GENERALIZED AMPLITUDE INTERPOLATION BY $\beta$-DIVERGENCE FOR VIRTUAL MICROPHONE ARRAY

Hiroki Katahira\textsuperscript{1}, Nobutaka Ono\textsuperscript{2,3}, Shigeki Miyabe\textsuperscript{1}, Takeshi Yamada\textsuperscript{1}, Shoji Makino\textsuperscript{1}

\textsuperscript{1}University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8577 Japan
\textsuperscript{2}National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo, 101-8430 Japan
\textsuperscript{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 $\beta$-divergence. Improvement of the array signal processing performance is obtained by appropriate tuning of the parameter $\beta$. 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) \cite{1}, direction of arrival (DOA) estimation \cite{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 \cite{3} and multichannel Wiener filter \cite{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 \cite{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 \cite{6} or spatial sound acquisition \cite{7,8}. To increasing the number of linearly independent observations, we employed a linear interpolation in the complex logarithmic domain \cite{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.

In this paper, we consider a generalization of this virtual microphone array technique. By introducing $\beta$ 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 $\beta$ 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 consist-
ing of both actual and virtual microphone signals as shown in
Fig. 1. Virtual microphone signals are generated as esti-
mates of signals observed with a virtual microphone placed
at a point where there is no actual microphone. A virtual mi-
crophone 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,

\[
v = (1 - \alpha)x_1 + \alpha x_2,
\]

where \( x_i(\omega, t) \) denotes the observed signal of the \( i \)-th
actual microphone at frequency \( \omega \), 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

\[
v = \exp\left((1 - \alpha) \log x_1 + \alpha \log x_2\right).
\]

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

\[
\log x_1 = \log |x_1| + j \angle x_1.
\]

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

\[
A_v = \exp\left((1 - \alpha) \log A_1 + \alpha \log A_2\right),
\]

\[
\phi_v = (1 - \alpha) \phi_1 + \alpha \phi_2,
\]

\[
v = A_v \exp(j \phi_v),
\]

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, re-
spectively. 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 rea-
sons for amplitude to be interpolated in the logarithmic scale.
Therefore, in the next section we consider the extension of amplitude interpolation method by introducing \( \beta \)-divergence.

### 3. INTRODUCTION OF \( \beta \)-DIVERGENCE FOR
AMPLITUDE INTERPOLATION

In this section, we introduce \( \beta \)-divergence for amplitude in-
terpolation. \( \beta \)-divergence is widely used criteria for margin
between nonnegative values such as amplitude. For instance,
\( \beta \)-divergence is used as the cost function of nonnegative ma-
trix factorization (NMF) [9, 10]. The \( \beta \)-divergence between
a virtual microphone signal amplitude \( A_v \) and the \( i \)-th actual
microphone signal amplitude \( A_i \) is defined as

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

\( \beta \) is continuous at \( \beta = 0 \) and \( \beta = 1 \). For \( \beta \)-divergence
based interpolation, we consider the amplitude \( A_v \) to mini-
mimize the sum \( \sigma_\beta \) of the \( \beta \)-divergence between the ampli-
tude values of an actual microphone signal and virtual micro-
phone signal weighted by a virtual microphone position \( \alpha \),

\[
\sigma_{D_\beta} = (1 - \alpha) D_\beta (A_v, A_i) + \alpha D_\beta (A_v, A_2),
\]

\( A_v, A_2 = \arg\min_{A_v, A_2} \sigma_{D_\beta} \).

By differentiating \( \sigma_{D_\beta} \) with \( A_v \) and setting it at 0, the ampli-
itude interpolation extended with \( \beta \)-divergence is obtained as

\[
A_v, \beta = \begin{cases} 
\exp\left((1 - \alpha) \log A_1 + \alpha \log A_2\right) & (\beta = 1) \\
\left((1 - \alpha) A_1^{\beta-1} + \alpha A_2^{\beta-1}\right)^{1/\beta} & (\text{otherwise})
\end{cases}
\]

Similarly to \( \beta \)-divergence \( D_\beta \), \( A_v, \beta \) is continuous at \( \beta = 1 \) as

\[
\lim_{\beta \to 1} \left((1 - \alpha) A_1^{\beta-1} + \alpha A_2^{\beta-1}\right)^{1/\beta} = \exp\left((1 - \alpha) \log A_1 + \alpha \log A_2\right).
\]

When \( \beta \) 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 \( \alpha \). Therefore, it also approaches the
maximum selection and minimum selection taking the limit
of \( \beta \to +\infty \) and \( \beta \to -\infty \) respectively as

\[
A_v, \beta = \max (A_1, A_2) (\beta \to +\infty),
\]

\[
A_v, \beta = \min (A_1, A_2) (\beta \to -\infty).
\]

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

\[
v = A_v, \beta \exp(j \phi_v).
\]

### 4. SPEECH ENHANCEMENT WITH MAXIMUM
SNR BEAMFORMER

We apply the virtual microphone array technique to a maxi-
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

$$w(\omega) = \begin{bmatrix} w_1(\omega), \ldots, w_M(\omega) \end{bmatrix}^T,$$

(15)

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

$$y(\omega, t) = w^H(\omega)x(\omega, t).$$

(16)

In a maximum SNR beamformer, the filter \( w(\omega) \) is designed to maximize the ratio \( \lambda(\omega) \) of the power between the target-only period \( \Theta_T \), and the interference-only period \( \Theta_I \):

$$\lambda(\omega) = \frac{w^H(\omega)R_T(\omega)w(\omega)}{w^H(\omega)R_I(\omega)w(\omega)},$$

(17)

where \( R_T(\omega) \) and \( R_I(\omega) \) represent the covariance matrices of the target-only period and interference-only periods, respectively. The covariance matrices are calculated as

$$R_T(\omega) = \frac{1}{|\Theta_T|} \sum_{t \in \Theta_T} x_T(\omega, t)x_T^H(\omega, t),$$

(18)

$$R_I(\omega) = \frac{1}{|\Theta_I|} \sum_{t \in \Theta_I} x_I(\omega, t)x_I^H(\omega, t),$$

(19)

where \( x_T \) is the observed signal vector in the target-only period and \( x_I \) is the observed signal vector in the interference-only period. The filter \( w(\omega) \) that maximizes the ratio \( \lambda(\omega) \) is given as an eigenvector corresponding to the maximum eigenvalue of the following generalized eigenvalue problem;

$$R_T(\omega)w(\omega) = \lambda(\omega)R_I(\omega)w(\omega).$$

(20)

4.2. Scaling compensation of beamformer

Since the maximum SNR beamformer \( w(\omega) \) has a scaling ambiguity, the beamformer is compensated in [12] as:

$$w(\omega) \leftarrow b_k(\omega)w(\omega),$$

(21)

where \( b_k(\omega) \) is the \( k \)-th component of \( b(\omega) \) given by

$$b(\omega) = \frac{R_x(\omega)w(\omega)}{w^H(\omega)R_x(\omega)w(\omega)},$$

(22)

$$R_x(\omega) = \frac{1}{T} \sum_{t=1}^{T} x(\omega, t)x^H(\omega, t).$$

(23)

5. EXPERIMENTS

We conducted speech enhancement experiments with the maximum SNR beamformer to evaluate the performance with different \( \beta \) 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 \( \alpha \) of the \( i \)-th virtual microphone was given as

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

(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 \( \beta \) with different virtual microphone layouts,
The distortion of output sound is reduced with the parameter $\beta$ set at $\beta = -20$. The decay of SDR with a parameter $\beta$ of around 0 is possibly caused by noise attributed to the rank deficiency of the covariance matrices. The parameter $\beta$ seems to have an effect to adjust the degree of rank deficiency. Thus, we need to examine the relationship between parameter $\beta$ and rank deficiency.

6. CONCLUSION

We proposed the generalization of the virtual microphone array technique that introduced $\beta$-divergence for the interpolation of amplitude. We compared the performance of a maximum SNR beamformer with different $\beta$ 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 $\beta$ set at $-20$ or $+20$. Therefore, we confirm the effectiveness of the introduction of $\beta$-divergence into the virtual microphone array technique.

7. REFERENCES


**Table 1: Experimental conditions**

<table>
<thead>
<tr>
<th># of actual microphones</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td># of virtual virtual microphones</td>
<td>0–9</td>
</tr>
<tr>
<td>Real microphone spacing</td>
<td>4 cm</td>
</tr>
<tr>
<td>Reverberation time</td>
<td>640 ms</td>
</tr>
<tr>
<td>Sampling rate</td>
<td>8 kHz</td>
</tr>
<tr>
<td>FFT frame length</td>
<td>1024 samples</td>
</tr>
<tr>
<td>FFT frame shift</td>
<td>256 samples</td>
</tr>
<tr>
<td>Test signal length</td>
<td>20 sec</td>
</tr>
<tr>
<td>Target only period length $</td>
<td>\theta_T</td>
</tr>
<tr>
<td>Interference only period length $</td>
<td>\theta_I</td>
</tr>
</tbody>
</table>


