宮部滋樹, 山田武志, 牧野昭二, “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,” 音講論 (秋), pp. 633–636, 2014.

[3] H. Katahira, N. Ono, S. Miyabe, T. Yamada and S. Makino, “Nonlinear speech enhancement by virtual increase of channels and maximum SNR beamformer,” EURASIP Journal on Advances in Signal Processing, vol. 2016, no. 1, pp. 1–8, Jan. 2016.

[4] H. L. Van Trees, *Optinum array processing*, John Wiley & Sons, 2002.

[5] E. Vincent, R. Gribonval and C. Févotte, “Performance measurement in blind audio source separation,” IEEE Trans. on Audio, Speech & Language Processing, vol. 14, no. 4, pp. 1462–1469, 2006.

[6] S. Nakamura, K. Hiyane, F. Asano, Y. Kaneda, T. Yamada, T. Nishiura, T. Kobayashi, S. Ise and H. Saruwatari, “Design and collection of acoustic sound data for hands-free speech recognition and sound scene understanding,” ICME 2002, vol. 2, pp. 161–164, Aug. 2002.