ヴァーチャルマイクロフォンの内挿における位相及び振幅補間の音声強調性能への影響の評価*

☆瀬川華子[1]，李莉[2]，牧野昭二[1,3]，山田武志[1]
[1] 筑波大学，[2] NTT コミュニケーション科学基礎研究所，[3] 早稲田大学

1 はじめに

近年，音声を用いたシステムの開発が多方面で進められ，それに伴って音声強調技術の研究が進められている。その一方で，マイクロフォンアレイを用いた音声強調技術の性能は，マイク数に大きく依存するため，利用できる場面が限られている。例えば，ビームフォーミングによる音声強調では，マイクの数より多い音源が存在する劣決定条件の場合には，強調性能が劣化することが知られている。故に，少数のマイクを用いて行われる音声強調は，マイク数の制約がある環境下でも用いることができるため，需要が高まっている。

このようなマイク数が制限された劣決定条件下において，音声強調性能を改善するための手法としてヴァーチャルマイクロフォン技術 [1][2] が提案されてきた。ヴァーチャルマイクロフォン技術は，2 つの実マイクロフォンの直線上に仮想的な信号を合成する技術であり，合成されたヴァーチャルマイクロフォン信号を用いることにより観測信号のチャネル数を人為的に増やすことで音声強調性能を改善する。ヴァーチャルマイクロフォン信号の合成は複素スペクトログラム領域において行われ，平面波の仮定に基づいた線形補完及び β ダイバージェンスを用いたルールベースの非線形補完により位相と振幅をそれぞれ求める。近年，深層学習を用いた振幅の内挿手法も提案されている [3][4]。このように位相と振幅をそれぞれ求めた合成信号を用いることで，ビームフォーミングの音声強調性能が改善できることが確認された一方で，それらが音声強調性能にどれほどの影響を及ぼすのかに関する調査は十分に行われていない。また，従来のヴァーチャルマイクロフォン技術は仮想信号が 2 つの実マイクの直線上にある場合しか適用できないため，利用できる場面が限られている。

そこで，本稿では，任意数のマイクから仮想信号を合成できるように従来法を拡張し，同一平面上にある 3 つの実マイクを用いて提案法の有効性を示す。また，2 つと 3 つの実マイクを用いた場合に，位相と振幅それぞれの補間が音声強調性能に与える影響をシミュレーションと車室内で測定したインパルス応答を用いて検討する。

2 2 マイクにおけるヴァーチャルマイクロフォンの内挿

本節では，2 マイクにおけるヴァーチャルマイクロフォンの内挿 [1][2] について紹介する。ヴァーチャルマイクロフォン技術では 2 つの実マイクロフォンの観測信号 $X_i(\omega,t)$ から，ヴァーチャルマイクロフォン信号 $V(\omega,t,\alpha)$ を生成する。i は生成に用いる実マイクロフォンの識別子 ($i=1,2$) であり，ω はそのマイクロフォンにおける周波数ビン，t は時間フレームを表す。α $(0<\alpha<1)$ はヴァーチャルマイクロフォンの補間係数であり，実マイクロフォン間の距離を 1 とした時の，1 番目の実マイクとヴァーチャルマイクロフォン間の距離の値である。Fig. 1 に実マイクロフォンとヴァーチャルマイクロフォンの関係を表す。複数の音が異なる方向から到来する場合，マイクロフォンの位置と波形の関係は複雑になり，補間が困難であるため，観測信号のスパース性を仮定する。すなわち，1 つの時間周波数ビンでは 1 つの音源のみが支配的であることを仮定する。これにより，複数の音が到来する時であっても各時間周波数ビン内では単一の音とみなすことができ，ヴァーチャルマイクロフォン信号の補間を行うことができる。

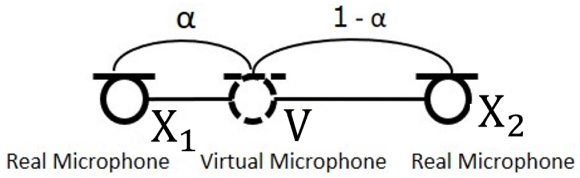

Fig. 1: Arrangement of 2 real and virtual microphones in interpolation technique

ヴァーチャルマイクロフォン技術においては位相と振幅は個別に補間される。実マイクロフォンでの観測信号の位相と振幅は以下のように表される。

$$\phi_i = \angle X_i(\omega,t) = \tan^{-1}\frac{\mathrm{Im}(X_i(\omega,t))}{\mathrm{Re}(X_i(\omega,t))}, \quad (1)$$

$$A_i = |X_i(\omega,t)|. \quad (2)$$

平面波の到来を仮定した時，ヴァーチャルマイクロフォンを内挿した時の位相は線形補間によって次のように表される。

$$\begin{aligned}\phi_v &= \phi_1+\alpha(\phi_2-\phi_1)\\ &= (1-\alpha)\phi_1+\alpha\phi_2. \end{aligned} \quad (3)$$

ここでは，信号の位相差が π を超えていないことを仮定して補間を行う。

$$|\phi_1-\phi_2|\le\pi. \quad (4)$$

振幅は多くの条件に依存しており，実際の振幅減衰を忠実にモデル化するのは困難である。そこで，従来法では，振幅補完を 2 つの実マイクの振幅との重みつ

*Evaluation of the effect of phase and amplitude interpolation on speech enhancement performance in virtual microphone interpolation. by Hanako SEGAWA (University of Tsukuba), Li LI (NTT), Shoji MAKINO (University of Tsukuba, Waseda University),Takeshi YAMADA(University of Tsukuba).

き β ダイバージェンスが最小となるような最適化問題として定式化し，その閉形式の解

$$A_v = \begin{cases} \exp\left((1-\alpha)\log A_1 + \alpha \log A_2\right) & (\beta = 1) \\ \left((1-\alpha)A_1^{\beta-1} + \alpha A_2^{\beta-1}\right)^{\frac{1}{\beta-1}} & (\text{otherwise}). \end{cases} \tag{5}$$

を補間ルールとして用いられた。β は振幅補完の非線形性を調整する役割を担った変数となる。式 (3) と式 (5) により求められた ϕ_v と A_v を用いて，合成されたヴァーチャルマイクロフォン信号は以下のように表すことができる。

$$V(\omega, t, \alpha) = A_v \exp\left(j\phi_v\right). \tag{6}$$

3 提案法：任意数のマイクを用いたヴァーチャルマイクロフォンの内挿

本節では，前節で述べた位相の線形補間と振幅の β ダイバージェンスによる補間からなるヴァーチャルマイクロフォンの内挿を同一平面上にある 3 つ以上のマイクに拡張する。

I 個の実マイクが存在し，i 番目のマイクの位置を座標 $p_i = [x_i, y_i, z_i]^\mathsf{T}$ で表し，ヴァーチャルマイクロフォン信号の座標を $p_v = [x_v, y_v, z_v]^\mathsf{T}$ で表すとする。まず，平面波の仮定に基づいた位相の線形補完を考える。マイクと到来する平面波が同一平面上にある場合には，非同一直線上にある 3 つの実マイクの座標のアフィン結合で任意位置のヴァーチャルマイクロフォン信号の座標を表すことができる。

$$p_v = \lambda_1 p_1 + \lambda_2 p_2 + \lambda_3 p_3, \tag{7}$$

$$\text{s.t.} \sum_i^3 \lambda_i = 1 \tag{8}$$

そこで，平面波の仮定に基づいて，位相の補完は

$$\phi_v = \sum_{i=1}^3 \lambda_i \phi_i. \tag{9}$$

のように表せる。ただし，λ_i は線形結合係数であり，以下の連立一次方程式

$$\mathbf{A}\boldsymbol{\lambda} = \mathbf{c} \tag{10}$$

の解である。ここで，$\mathbf{A}$, $\mathbf{c}$ と $\boldsymbol{\lambda}$ はそれぞれ以下のように定義する。

$$\mathbf{A} = \begin{bmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{bmatrix}, \mathbf{c} = \begin{bmatrix} x_v \\ y_v \\ 1 \end{bmatrix}, \boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} \tag{11}$$

音源とマイクが同一平面上に存在するため，$z_i = 0$ 及び $z_v = 0$ であり，任意の解を $\sum_i \lambda_i z_i = z_v$ に代入しても成立することが自明である。

振幅は，2 マイクにおける β ダイバージェンスを用いた補間を任意数のマイクに拡張する。各実マイクロフォンからヴァーチャルマイクロフォンの座標までの距離を $\hat{\alpha}$ とし，以下のように定義する。

$$\hat{\alpha}_i = \sqrt{(x_i - x_v)^2 + (y_i - y_v)^2 + (z_i - z_v)^2}, \tag{12}$$

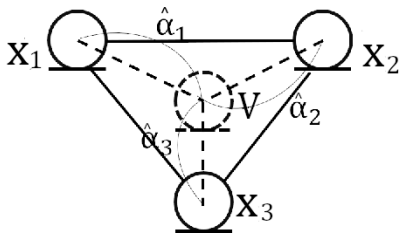

Fig. 2: Arrangement of 3 real and virtual microphones in interpolation technique

この時，各実マイクの振幅との β ダイバージェンスの重みを表す係数である α_i は，近い実マイクロフォンほど重みが大きくなように距離 $\hat{\alpha}$ に反比例することを仮定する。また，$\sum_i \alpha_i = 1$ となるように正規化を行うと重み係数 α_i は次式で表される。

$$\alpha_i = \frac{\hat{\alpha}_1 \hat{\alpha}_2 \hat{\alpha}_3}{\hat{\alpha}_i(\hat{\alpha}_1\hat{\alpha}_2 + \hat{\alpha}_2\hat{\alpha}_3 + \hat{\alpha}_1\hat{\alpha}_3)} \tag{13}$$

Fig. 2 に α_i とマイクロフォンの位置の関係を表す。ヴァーチャルマイクロフォン信号の振幅 A_v と，i チャネル目の実マイクロフォン信号の振幅 A_i の間の β ダイバージェンス $D_\beta(A_v, A_i)$ は次式のように定義される。

$$D_\beta(A_v, A_i) = \tag{14}$$
$$\begin{cases} A_v(\log A_v - \log A_i) + (A_i + A_v) & (\beta = 1) \\ \frac{A_v}{A_i} - \log \frac{A_v}{A_i} - 1 & (\beta = 0) \\ \frac{A_v^\beta}{\beta(\beta-1)} + \frac{A_i^\beta}{\beta} - \frac{A_v A_i^{\beta-1}}{\beta-1} & (\text{otherwise}) \end{cases}$$

β ダイバージェンスを用いた振幅補間は，次式のような最適化を行う。

$$A_v = \text{argmin}_{A_v} \sum_{i=1}^I \alpha_i D_\beta(A_v, A_i) \tag{15}$$

ただし，位相の補完と異なり，振幅の補完にはすべての I 個の実マイクを用いることができる。上式を最適化する解は

$$A_v = \begin{cases} \exp\{\sum_{i=1}^I \alpha_i \log A_i\} & (\beta = 1), \\ (\sum_{i=1}^I \alpha_i A_i^{\beta-1})^{\frac{1}{\beta-1}} & (\text{otherwise}). \end{cases} \tag{16}$$

で表せる。この補完ルールは式 (5) を一般化したものであることが分かる。これにより，ヴァーチャルマイクロフォン信号は ϕ_v と A_v を用いて以下のように表すことができる。

$$V(\omega, t, \Xi) = A_v \exp\left(j\phi_v\right). \tag{17}$$

ここで，$\Xi = \{\alpha_i, \lambda_i\}_i$ はすべての線形結合係数 λ_i と重み係数 α_i の集合である。

4 評価実験

4.1 実験概要

本実験では， 4 種類のシミュレーションインパルス応答と， 1 種類の車室内実測インパルス応答を用いて実験を行った。はじめに，三角マイクロフォンア

Table 1: sound source direction and reverberation time

RT	target	Interf 1	Interf 2	Interf 3
60	80°	160°	40°	130°
120	90°	145°	50°	10°
250	110°	150°	20°	70°
400	70°	100°	30°	150°

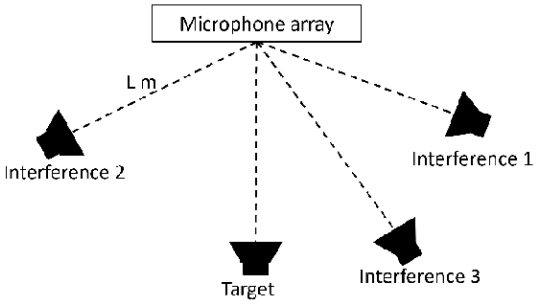

Fig. 3: Arrangement of sound sources in simulation

レイに対して従来法を 2 回用いて提案法と同座標にヴァーチャルマイクロフォン信号を合成する方法および提案法それぞれで内挿を行い，音声強調性能を比較した。次に，2 マイクロフォンと 3 マイクロフォンそれぞれのヴァーチャルマイクロフォン信号の補間について，位相・振幅を共に補間する場合，位相のみを補間する場合，振幅のみ補間する場合，補間を行わない場合（実マイクロフォン信号）を比較することで，振幅・位相それぞれの補間による音声強調性能への影響を調査した。

4.2 実験条件

本実験では, ATR デジタル音声データベースのセット B に収録されている, 全 503 文の音素バランス文の男性 6 話者, 女性 4 話者の計 10 話者分のデータを使用した [7]。このデータベースの中から，ランダムに 3 人または 4 人を 25 パターン選択し，4 種類のシミュレーションのインパルス応答を畳み込むことでシミュレーションインパルス応答を用いた観測信号をそれぞれ 100 パターン作成した。さらに，同様に 1 種類の実測インパルス応答を 1 種類畳み込むことで観測信号を 25 パターン作成した。シミュレーションの音源の配置を Fig. 3 と Table. 1 に示す。Table. 1 における RT は，残響時間 (Reverberation time) を示す。インパルス応答の音源の配置を Fig. 4 に示す。また，マイクロフォンアレイの配置を Fig. 5，Fig. 6 に示す。3 音源に対して 2 つの実マイクロフォン信号を用いて音声強調を行う実験では，Target, Interference 1, Interference 2 を実験に用い, 4 音源に対して 3 つの実マイクロフォン信号を用いて音声強調を行う実験では,Target, Interference 1, Interference 2, Interference 3 を実験に用いる。β ダイバージェンスにおける β は，2 つの実マイクロフォンを用いた実験と従来法においては 20 を用いた。3 つの実マイクロフォンを用いた提案法では，140 を用いた。評価指針として，signal-to-distortion ratio (SDR) Improvement ,source-to-interference ratio(SIR), sources-to-artifacts ratio (SAR) を用いる。音声強調には，maxSNR ビームフォーマを用いる。[5] [6] その他の実験条件を Table. 2 に示す。

Table 2: experimental conditions

Sampling rate	8 kHz
Signal-to-noise ratio(SNR)	0 dB
FFT frame length	1024 samples
FFT shift	256 samples
Beamformer	maxSNR

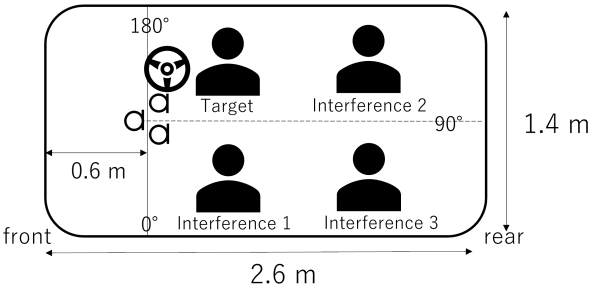

Fig. 4: Arrangement of sound sources in real car

4.3 実験結果と考察

はじめに，3 つの実マイクロフォンを用いたヴァーチャルマイクロフォンの補間について従来法と提案法の実験結果を Table. 3 に示す。prev((A,B),C) は，従来の 2 つの実マイクロフォンからヴァーチャルマイクロフォン信号を補間する手法を用いて，実マイクロフォン A，実マイクロフォン B からヴァーチャルマイクロフォンを合成し，合成したヴァーチャルマイクロフォンと実マイクロフォン C を用いて最終的に音声強調に使用するヴァーチャルマイクロフォン信号を合成することを示す。シミュレーションにおいては，従来法は実マイクロフォンの選択順にかかわらずほぼ SDR が変わらず，提案法においても従来法と同程度の SDR を示した。実測インパルス応答を用いた実験では，従来法において実マイクロフォンの選択順によって SDR にわずかにばらつきが見られ，提案法はその平均程度の SDR を示した。提案法においては，従来法におけるマイクロフォンの選択順による性能の変動を無くすことができ，かつ任意のマイク数への拡張が容易であるという利点がある。

次に，3 音源に対して 2 つの実マイクロフォン信号を用いて音声強調を行った実験結果を Table. 4 に示す。Phase, Amp における R, V は，ヴァーチャルマイクロフォン位置の信号における位相や振幅において，実位相，実振幅を与えている場合は R，ヴァーチャルマイクロフォン技術によって補間された位相，振幅を与えている場合には V で表す。位相，振幅ともにヴァーチャルマイクロフォン技術で補間した信号を 2 つの実マイクロフォンに加えて音声強調に用いた場合，2 つの実マイクロフォンの信号のみを用いて音声強調を行った場合と比較して一定の SDR の向上が確認された。また，実位相のみを与えた場合には，位相，振幅ともにヴァーチャルマイクロフォン技術で補間した場合と SDR においてほぼ差が見られなかった一方で，実振幅を与えた場合には SDR がシミュレーションでは約 1.9dB，実測インパルス応答では約 3.7dB 向上した。これは，2 つの実マイクロフォンを用いた場合の補間において，補間した振幅は実振幅と大きく異なっており，かつ振幅が音声強調性能に与える影響が大きいと考えられ，実振幅を与え

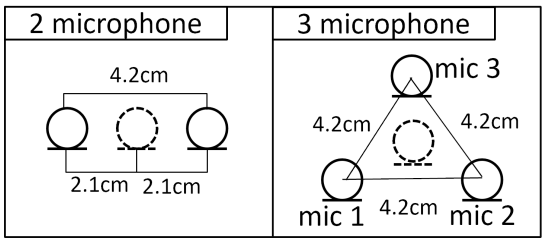

Fig. 5: Arrangement of microphones in simulation

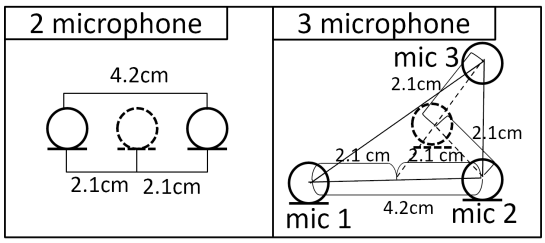

Fig. 6: Arrangement of microphones in real car

ることで大きく音声強調性能が改善したと考えられる。一方で，3 つの実マイクロフォン信号を用いた決定条件における音声強調と，2 つの実マイクロフォン信号と位相のみをヴァーチャルマイクロフォンで補間した信号で音声強調を行った場合で SDR にシミュレーションでは約 11dB，実測インパルス応答では約 6.7dB の差が見られたことから，位相の補間や位相と振幅の整合性など今まで検討されてこなかった部分についても検討が必要であると考えられる。

次に，4 音源に対して 3 つの実マイクロフォン信号を用いて音声強調を行った実験結果を Table. 5 に示す。4 音源に対する実験においても，位相，振幅ともにヴァーチャルマイクロフォン技術で補間した信号を実マイクロフォン信号に加えて用いた場合に一定の SDR の向上が確認された。一方で，実位相のみ，実振幅のみを与えた場合ともに，位相，振幅ともにヴァーチャルマイクロフォン技術で補間した場合と比較して SDR にほとんど差が見られなかった。このことから，3 マイクにおいても位相と振幅の整合性など振幅と位相の補間以外の観点からの検討を行う必要があると考えられる。

5 結論

本稿では，ヴァーチャルマイクロフォン技術を同一平面上にある 3 つ以上の実マイクロフォン信号に拡張することを提案した。また，劣決定条件下においてヴァーチャルマイクロフォン信号に実位相のみ・実振幅のみを与えた場合や実マイクロフォン信号を与えた場合と比較することで，ヴァーチャルマイクロフォン信号の補間による音声強調性能への影響について検討した。実験の結果，提案法でも劣決定条件下における音声強調性能の改善が確認された。提案法は，3 つ以上のマイク数への拡張を容易に行うことができる。また，ヴァーチャルマイクロフォン技術における位相と振幅の補間が音声強調性能に及ぼす影響の内容が確認された。

謝辞 本研究は科研費 19H04131 の助成を受けた。

Table 3: Proposed method and previous method in simulation

	method	SDRi	SIR	SAR
simu	proposed	6.74	6.46	5.00
	prev((1,2),3)	6.73	6.44	5.01
	prev((1,3),2)	6.73	6.43	5.01
	prev((2,3),1)	6.73	6.43	5.01
real	proposed	11.71	14.58	8.24
	prev((1,2),3)	11.86	14.96	8.30
	prev((1,3),2)	11.51	14.29	8.04
	prev((2,3),1)	11.86	14.98	8.30

Table 4: experiment of 2 microphone

	mic	Phase	Amp.	SDRi	SIR	SAR
simu	2mic	-	-	1.97	0.44	7.95
	2mic+VM	V	V	3.37	2.62	6.61
	2mic+VM	R	V	3.61	3.01	6.51
	2mic+VM	V	R	5.38	5.05	7.30
	3mic	R	R	17.11	21.74	15.85
real	2mic	-	-	4.88	6.07	6.37
	2mic+VM	V	V	7.06	10.45	6.86
	2mic+VM	R	V	7.77	11.51	7.34
	2mic+VM	V	R	10.80	16.07	9.97
	3mic	R	R	17.48	25.25	15.77

Table 5: experiment of 3 microphone

	mic	Phase	Amp.	SDRi	SIR	SAR
simu	3mic	-	-	5.78	5.68	4.87
	3mic+VM	V	V	6.25	6.46	5.00
	3mic+VM	R	V	6.38	6.71	5.02
	3mic+VM	V	R	6.30	6.41	5.09
	4mic	R	R	10.31	11.29	8.17
real	3mic	-	-	10.39	12.59	7.16
	3mic+VM	V	V	11.71	14.58	8.24
	3mic+VM	R	V	11.72	14.93	8.13
	3mic+VM	V	R	11.91	15.00	8.40
	4mic	R	R	18.56	23.41	14.55

参考文献

[1] H. Katahira et al., “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, 2016.

[2] K. Yamaoka et al., “Performance evaluation of nonlinear speech enhancement based on virtual increase of channels in reverberant environments,” In Proc. EUSIPCO, pp. 2324–2328, 2017.

[3] K. Yamaoka et al., “CNN-based virtual microphone signal estimation for MPDR beamforming in underdetermined situations,“ In Proc. EUSIPCO, pp. 1-5, 2019.

[4] R. Takahashi et al. “VMInNet: Interpolation of Virtual Microphones in Optimal Latent Space Explored by Autoencoder.“ In Proc, NCSP, pp. 93-96, 2021.

[5] H.L van Trees., “Optimum array processing,” 2002.

[6] S. Araki et al., “Blind speech separation in a meeting situation with maximum snr beamformers,” In Proc. ICASSP, Vol. 1, pp. 41-44, 2007.

[7] A. Kurematsu et al., “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech communication, vol. 9, no. 4, pp. 357–363, 1990.

[8] 山岡洸瑛他.,”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,” 音講論 (秋), pp. 379-382, 2016.