GENERALIZED AMPLITUDE INTERPOLATION BY β-DIVERGENCE FOR VIRTUAL MICROPHONE ARRAY

Hiroki Katahira[1], *Nobutaka Ono*[2,3], *Shigeki Miyabe*[1], *Takeshi Yamada*[1], *Shoji Makino*[1]

[1]University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8577 Japan
[2]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo, 101-8430 Japan
[3]The Graduate University for Advanced Studies (Sokendai)
katahira@mmlab.cs.tsukuba.ac.jp, onono@nii.ac.jp,
{miyabe, maki}@tara.tsukuba.ac.jp, takeshi@cs.tsukuba.ac.jp

ABSTRACT

In this paper, we present a generalization of the virtual microphone array we previously proposed to increase the microphone elements by nonlinear interpolation. In the previous work, we generated a virtual observation from two actual microphones by an interpolation in the logarithmic domain. This corresponds to a linear interpolation of the phase and the geometric mean of the amplitude. In this paper, we generalize this interpolation using a linear interpolation of the phase and a nonlinear interpolation of the amplitude with adjustable nonlinearity based on β-divergence. Improvement of the array signal processing performance is obtained by appropriate tuning of the parameter β. We evaluate the improvement in speech enhancement using a maximum SNR beamformer.

Index Terms— virtual microphone, array signal processing, speech enhancement, maximum SNR beamformer

1. INTRODUCTION

A microphone array is a signal processing framework based on multichannel observation and it is important for blind source separation (BSS) [1], direction of arrival (DOA) estimation [2] and speech enhancement. The array signal processing performance depends on the number of microphones. Although several methods such as time-frequency (T-F) masking [3] and multichannel Wiener filter [4] can work well with a small number of microphones, better performance can be expected if more microphones are available.

Therefore, we have investigated an approach for improving the performance of array signal processing by virtually increasing the number of channels [5]. The concept of a "virtual microphone array", which is an attempt to estimate or create an acoustic observation at a place where there are no actual microphones, can be found in other contexts such as introducing higher order statistics [6] or spatial sound acquisition [7, 8]. To increasing the number of linearly independent observations, we employed a linear interpolation in the complex logarithmic domain [5]. We also confirmed that there was an improvement in the speech enhancement performance when we used a maximum SNR beamformer. In a linear interpolation in the complex logarithmic domain, the phase is linearly interpolated because the phase is defined as the imaginary part of the complex logarithm. It is appropriate because the phase changes linearly between two microphones if a single plane wave arrives. However, there is no reason for the amplitude to be interpolated in the logarithmic scale.

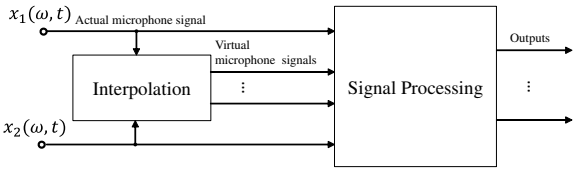


Fig. 1: Block diagram of signal processing with virtual microphone array technique

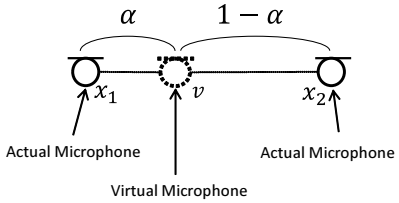


Fig. 2: Arrangement of actual and virtual microphones

In this paper, we consider a generalization of this virtual microphone array technique. By introducing β divergence, we propose a new nonlinear amplitude interpolation with adjustable nonlinearity. The logarithmic interpolation used in the previous work is given by $\beta = 1$, and the linear interpolation in the amplitude domain is given by $\beta = 2$. The speech enhancement performance with the maximum SNR beamformer is evaluated experimentally with different β values and different numbers of virtual channels.

2. VIRTUAL MICROPHONE BY INTERPOLATION

In our virtual microphone array technique, we create arbitrary channels of virtual microphone signals as a synthesis of 2

channels of actual microphone signals, and then we perform the array signal processing with an observed signal consisting of both actual and virtual microphone signals as shown in Fig. 1. Virtual microphone signals are generated as estimates of signals observed with a virtual microphone placed at a point where there is no actual microphone. A virtual microphone signal $v = v(\omega, t, \alpha)$ is generated as an estimated observation obtained with a virtual microphone placed at the $\alpha : (1-\alpha)$ internal division point of the positions of two actual microphones (Fig. 2). The simplest approach is linear interpolation as,

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

where $x_i = x_i(\omega, t)$ denotes the observed signal of the i-th actual microphone at frequency ω, and time t. However, virtual microphone signals derived from linear interpolation are linearly dependent and the linear interpolation does not provide any new statistical information for array signal processing. Thus, we generate a virtual microphone signal as an interpolation in the nonlinear function domain and we have already proposed a virtual microphone derived from complex logarithmic interpolation [5] written as

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

The complex logarithmic function derives the logarithmic amplitude and phase of signal as real and imaginary parts respectively as follows,

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

The interpolation consists of the linear interpolation of the phase angle and logarithmic interpolation of the amplitude and Eq. (2) is rewritten as

$$A_{\mathrm{v}} = \exp\left((1-\alpha)\log A_1 + \alpha \log A_2\right), \tag{4}$$
$$\phi_{\mathrm{v}} = (1-\alpha)\phi_1 + \alpha\phi_2, \tag{5}$$
$$v = A_{\mathrm{v}} \exp(j\phi_{\mathrm{v}}), \tag{6}$$

where $A_i = |x_i(\omega, t)|$ and $\phi_i = \angle x_i(\omega, t)$ represent for the amplitude and phase of the i-th actual microphone signal, respectively. The linear interpolation of phase angle in Eq. (5) is assumed to be appropriate with the assumption of a single plane wave propagation model because the phase difference between microphones is in linear relation to the microphone positions. In contrast, the logarithmic amplitude interpolation in Eq. (4) is not based on a specific model or there are no reasons for amplitude to be interpolated in the logarithmic scale. Therefore, in the next section we consider the extension of amplitude interpolation method by introducing β-divergence.

3. INTRODUCTION OF β-DIVERGENCE FOR AMPLITUDE INTERPOLATION

In this section, we introduce β-divergence for amplitude interpolation. β-divergence is widely used criteria for margin between nonnegative values such as amplitude. For instance, β-divergence is used as the cost function of nonnegative matrix factorization (NMF) [9, 10]. The β-divergence between a virtual microphone signal amplitude A_{v} and the i-th actual microphone signal amplitude A_i is defined as

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

D_β is continuous at $\beta = 0$ and $\beta = 1$. For β-divergence based interpolation, we consider the amplitude A_{v} to minimize the sum σ_β of the β-divergence between the amplitude values of an actual microphones signal and virtual microphone signal weighted by a virtual microphone position α,

$$\sigma_{D_\beta} = (1-\alpha) D_\beta(A_{\mathrm{v}}, A_1) + \alpha D_\beta(A_{\mathrm{v}}, A_2), \tag{8}$$
$$A_{\mathrm{v}\beta} = \mathrm{argmin}_{A_{\mathrm{v}}} \sigma_{D_\beta}. \tag{9}$$

By differentiating σ_{D_β} with A_{v} and setting it at 0, the amplitude interpolation extended with β-divergence is obtained as

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

Similarly to β-divergence D_β, $A_{\mathrm{v}\beta}$ is continuous at $\beta = 1$ as

$$\lim_{\beta\to 1}\left((1-\alpha) A_1^{\beta-1} + \alpha A_2^{\beta-1}\right)^{\frac{1}{\beta-1}} = \exp\left((1-\alpha)\log A_1 + \alpha \log A_2\right). \tag{11}$$

When β is set at 1, the interpolation is equal to the logarithmic interpolation noted in Eq. (4). The amplitude interpolation in Eq. (10) is assumed to be the $\beta - 1$ norm of a vector composed of amplitude weighted by α. Therefore, it also approaches the maximum selection and minimum selection taking the limit of $\beta \to +\infty$ and $\beta \to -\infty$ respectively as

$$A_{\mathrm{v}\beta} = \max(A_1, A_2)\,(\beta \to +\infty), \tag{12}$$
$$A_{\mathrm{v}\beta} = \min(A_1, A_2)\,(\beta \to -\infty). \tag{13}$$

The phase of a signal is linearly interpolated in a similar way to complex logarithmic interpolation as in Eq. (5), and a virtual microphone signal is obtained similarly to Eq. (6),

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

4. SPEECH ENHANCEMENT WITH MAXIMUM SNR BEAMFORMER

We apply the virtual microphone array technique to a maximum SNR beamformer [11] to evaluate the performance. A

maximum SNR beamformer requires the covariance matrices of the target-only period and the interference-only period as prior information of speech enhancement. However, it requires no information about sound direction such as steering vectors.

4.1. Construction of maximum SNR beamformer

Speech enhancement with a beamformer is realized by constructing a multichannel filter given by

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

to reduce the contamination of interference sources, where $w_i^*\left(\omega\right)$ is a filter for the i-th channel and $\{\cdot\}^*$ denotes a complex conjugation. The enhanced signals $y\left(\omega, t\right)$ are given as the inner product of the filter and the observed signal vector,

$$y\left(\omega, t\right) = \mathbf{w}^{\mathrm{H}}\left(\omega\right)\mathbf{x}\left(\omega, t\right). \tag{16}$$

In a maximum SNR beamformer, the filter $\mathbf{w}\left(\omega\right)$ is designed to maximize the ratio $\lambda\left(\omega\right)$ of the power between the target-only period Θ_{T}, and the interference-only period Θ_{I}:

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

where $\mathbf{R}_{\mathrm{T}}\left(\omega\right)$ and $\mathbf{R}_{\mathrm{I}}\left(\omega\right)$ represent the covariance matrices of the target-only period and interference-only periods, respectively. The covariance matrices are calculated as

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

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

where $\mathbf{x}_{\mathrm{T}}$ is the observed signal vector in the target-only period and $\mathbf{x}_{\mathrm{I}}$ is the observed signal vector in the interference-only period. The filter $\mathbf{w}\left(\omega\right)$ that maximizes the ratio $\lambda\left(\omega\right)$ is given as an eigenvector corresponding to the maximum eigenvalue of the following generalized eigenvalue problem;

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

4.2. Scaling compensation of beamformer

Since the maximum SNR beamformer $\mathbf{w}\left(\omega\right)$ has a scaling ambiguity, the beamformer is compensated in [12] as:

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

where $b_k\left(\omega\right)$ is the k-th component of $\mathbf{b}\left(\omega\right)$ given by

$$\mathbf{b}(\omega) = \frac{\mathbf{R}_{\mathbf{x}}(\omega)\mathbf{w}(\omega)}{\mathbf{w}^{\mathrm{H}}(\omega)\mathbf{R}_{\mathbf{x}}(\omega)\mathbf{w}(\omega)}, \tag{22}$$

$$\mathbf{R}_{\mathbf{x}}(\omega) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}(\omega, t)\mathbf{x}^{\mathrm{H}}(\omega, t). \tag{23}$$

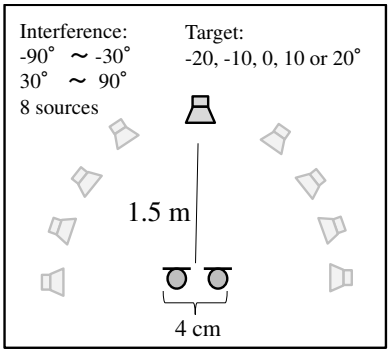


Fig. 3: Source and microphone layout in experiment

5. EXPERIMENTS

We conducted speech enhancement experiments with the maximum SNR beamformer to evaluate the performance with different β values.

5.1. Experimental conditions

The layout of the sources and actual microphones is shown in Fig. 3, and other experimental conditions are shown in Table 1. We used the 3 samples of Japanese and English speech for the target signals, and we performed 5 directions of arrival (DOA) experiments for each sample target signal, giving a total of 15 combinations of target DOA and speech samples. We used a mixture of 8 speech signals for the interference signal. The speech signals arrive from 8 different directions simultaneously. The observed signals were formed as the convolutive mixture of measured impulse responses and speech signals. We placed virtual microphones between two actual microphones at regular intervals, thus the parameter α of the i-th virtual microphone was given as

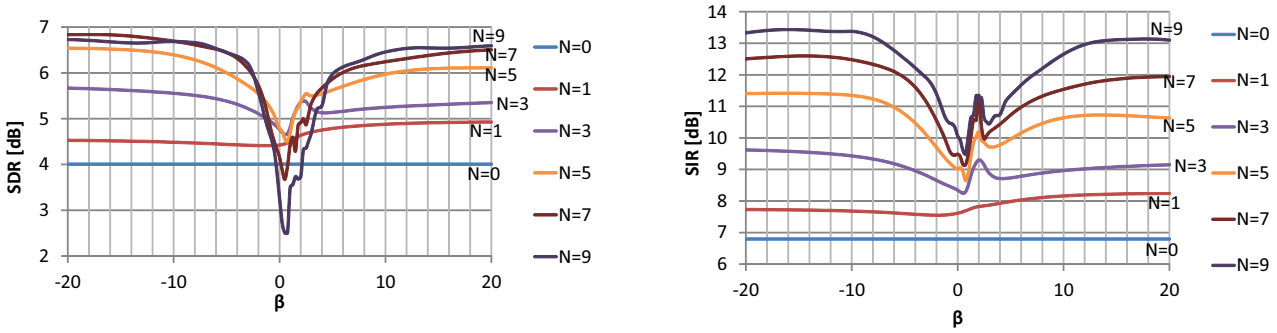
$$\alpha = \frac{i}{N+1}, \tag{24}$$

where N is the number of inserted virtual microphones. Speech enhancement was conducted with microphone arrays consisting of 2 actual microphones and N virtual microphones, thus giving $(N+2)$ channels in total. In this experiment, the first microphone was chosen as the reference ($k=1$) for scale compensation described in section 4.2. Unlike our previous work [5], we here performed the experiment without any regularization to the covariance matrices.

To evaluate the performance of the beamformer, we used an objective criterion, the signal-to-distortion ratio (SDR) and the signal-to-interference ratio (SIR) [13]. We show the mean SDR and SIR values for 15 combinations of target DOA and speech samples.

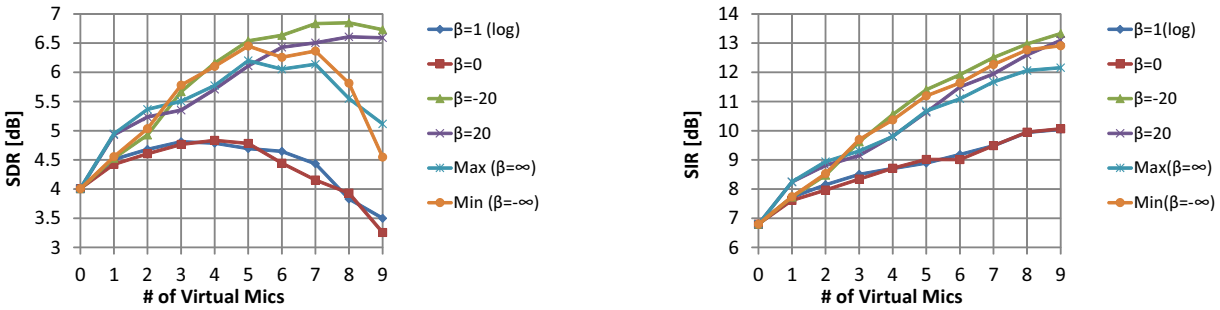
5.2. Results and discussion

Figure 4 shows the relation between the speech enhancement performance and β with different virtual microphone layouts,



(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

Table 1: Experimental conditions

# of actual microphones	2
# of virtual virtual microphones	0–9
Real microphone spacing	4 cm
Reverberation time	640 ms
Sampling rate	8 kHz
FFT frame length	1024 samples
FFT frame shift	256 samples
Test signal length	20 sec
Target only period length $\lvert\theta_{\mathrm{T}}\rvert$	10 sec
Interference only period length $\lvert\theta_{\mathrm{I}}\rvert$	10 sec

and Fig. 5 shows the relationship between performance and number of virtual microphones with several β values. Please note that when the number of virtual microphones is zero it means that the beamformer is processed solely with actual microphone signals and without virtual microphones. The SDR is improved when a few virtual microphones are inserted with any parameter β. The SIR is improved as the number of virtual microphones is increased. In contrast, the SDR is improved particularly when the parameter β is far from 0, and the SDR decreases with a parameter β of around 0 including a logarithmic interpolation ($\beta = 1$). The output sound of beamformer becomes distorted when a large number of virtual microphones is introduced with the parameter $\beta = 1$. However the distortion of output sound is reduced with the parameter is set at $\beta = -20$. The decay of SDR with a parameter β of around 0 is possibly caused by noise attributed to the rank deficiency of the covariance matrices. The parameter β seems to have an effect to adjust the degree of rank deficiency. Thus, we need to examine the relationship between parameter β and rank deficiency.

6. CONCLUSION

We proposed the generalization of the virtual microphone array technique that introduced β-divergence for the interpolation of amplitude. We compared the performance of a maximum SNR beamformer with different β values. With a conventional complex logarithmic interpolation ($\beta = 1$), the SDR has a peak about 1 dB higher than the actual microphone array. In contrast, the SDR is improved about 2.8 dB compared with a real microphone array with β set at -20 or $+20$. Therefore, we confirm the effectiveness of the introduction of β-divergence into the virtual microphone array technique.

7. REFERENCES

[1] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*, Springer, 2007.

[2] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans.*, vol. 34, no. 3, 1986.

[3] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[4] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Underdetermined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans.on Audio Speech Language Process*, vol. 18, no. 7, pp. 1830–1840, 2010.

[5] H. Katahira, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Virtually increasing microphone array elements by interpolation in complex-logarithmic domain," *Proc. EUSIPCO*, vol. TH-L 5.3, pp. 1–5, 2013.

[6] P. Chevalier, L. Albera, A. Ferréol, and P. Comon, "On the virtual array concept for higher order array processing," *IEEE Trans. Signal Processing*, vol. 53, no. 4, pp. 1254–1271, 2005.

[7] G. Del Galdo, O. Thiergart, T. Weller, and E. A. P. Habets, "Generating virtual microphone signals using geometrical information gathered by distributed arrays," *Proc. HSCMA*, pp. 185–190, 2011.

[8] K. Kowalczyk, A. Craciun, and E. A. P. Habets, "Generating virtual microphone signals in noisy environments," *Proc. EUSIPCO*, 2013.

[9] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," *Proc. IEEE MLSP*, pp. 283–288, 2010.

[10] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, 2011.

[11] H. L. Van Trees, *Optinum Array Processing*, John Wiley & Sons, 2002.

[12] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," *Proc. ICASSP*, vol. I, pp. 41–45, 2007.

[13] 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.