APSIPA Transactions on Signal and Information Processing, 2023, 12, e21

Original Paper

Wavelength-Proportional Interpolation and Extrapolation of Virtual Microphone for Underdetermined Speech Enhancement

Ryoga Jinzai[1], Kouei Yamaoka[1,2], Shoji Makino[1,3*], Nobutaka Ono[2], Mitsuo Matsumoto[1] and Takeshi Yamada[1]

[1] *University of Tsukuba, Japan*
[2] *Tokyo Metropolitan University, Japan*
[3] *Waseda University, Japan*

ABSTRACT

We previously proposed the virtual microphone technique to improve speech enhancement performance in underdetermined situations, in which the number of channels is virtually increased by estimating extra microphone signals at arbitrary positions along the straight line formed by real microphones. The effectiveness of the interpolation of virtual microphone signals for speech enhancement was experimentally confirmed. In this work, we apply the extrapolation of a virtual microphone as preprocessing of the maximum signal-to-noise ratio (SNR) beamformer and compare its speech enhancement performance (the signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR)) with that of using the interpolation of a virtual microphone. Furthermore, we aim to improve speech enhancement performance by solving a trade-off relationship between performance at low and high frequencies, which can be controlled by adjusting the virtual microphone interval. We propose a new arrangement where a virtual microphone is placed at a distance from the reference real microphone proportional to the

*Corresponding author: Shoji Makino, s.makino@ieee.org.

Received 22 November 2022; Revised 13 May 2023
ISSN 2048-7703; DOI 10.1561/116.00000078

wavelength at each frequency. From the results of our experiment in an underdetermined situation, we confirmed speech enhancement performance using the extrapolation of a virtual microphone is higher than that of using the interpolation of a virtual microphone. Moreover, the proposed wavelength-proportional interpolation and extrapolation method improves speech enhancement performance compared with the interpolation and extrapolation. Furthermore, we present the directivity patterns of a spatial filter and confirmed the behavior that improves speech enhancement performance.

Keywords: Virtual microphone, underdetermined situation, speech enhancement, beamforming, array signal processing.

1 Introduction

Signal processing using a microphone array includes various techniques such as blind source separation [5, 11, 12, 19, 20, 25, 28, 33], direction of arrival (DoA) estimation [3, 27, 30, 31, 40, 43, 44], and speech enhancement using a beamformer [9, 13, 17, 18, 37, 46, 47]. Basically, the performance of these techniques depends on the number of microphones. In other words, performance may degrade when the number of microphones is smaller than that of sound sources, which is called an underdetermined situation. On the other hand, portable recording devices such as smartphones and voice recorders, which usually have a small number of microphones, are widely used. Consequently, these techniques are prone to face an underdetermined situation in real environments. Several methods such as time–frequency masking [1, 4, 6–8, 24, 38, 39, 49, 53, 54], multichannel Wiener filtering [10, 14, 15, 41] and multichannel non-negative matrix factorization [36, 42] are effective in an underdetermined situation, although these methods tend to increase the distortion or computational complexity when trying to achieve high separation performance.

A simple solution to the above problem is to increase the number of microphones. However, this requires costly special equipment, such as a synchronized A/D converter, and a large amount of wiring. For this reason, we previously proposed the *virtual microphone* technique, in which the number of microphones is not actually but virtually increased [22, 26, 32, 52]. In this technique, additional observed signals, namely, virtual microphone signals, are estimated at arbitrary positions along the straight line formed by real microphones. Signal processing using a virtually extended microphone array is possible by using virtual microphone signals in addition to real microphone signals. The virtual microphone technique can be viewed as the time-frequency

switching beamformers [50, 51] which utilises the sparseness of the speech signals. Deep neural network-based virtual microphone technique, which used convTasnet, has also been proposed [35, 45].

The virtual microphone technique involves the interpolation and extrapolation of a virtual microphone depending on its position. In our previous studies, the interpolation was mainly used for speech enhancement [26, 52] and the extrapolation was mainly used for sound image localization [22, 32]. Thus, we apply the extrapolation of a virtual microphone to speech enhancement and compare its speech enhancement performance with that of using the interpolation of a virtual microphone.

The actual microphone array has an optimal placement of microphone for frequency. In general, at high frequencies, i.e., for a signal with a short wavelength, a shorter microphone interval is advantageous for preventing spatial aliasing and thus avoiding the degradation of speech enhancement performance. Conversely, at low frequencies, i.e., for a signal with a long wavelength, a longer interval is advantageous for obtaining a sufficient time difference, or equivalently, a sufficient phase difference to construct a spatial filter, which improves speech enhancement performance. This means that there is a trade-off relationship between performance at low and high frequencies, and it can be controlled by adjusting the microphone interval. To maximize observed phase differences while avoiding spatial aliasing, a microphone should be placed so that the microphone interval becomes half the wavelength at each frequency. In actual microphone array, a nonuniform-spacing microphone array has been used to deal with the trade-off relationship [16, 21, 29]. In this technique, a number of microphones are placed at nonuniform intervals, and signal processing is performed using a microphone pair with an appropriate microphone interval for each frequency band. This allows the microphone interval to be optimized for each frequency band, but inevitably requires many microphones, increasing the cost. Therefore, it is difficult to implement this technique on widely used small devices such as smartphones and voice recorders. On the other hand, in the virtual microphone technique, the virtual microphone can be placed at any position on the same straight line as the real microphones by the interpolation and extrapolation. In addition, since the virtual microphone signal is independently estimated at each frequency bin, it is possible to change the position of the virtual microphone at each frequency. On this basis, we propose a new technique of virtual microphones that solves the trade-off relationship between performance at low and high frequencies, namely, wavelength-proportional virtual microphone (WPVM) technique [23]. Here, the virtual microphone is placed at a distance from the reference real microphone proportional to the wavelength at each frequency for speech enhancement, and both interpolation and extrapolation are used.

In this study, we evaluate speech enhancement performance using the maximum signal-to-noise ratio (SNR) beamformer [2, 46] by the virtual microphone

technique, where we use the extrapolation of a virtual microphone and WPVM technique. First, to examine the effectiveness and robustness against the directions of target and interferer sound sources, we perform speech enhancement with the extrapolated virtual microphones and WPVM in various acoustic environments. Next, we present the directivity patterns of spatial filters to illustrate the effect of WPVM technique on filter design.

The structure of this paper is as follows. In Section 2, we explain the virtual microphone technique. In Section 3, we propose WPVM technique. In Section 4, we explain the maximum SNR beamformer. In Section 5, we experimentally evaluate the performance of the extrapolation of the virtual microphone and WPVM technique. Additionally, we present the directivity patterns to confirm the behavior of those methods. Finally, the paper is concluded in Section 6.

2 Virtual Microphone Technique

2.1 Preliminary

In this section, we introduce the virtual microphone technique involving interpolation based on β-divergence [26, 52] and extrapolation of a virtual microphone [22]. In this paper, the interpolation and extrapolation of a virtual microphone, which are conventional methods using the same position of the virtual microphone at all frequencies, are collectively referred to as the fixed virtual microphone technique. In this technique, all microphone signals are processed in the time–frequency domain. A virtual microphone signal $v(\omega, t)$ is generated from the observed signals of two real microphones $x_i(\omega, t)$, where $x_i(\omega, t)$ is the ith microphone signal $(i = 1, 2)$ at angular frequency ω in the tth time frame. The number of channels of the microphone array is virtually increased by using virtual microphone signals in addition to the real microphone signals. The arrangement of the real and virtual microphones is shown in Figure 1, where α is a coefficient that determines the position of the virtual microphone.

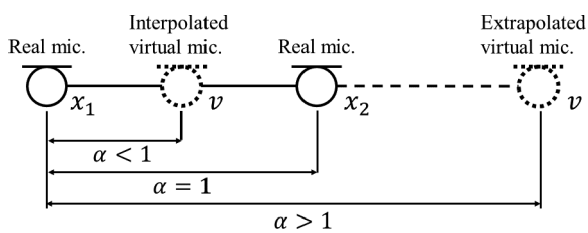


Figure 1: Arrangement of real and virtual microphones.

In an environment where there are multiple sounds arriving from different directions, the relationship between the microphone position and the observed signal is generally complicated. In the virtual microphone technique, by assuming W-disjoint orthogonality (W-DO)[53] for mixed signals, we can simplify the model of the observed signal. W-DO indicates the strong sparsity of a signal in the time–frequency domain, i.e., the component from a sound source dominates one time–frequency slot. By assuming W-DO, even when multiple sounds arrive, we can regard them as a single sound in each time–frequency slot.

In this technique, the phase and amplitude of a virtual microphone signal are estimated individually. Here, different models can be applied for the phase and amplitude estimation, making the generation of the virtual microphone signals simple. Additionally, this formulation naturally leads to the nonlinearity of generation of virtual microphone signals, which is an essential property to apply this technique as preprocessing in linear signal processing. Here, the phase and amplitude of $x_i(\omega, t)$ are respectively defined as

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

$$A_i = |x_i(\omega, t)|. \tag{2}$$

2.2 Estimation of Phase of Virtual Microphone Signal

When a sound wave arrives from a sufficient distance relative to the microphone interval, the propagating wave can be approximated as a plane wave. In both interpolation and extrapolation, we can estimate the phase ϕ_v of the virtual microphone signal using the linear equation

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

The phase has the value $\phi_i \pm 2\pi n$, where n is an arbitrary natural numbers. Thus, the phase of the virtual microphone signal is estimated under the assumption of

$$|\phi_1 - \phi_2| \leq \pi. \tag{4}$$

2.3 Estimation of the Amplitude of Virtual Microphone Signal

In the estimation of the amplitude of the virtual microphone signal, the formulas are different for interpolation and extrapolation.

The physical modeling of the amplitude difference is not as simple as that of the phase difference because the amplitude depends on the distance between the source and the microphones in addition to the DOA. Thus, instead of

interpolation based on some physical assumption, amplitude interpolation based on β-divergence, which has simple processing and parameter adjustment, was proposed [26].

β divergence is a widely used distance measure for nonnegative values such as amplitude. For instance, β divergence is used as the cost function for nonnegative matrix factorization (NMF). β divergence is equivalent to Itakura-Saito divergence ($\beta = 0$), Kullback-Leibler divergence ($\beta = 1$), and Euclidean divergence ($\beta = 2$). Note that β divergence also corresponds to the far-field model ($\beta = 2$) and the near-field model. The β divergence between the signal amplitude of a virtual microphone A_{v} and that of the ith real microphone A_i is defined as

$$D_\beta\left(A_{\mathrm{v}}, A_i\right) = \begin{cases} A_{\mathrm{v}}\left(\log A_{\mathrm{v}} - \log A_i\right) + \left(A_i - A_{\mathrm{v}}\right) & (\beta = 1), \\ \dfrac{A_{\mathrm{v}}}{A_i} - \log\dfrac{A_{\mathrm{v}}}{A_i} - 1 & (\beta = 0), \\ \dfrac{A_{\mathrm{v}}^\beta}{\beta\left(\beta - 1\right)} + \dfrac{A_i^\beta}{\beta} - \dfrac{A_{\mathrm{v}} A_i^{\beta-1}}{\beta - 1} & (\text{otherwise}). \end{cases} \tag{5}$$

Note that D_β is continuous at $\beta = 0$ and $\beta = 1$. For β-divergence-based interpolation, we derive the amplitude A_{v} that minimizes the sum σ_{D_β} of the β divergence between the amplitude of a real microphones signal and a virtual microphone signal weighted by the virtual microphone interpolation parameter α,

$$\sigma_{D_\beta} = \left(1 - \alpha\right) D_\beta\left(A_{\mathrm{v}}, A_1\right) + \alpha D_\beta\left(A_{\mathrm{v}}, A_2\right), \tag{6}$$

$$A_{\mathrm{v}\beta} = \mathrm{argmin}_{A_{\mathrm{v}}} \sigma_{D_\beta}. \tag{7}$$

Differentiating σ_{D_β} with respect to A_{v} and setting it to 0, the interpolated amplitude extended using β divergence is obtained as

$$A_v = \begin{cases} \exp\left(\left(1 - \alpha\right)\log A_1 + \alpha \log A_2\right) & (\beta = 1) \\ \left(\left(1 - \alpha\right) A_1^{\beta-1} + \alpha A_2^{\beta-1}\right)^{\frac{1}{\beta-1}} & (\text{otherwise}), \end{cases} \tag{8}$$

where $0 < \alpha < 1$.

The result of β-divergence-based interpolation is assumed to be the $\beta - 1$ norm of the vector $\left[\left(1 - \alpha\right) x_1, \alpha x_2\right]^{\mathrm{T}}$, which is composed of the amplitude weighted by α. Therefore, taking the limits of $\beta \to +\infty$ and $\beta \to -\infty$, the interpolation corresponds to the selection of the following maximum and minimum values, respectively:

$$A_{\mathrm{v}} = \begin{cases} \max\left(A_1, A_2\right) & (\beta \to +\infty), \\ \min\left(A_1, A_2\right) & (\beta \to -\infty). \end{cases} \tag{9}$$

Note that the linear interpolation of the phase angle is defined in the domain of arbitrary real numbers α, not only in the range $0 \leq \alpha \leq 1$. On the other hand, the β-divergence-based interpolation of the amplitude is defined only in the domain of $0 \leq \alpha \leq 1$ when β is set to $\beta \neq 1$.

For the extrapolation, the conceivable amplitude of the virtual microphone is more complex than that for the interpolation. When (8) is applied to extrapolation, it may output unrealistic amplitudes such as a complex amplitude, a negative amplitude, or an amplitude diverging to positive infinity except for $\beta = 1$. Therefore, in this study, as the simplest way to avoid these problems, we restricted to using the amplitude of the signal of a real microphone that is closer to the position of the virtual microphone. Thus, the amplitude of the extrapolated virtual microphone signal is [22, 32]

$$A_v = \begin{cases} A_1 & (\alpha < 0) \\ A_2 & (\alpha > 1). \end{cases} \tag{10}$$

2.4 Estimation of Virtual Microphone Signal

From the above, the virtual microphone signal $v(\omega, t, \alpha)$ is represented as

$$v(\omega, t, \alpha) = A_v \exp(j\phi_v). \tag{11}$$

When we need many virtual microphones, we can use an arbitrary number of α values to generate the same number of virtual microphones.

3 Wavelength-Proportional Virtual Microphone

As mentioned in the introduction, there is a trade-off relationship between performances at low and high frequencies in array signal processing techniques. For example, at high frequencies and short wavelengths, a shorter microphone interval prevents spatial aliasing. Conversely, at low frequencies and long wavelengths, a longer microphone interval provides a sufficient phase difference as spatial information.

In this paper, we propose a new arrangement, in which the position of the virtual microphone is proportional to the wavelength at each frequency [23]. We call the proposed technique the wavelength-proportional virtual microphone (WPVM). The arrangement of real and virtual microphones is shown in Figure 2.

In this method, the coefficient of the position of the virtual microphone α is given by

$$\alpha(\omega) = \frac{\lambda(\omega)k}{d} = \frac{2\pi ck}{\omega d}, \tag{12}$$

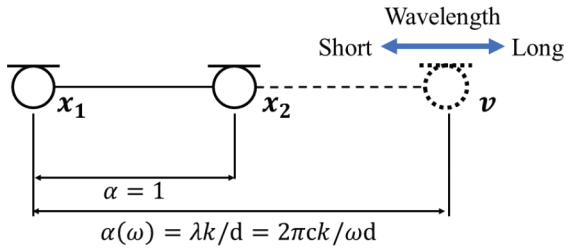


Figure 2: Arrangement of real microphones and a wavelength-proportional virtual microphone.

where λ is wavelength, d is the distance between the real microphones, k is a wavelength coefficient, and c is the speed of sound. The wavelength coefficient k is the interval between reference microphone x_1 and the virtual microphone v relative to the wavelength $\lambda(\omega)$. This equation means that the virtual microphone is placed at a position k times the wavelength corresponding to the frequency to be processed; thus, the total length of the microphone array including the virtual microphone is large at low frequencies and small at high frequencies. For example, when $k = 0.5$, the position of the virtual microphone is 42.5 cm at 400 Hz, 17 cm at 1 kHz, and 4.25 cm at 4 kHz. In this case, the maximum phase difference between x_1 and v is π, so spatial aliasing does not occur at all frequencies.

4 Maximum SNR Beamformer

In this study, to evaluate the performance of the extrapolation of the virtual microphone and WPVM technique, we carry out the extrapolation and WPVM technique as preprocessing of the maximum SNR beamformer [2, 46]. The advantage of this beamformer is that it does not explicitly require the direction of sound sources.

In speech enhancement by a beamformer, the multichannel filter $\boldsymbol{w}(\omega)$ is constructed for the N-channel observation signals $\boldsymbol{x}(\omega, t)$.

$$\boldsymbol{x}(\omega, t) = [x_1(\omega, t), \cdots, x_N(\omega, t)]^T \tag{13}$$

$$\boldsymbol{w}(\omega) = [w_1(\omega, t), \cdots, w_N(\omega, t)]^T \tag{14}$$

where ^{T} stands for the transposition of a vector. The sound $y(\omega, t)$ in which the target sound is enhanced can be obtained by applying filter $\boldsymbol{w}(\omega)$ to observed signal $\boldsymbol{x}(\omega, t)$ as follows:

$$y(\omega, t) = \boldsymbol{w}^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega, t). \tag{15}$$

The construction of the maximum SNR beamformer requires prior information on the spatial covariance matrices of the target-active period $\mathbf{R}_{\mathrm{T}}(\omega)$ and target-inactive period $\mathbf{R}_{\mathrm{I}}(\omega)$. From this information, the maximum SNR beamformer constructs a filter so that the SNR, $\gamma(\omega)$, of the target to the interference signal becomes maximum as follows:

$$\gamma(\omega) = \frac{\boldsymbol{w}^{\mathrm{H}}(\omega)\mathbf{R}_{\mathrm{T}}(\omega)\boldsymbol{w}(\omega)}{\boldsymbol{w}^{\mathrm{H}}(\omega)\mathbf{R}_{\mathrm{I}}(\omega)\boldsymbol{w}(\omega)}. \tag{16}$$

Although, a constructed spatial filter $\boldsymbol{w}(\omega)$ has a scaling ambiguity in the maximum SNR beamformer, a compensation method was proposed in [2].

When the virtual microphone technique is used, the observed signals including virtual microphone signal and the constructed filters are

$$\boldsymbol{x}(\omega, t) = [x_1(\omega, t), \cdots, x_N(\omega, t), x_v(\omega, t)]^T \tag{17}$$

$$\boldsymbol{w}(\omega) = [w_1(\omega, t), \cdots, w_N(\omega, t), w_v(\omega, t)]^T. \tag{18}$$

Thus, the enhanced signal can be obtained by (15). The virtual microphone technique can be similarly applied to other microphone array signal processing techniques as well as the maximum SNR beamformer.

5 Experiment

In the experiment, we compared speech enhancement performance of the maximum SNR beamformer using the extrapolation of the virtual microphone with that using the interpolation. Furthermore, we also evaluated the enhancement performance with the WPVM technique.

5.1 Experimental Conditions

The layout of the sound sources, which is set up to consider speech enhancement in a conversational scene, is shown in Figure 3. One target speaker and two interferers are assumed to be in the scene. Furthermore, two real microphones, M_1 and M_2, and one virtual microphone, M_{v}, are assumed, as shown in the figure. Other experimental conditions are listed in Table 1. The sampling frequency is 8 kHz, so spatial aliasing would occur if the interval between the two real microphones were longer than 4.25 cm. The interval between the real microphones is 2.83 cm, thus there is no spatial aliasing between them. We use four female speeches and four male speeches as a target source, and a female speech and a male speech as each interference source. Half the speeches are in Japanese and the other half are in English. In total, 32 ($8 \times 2 \times 2$) combinations of target and interference speeches are used for the experiment.

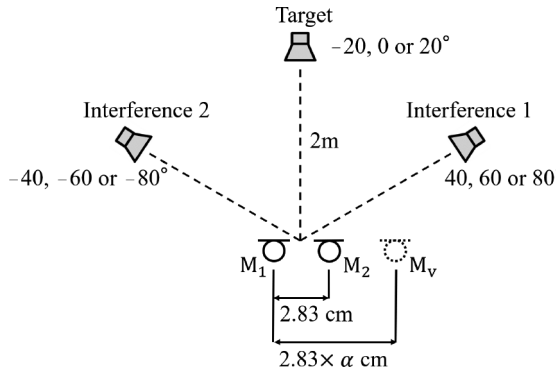


Figure 3: Layout of sound sources and microphones.

Table 1: Experimental Conditions.

Number of real microphones	2
Number of virtual microphones	1
Input SNR	0 dB
Sampling rate	8 kHz
Interval between real microphones	2.83 cm
Reverberation time T_R	300 ms
FFT frame length/shift	1024/256 samples
Number of target speech types	8
Number of interference speech types	2

The target speaker is located in three directions, that is, at azimuth of 0° (front), −20° (left), and 20° (right), as shown in Figure 3. The same applies to the two interferers. Therefore, a total of 27 (3×3×3) combinations of target and interferer directions are examined. The observed signals are simulated by convolving the speeches and a set of measured impulse responses in the RWCP Sound Scene Database [34]. The impulse responses are measured in a room (T_{60} = 300 ms).

In the experiment to compare speech enhancement performance between the case of using interpolation and extrapolation, the coefficient of the position of the virtual microphone α is varied from 0.1 to 30 (i.e., the interval between M_1 and M_v is varied from 0.283 cm to 84.9 cm), where $0 < \alpha < 1$ indicates interpolation and $\alpha > 1$ indicates extrapolation. $\alpha = 1$ indicates that no virtual microphone is used (i.e., only the two real microphones are used). In the interpolation, since it has been experimentally confirmed that $\beta = -20$ provides the highest performance [26], we set β to −20. In the evaluation of speech enhancement performance with the WPVM technique, to compare

differences in performance owing to k, the wavelength coefficient k is set to 0.25, 0.5, 1, and 2. The SNR of the target signal to interference signals is set to 0 dB. To evaluate speech enhancement performance, we use the signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) as objective evaluation criteria [48]. A concise representation of the results is obtained by averaging these criteria over 864 (32×27) trials for speakers and directions.

5.2 Results

Figure 4 shows the relationship between the coefficient of the virtual microphone α and the SDR and SIR. Note that the horizontal axis has a logarithmic scale.

The curved line indicates speech enhancement performance using the fixed virtual microphone technique, which uses the same value of α for all frequencies. According to Figure 4, the SDR was improved by up to 1.5 dB compared with that without the virtual microphone ($\alpha = 1$) by using interpolation ($\alpha < 1$), whereas it was improved by up to about 2.5 dB by using extrapolation ($\alpha > 1$), i.e., the SDR is 1 dB higher when using extrapolation than when using interpolation. Similarly, the SIR was improved by 2.5 and 4.5 dB by using interpolation and extrapolation, respectively. From these results, it can be seen that the extrapolation of the virtual microphone is more effective than the interpolation.

5.3 Discussion

To clarify the reason underlying these results, we illustrate the directivity patterns of the spatial filter of the maximum SNR beamformer. We focused on a specific combination of directions of the target and interference sources: 0° for the target, 60° for interference 1, and -60° for interference 2. Speech enhancement performance of each method is shown in Figure 5, and is similar to that in Figure 4.

The other straight lines show the enhancement performance when using WPVM technique with wavelength coefficient k. In the evaluation of WPVM technique, for the SDR, it is confirmed that the performance is highest for $k = 0.5$. Its performance was improved by up to 1.3 and 0.3 dB compared with those of interpolation and extrapolation, respectively. In contrast, for the SIR, the performance is highest for $k = 1$. Its performance was improved by up to 3 and 1 dB compared with those of interpolation and extrapolation, respectively. In contrast, the results for $k = 2$ and $k = 0.25$ were inferior to that using the fixed virtual microphone technique for some values of α.

For the same combination of directions as above, the directivity patterns of the maximum SNR beamformer obtained by using interpolation and extrapolation, and WPVM technique are respectively shown in Figures 6 and 7,

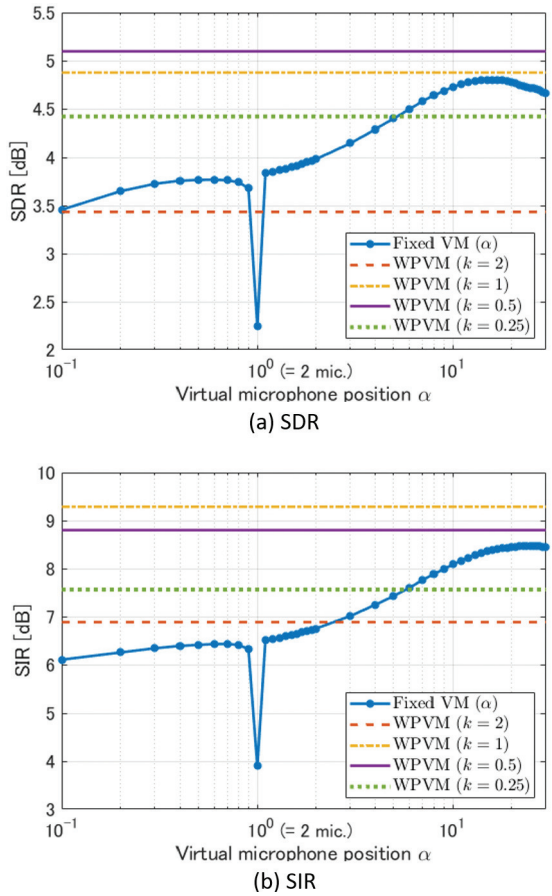


Figure 4: Relationship between coefficient of virtual microphone α and average speech enhancement performance.

where the values of α with the best enhancement performance were selected for interpolation and extrapolation.

According to Figure 6 (a), the spatial filter with the interpolated virtual microphone ($\alpha = 0.4$ at all frequencies) has nulls in the frequency range from 1 to 4 kHz and no nulls at frequencies below 1 kHz. This means that sounds below 1 kHz cannot be sufficiently suppressed. According to Figure 6 (b), the spatial filter with the extrapolated virtual microphone ($\alpha = 13$ at all frequencies) has many sharp nulls, which implies the occurrence of spatial aliasing. As a result, sounds from various directions, such as those near the target source, are suppressed in addition to the interference sound. However, unlike in interpolation, nulls exist even at frequencies below 1 kHz, which means

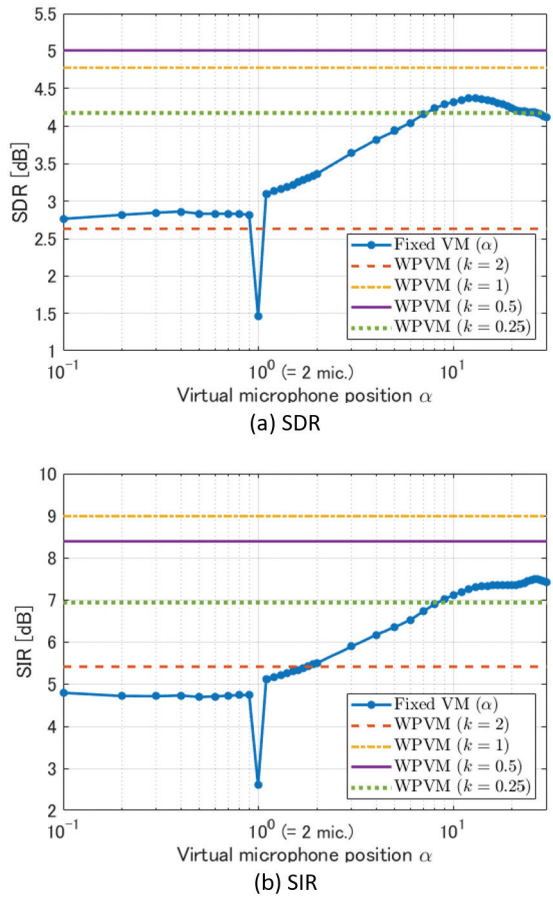


Figure 5: Relationship between coefficient of virtual microphone α and speech enhancement performance in the situation where 0° for target, 60° for interference 1, and −60° for interference 2.

that sounds below 1 kHz can be appropriately suppressed. In general, human speech has more energy at low frequencies than at high frequencies. Since the beamformer with extrapolation can improve the performance at low frequencies by widening the microphone interval, we conclude that the extrapolation contributes to the improvement of speech enhancement performance.

For the beamformer with WPVM technique (Figure 7), four sharp nulls are found in the directivity pattern for $k = 2$ (Figure 7 (a)). This indicates the occurrence of spatial aliasing at all frequencies. On the other hand, two fuzzy nulls are found in the directivity pattern for $k = 0.25$ (Figure 7 (d)), which indicates that the phase difference between each microphone is too small at all frequencies to construct a spatial filter with sharp nulls. In contrast, two nulls

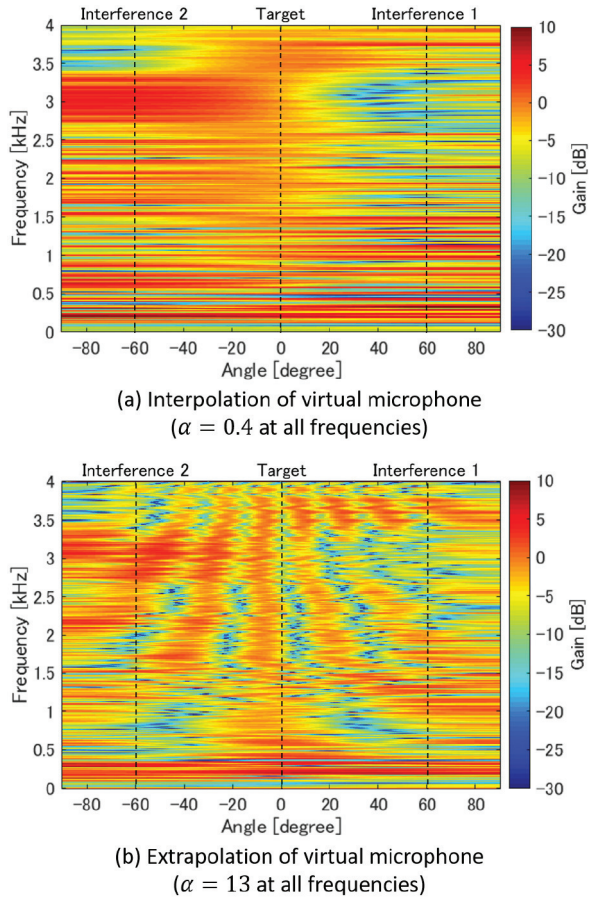
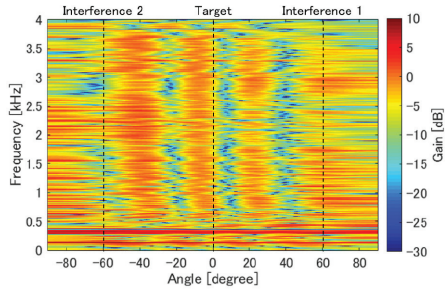


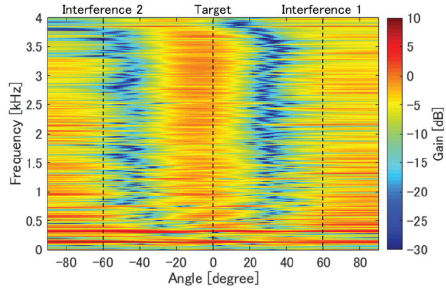
Figure 6: Directivity patterns of beamformer with fixed virtual microphone.

are clearly observed for $k = 0.5$ (Figure 7 (c)). Moreover, two belt-shaped nulls are clearly observed for $k = 1$ (Figure 7 (b)) indicating that no spatial aliasing occurs. As the reason for the improved speech enhancement performance, by using an appropriate k, it is possible to maximize the observed phase difference within a range where spatial aliasing does not occur, thereby making it possible for the beamformer to generate sharp nulls. As a feature of the directivity patterns for WPVM technique, similar directivity is found at all frequencies, indicating that it has directivity characteristics independent of frequency.

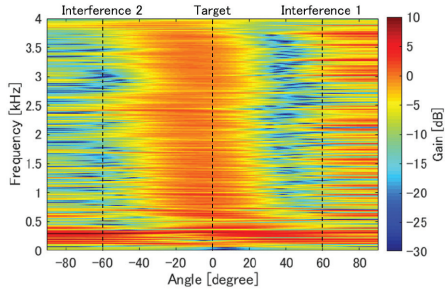
For these results, the nulls tend to slightly deviate from the direction of the interference sound sources. We attribute this to the effect of room reverberation, which is known to introduce bias.



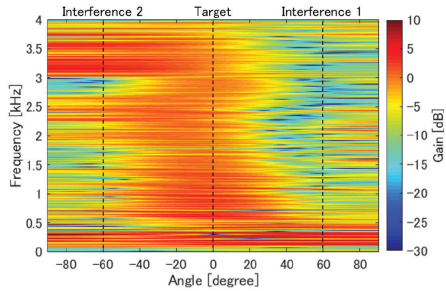
(a) $k = 2$



(b) $k = 1$



(c) $k = 0.5$



(d) $k = 0.25$

Figure 7: Directivity patterns of beamformer with fixed virtual microphone.

Summarizing this discussion, when the microphone interval is small, an insufficient phase difference between microphones exists at low frequencies, resulting in nulls not being properly generated. In contrast, when the microphone interval is large, spatial aliasing occurs at high frequencies. WPVM technique using an appropriate wavelength coefficient k can cope with these two problems; thus, this method shows the highest performance.

6 Conclusion

In this paper, we applied extrapolation of a virtual microphone with the maximum SNR beamformer to speech enhancement in an underdetermined situation, and confirmed that its speech enhancement performance is better than that with interpolation of a virtual microphone. In addition, we proposed a new arrangement where a virtual microphone is placed at a distance from the reference real microphone proportional to the wavelength at each frequency. The advantages of this method are that no spatial aliasing occurs and the phase difference between microphones is sufficient to construct a spatial filter at all frequencies by setting an appropriate wavelength coefficient k.

In the experiment, we evaluated speech enhancement performance on the basis of the SDR and SIR in an underdetermined situation. By comparing the proposed method with the conventional method, we found that the SDR was improved by about 1.3 dB and the SIR by about 3 dB. These results indicate that the proposed WPVM technique is effective for speech enhancement using the maximum SNR beamformer in an underdetermined situation.

Abbreviations

SNR: Signal-to-noise-ratio; VM: Virtual microphone; WPVM: Wavelength-proportional virtual microphone; W-DO: W-disjoint orthogonality; SDR: Signal-to-distortion ratio; SIR: Signal-to-interference ratio

Availability of Data and Materials

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Author's Contributions

Ryoga Jinzai performed the experiments and wrote the majority of the manuscript, and other authors reviewed and revised the manuscript. All

authors made contributions to the conception and design of the work, analyzed the data, and interpreted the results. All authors read and approved the final manuscript.

Funding

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI through a Grant-in-Aid for Scientific Research under Grants 16H01735 and 19H04131 and 19J20420, and the SECOM Science and Technology Foundation.

Biographies

Ryoga Jinzai received M.Eng. degrees from University of Tsukuba in 2020. His research interests include acoustic signal processing, specifically, microphone array signal processing. He had been engaged in research of speech enhancement with microphone array while he was in graduate school.

Kouei Yamaoka received the B.Sc. and M.E. degrees in information engineering and engineering from the University of Tsukuba, Tsukuba, Japan, in 2017 and 2019, respectively. He is currently working toward the Ph.D. degree with Tokyo Metropolitan University, Hino, Japan. His research interests include acoustic signal processing, signal enhancement, source localization, and asynchronous distributed microphone array. Mr. Yamaoka is a member of the Acoustical Society of Japan.

Shoji Makino received the B.E., M.E., and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1979, 1981, and 1993, respectively. He joined NTT in 1981, and the University of Tsukuba in 2009. He is currently a Professor at Waseda University. He has authored or coauthored more than 400 papers in journals and conference proceedings and is responsible for more than 200 patents. His research interests include adaptive filtering technologies, the realization of acoustic echo cancellation, blind source separation of convolutive speech mixtures, and acoustic signal processing for speech and audio applications. He was a recipient of 30 Awards, including the IEEE SPS Leo L. Beranek Meritorious Service Award in 2022, the IEEE SPS Best Paper Award in 2014, the IEEE MLSP Competition Award in 2007, and the ICA Unsupervised Learning Pioneer Award in 2006. He was an IEEE SPS Distinguished Lecturer (2009–2010), an IEEE Fellow, an IEICE Fellow, and an APSIPA Board of Governor (2023–).

Nobutaka Ono received the B.E., M.S., and Ph.D. degrees from the University of Tokyo, Japan, in 1996, 1998, 2001, respectively. He became a research associate in 2001 and a lecturer in 2005 in the University of Tokyo. He moved to the National Institute of Informatics in 2011 as an associate professor, and moved to Tokyo Metropolitan University in 2017 as a full professor. His research interests include acoustic signal processing, machine learning, and optimization algorithms for them. He was a chair of Signal Separation Evaluation Campaign evaluation committee in 2013 and 2015, and an Associate Editor of the IEEE Transactions on Audio, Speech and Language Processing during 2012 to 2015. He is a senior member of the IEEE Signal Processing Society and a member of IEEE Audio and Acoustic Signal Processing Technical Committee from 2014. He received the unsupervised learning ICA pioneer award from SPIE.DSS in 2015.

Mitsuo Matsumoto received M.S. degree in computer science from Oita University. In 1985, he joined Victor Company of Japan Ltd. (JVC) as an audio and acoustical engineer; where he engaged in research of sound image localization, sound field control, concert hall acoustics, and watermarking technology. In 2006, he received Ph.D. degree from Chiba Institute of Technology. He lectured in Universities and national institute of technology. He joined the University of Tsukuba in 2017 as a research member. He was awarded as an inventor of a profitable patent by Japan Institute of Invention and Innovation. He was a director of the AES Japan Section from 1997 to 2001.

Takeshi Yamada received the B.Eng. degree from Osaka City University, Japan, in 1994, and the M.Eng. and Dr.Eng. degrees from the Nara Institute of Science and Technology, Japan, in 1996 and 1999, respectively. He is currently an associate professor with the Institute of Systems and Information Engineering, University of Tsukuba, Japan. His research interests include speech recognition, sound scene understanding, multichannel signal processing, and media quality assessment. He is a member of the IEEE, IEICE, IPSJ, and ASJ.

References

[1] F. Abrard and Y. Deville, "A time–frequency blind signal separation method applicable to underdetermined mixtures of dependent sources," *Signal Processing*, 85(7), 2005, 1389–403.

[2] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," in *Proc. ICASSP*, Vol. 1, 2007, 41–4.

[3] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proc. ICASSP*, Vol. 5, 2006, 33–6.

[4] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, 87(8), 2007, 1833–47.

[5] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*, Springer Science & Business Media, 2005.

[6] P. Bofill, "Underdetermined blind separation of delayed sound sources in the frequency domain," *Neurocomputing*, 55(3-4), 2003, 627–41.

[7] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform," in *Proc. ICA*, Vol. 2000, 2000, 87–92.

[8] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, 81(11), 2001, 2353–62.

[9] M. Brandstein and D. Ward, *Microphone Arrays*, Springer, 2001.

[10] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. on Audio, Speech, and Language Processing*, 14(4), 2006, 1218–34.

[11] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, 36(3), 1994, 287–314.

[12] P. Comon, C. Jutten, and J. Herault, "Blind separation of sources, Part II: Problems statement," *Signal Processing*, 24(1), 1991, 11–20.

[13] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, 32(2), 2015, 18–30.

[14] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. on Signal Processing*, 50(9), 2002, 2230–44.

[15] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. on Audio, Speech, and Language Processing*, 18(7), 2010, 1830–40.

[16] G. W. Elko, "Microphone array systems for hands-free telecommunication," *Speech Communication*, 20(3-4), 1996, 229–40.

[17] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, 60(8), 1972, 926–35.

[18] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas and Propagation*, 30(1), 1982, 27–34.

[19] S. Haykin, *Unsupervised adaptive filtering, volume I: blind source separation*, John Wiley & Sons, 2000.

[20] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, John Wiley & Sons, 2001.

[21] M. Inoue, S. Nakamura, T. Yamada, and K. Shikano, "Microphone array design measures for hands-free speech recognition," in *Proc. European Conference on Speech Communication and Technology*, 1997, 331–4.

[22] R. Jinzai, K. Yamaoka, M. Matsumoto, T. Yamada, and S. Makino, "Microphone position realignment by extrapolation of virtual microphone," in *Proc. APSIPA ASC*, 2018, 367–72.

[23] R. Jinzai, K. Yamaoka, M. Matsumoto, T. Yamada, and S. Makino, "Wavelength proportional arrangement of virtual microphones based on interpolation/extrapolation for underdetermined speech enhancement," in *Proc. EUSIPCO*, 2019, 1–5.

[24] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. ICASSP*, Vol. 5, 2000, 2985–8.

[25] C. Jutten and J. Herault, "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, 24(1), 1991, 1–10.

[26] 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*, 2016(1), 2016, 11.

[27] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 24(4), 1976, 320–7.

[28] T.-W. Lee, *Independent component analysis; Theory and Applications*, Kluwer Academic Publishers, 1998.

[29] Q.-G. Liu, B. Champagne, and P. Kabal, "A microphone array processing technique for speech enhancement in a reverberant space," *Speech Communication*, 18(4), 1996, 317–34.

[30] A. Lombard, H. Buchner, and W. Kellermann, "Multidimensional localization of multiple sound sources using blind adaptive MIMO system identification," in *Proc. IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2006, 7–12.

[31] A. Lombard, T. Rosenkranz, H. Buchner, and W. Kellermann, "Multidimensional localization of multiple sound sources using averaged directivity patterns of blind source separation systems," in *Proc. ICASSP*, 2009, 233–6.

[32] N. Mae, K. Yamaoka, Y. Mitsui, M. Matsumoto, S. Makino, D. Kitamura, N. Ono, T. Yamada, and H. Saruwatari, "Ego noise reduction and sound localization adapted to human ears using hose-shaped rescue robot," in *Proc. International Workshop on Nonlinear Circuits, Communications and Signal Processing*, 2018, 371–4.

[33] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*, Springer, 2007.

[34] 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," in *Proc. IEEE International Conference on Multimedia and Expo*, Vol. 2, 2002, 161–4.

[35] T. Ochiai, M. Delcroix, T. Nakatani, R. Ikeshita, K. Kinoshita, and S. Araki, "Neural network-based virtual microphone estimator," in *Proc. ICASSP*, 2021, 6114–8.

[36] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, 18(3), 2010, 550–63.

[37] S. U. Pillai, *Array signal processing*, Springer-Verlag, 1989.

[38] S. Rickard, "Sparse sources are separated sources," in *Proc. EUSIPCO*, 2006, 1–5.

[39] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. ICASSP*, Vol. 1, 2002, 529–32.

[40] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 37(7), 1989, 984–95.

[41] R. Sakanashi, S. Miyabe, T. Yamada, and S. Makino, "Comparison of superimposition and sparse models in blind source separation by multichannel Wiener filter," in *Proc. APSIPA ASC*, 2012, 1–6.

[42] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. on Audio, Speech, and Language Processing*, 21(5), 2013, 971–82.

[43] H. Sawada, R. Mukai, and S. Makino, "Direction of arrival estimation for multiple source signals using independent component analysis," in *Proc. International Symposium on Signal Processing and Its Applications*, Vol. 2, 2003, 411–4.

[44] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Antennas and Propagation*, 34(3), 1986, 276–80.

[45] H. Segawa, T. Ochiai, M. Delcroix, T. Nakatani, R. Ikeshita, S. Araki, T. Yamada, and S. Makino, "Neural virtual microphone estimator: Application to multi-talker reverberant mixtures," in *Proc. APSIPA ASC*, 2022, 293–9.

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

[47] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, 5, 1988, 4–24.

[48] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, 14(4), 2006, 1462–9.

[49] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization," *EURASIP Journal on Advances in Signal Processing*, 2007, 2006, Article ID 024717.

[50] K. Yamaoka, N. Ono, and S. Makino, "Time-frequency-bin-wise linear combination of beamformers for distortionless signal enhancement," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, 29, 2021, 3461–75.

[51] K. Yamaoka, N. Ono, S. Makino, and T. Yamada, "Time-frequency-bin-wise switching of minimum variance distortionless response beamformer for underdetermined situations," in *Proc. ICASSP*, 2019, 7908–12.

[52] K. Yamaoka, S. Makino, N. Ono, and T. Yamada, "Performance evaluation of nonlinear speech enhancement based on virtual increase of channels in reverberant environments," in *Proc. EUSIPCO*, 2017, 2324–8.

[53] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, 52(7), 2004, 1830–47.

[54] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computation*, 13(4), 2001, 863–82.