Applying Virtual Microphones to Triangular Microphone Array in in-Car Communication

Hanako Segawa, Riki Takahashi, Ryoga Jinzai, Shoji Makino, Takeshi Yamada
University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8577 Japan

***Abstract*—We have evaluated the application of the wavelength–proportional arrangement of virtual microphones (WPVM) to a triangular microphone array in a real car environment. Beamforming is effective for speech enhancement in in-car communication. However, it is difficult to install a sufficient number of microphones for beamforming in a car. In this paper, we propose a virtual microphone technique that increases the number of microphones virtually to achieve high speech enhancement performance with a small number of microphones in a car. We applied the virtual microphone technique to a triangular microphone array in a car and evaluated the adaptive beamforming performance. As a result, the WPVM was shown to be effective for improving the performance of speech enhancement in a car.**

I. Introduction

Information and communication technology (ICT) is developing rapidly. Recently, several studies of the application of ICT to in-car communication (ICC) have been carried out [1] [2]. Owing to the seat arrangement and background noise, communication within a car is difficult. In particular, backseat passengers may feel uncomfortable listening to the speech from front seat passengers. The situation can be improved by enhancing particular parts of speech by a loudspeaker located near the back seats.

Beamforming is one of the methods used to enhance target speech. In particular, beamforming speech enhancement is effective because the positions of people, or the directions of speech arrival, are easily predictable in a car. However, beamformers do not work properly when there are fewer microphones than sources, i.e., underdetermined conditions. The performance of beamforming degrades in underdetermined situations. In ICC, the performance degrades because of the inability to cope with an increase in the number of speakers and background noise.

A simple solution to this problem is to increase the number of microphones. However, in a car, this means that the microphone array will be larger and more expensive. Considering the increased costs and the balance with the design and other equipment in a car, this solution is not practical.

In this situation, the virtual microphone technique is one of the effective methods. A virtual microphone signal is generated from the observed signals of two real microphones, and a virtual microphone is deployed on the line connecting the two real microphones. By using the virtual signal, we can increase the number of channels of the beamformer. In experiments with two microphones, it has been shown that the performance of speech enhancement with a virtual microphone is better than that without the virtual microphone [3]. The advantage of the virtual microphone technique is that it is easy to install because it has a low cost and requires no space.

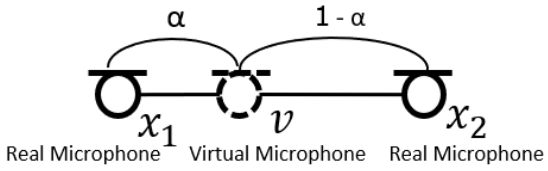

Fig. 1. Arrangement of real and virtual microphones in interpolation technique

In this paper, we consider the virtual microphone technique in ICC. Specifically, we generate virtual signals from the signals obtained by a triangular microphone array in a car and evaluate the adaptive beamforming performance. In an experiment, we use the impulse response measured in a car. We examine the effectiveness of the wavelength–proportional arrangement of virtual microphones (WPVM) for speech enhancement in ICC.

II. Increasing number of channels by virtual microphone technique

A. Interpolation of Virtual Microphone Signals

In this section, we introduce interpolation by the virtual microphone technique [4] [5]. In the virtual microphone technique, a virtual microphone signal $v(\omega, t, \alpha)$ is generated from observed signals of two real microphones x_1, x_2 in the time-frequency domain, where $x_i(\omega, t)$ is the ith microphone signal $(i = 1, 2)$ at the angular frequency ω in the tth time frame. α is the coefficient of the interpolation of the virtual microphone; a virtual microphone is located at the point obtained by internally dividing the line joining the two real microphones in the ratio α to $(1 - \alpha)$. The arrangement of real and virtual microphones is shown in Fig. 1.

In a situation where there are multiple sounds arriving from different directions, the relationship between the microphone position and the waveform is complicated and is difficult to interpolate. Therefore, in this method, by assuming the W-disjoint orthogonality (W-DO) [6] of the observed signals, we simplify the modeling of the relationship. W-DO implies the strong sparsity of the signal in the time-frequency domain, assuming that the component from a sound source dominates

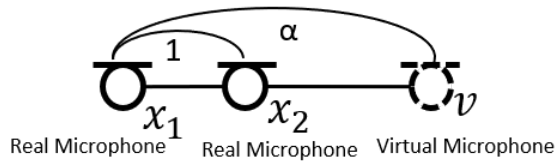

Fig. 2. Arrangement of real and virtual microphones in extrapolation technique

one time–frequency slot of a discrete short–time Fourier transform (STFT).

By assuming this, when multiple sounds arrive, they can be regarded as a single sound in each time–frequency slot. Thus, we can interpolate the virtual microphone signal. In the virtual microphone, the phase and amplitude are interpolated individually. The phase and amplitude of the observed signal $x_i(\omega, t)$ are respectively denoted as

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

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

The phase ϕ_i of the virtual microphone signal is interpolated linearly as follows.

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

The values of the phase are $\phi_i \pm 2n\pi$ for an arbitrary natural number n. Therefore, the phase of the virtual microphone is interpolated with the assumption that

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

The appropriate interpolation of the amplitude of the virtual microphone depends on reverberation and the distance between the source and the microphones in addition to the direction of arrival. Thus, it is difficult to faithfully model the actual amplitude attenuation. However, The signal amplitude must also be interpolated in accordance with an appropriate rule. Consequently, this method uses β-divergence for amplitude interpolation instead of some physical model [4] [5]. β-divergence is used distance measure for nonnegative values such as amplitude. The amplitude of the virtual microphone is interpolated as

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

Using the parameter β, it is possible to nonlinearly interpolate the amplitude of the virtual microphone using the amplitudes of the two real microphones.

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

$$v(\omega, t, \alpha) = A_{\mathrm{v}} \exp(j\phi_{\mathrm{v}}). \quad (6)$$

B. Extrapolation of Virtual Microphone Signals

Next in this section, we introduce extrapolation by the virtual microphone technique [7]. The arrangement of real and virtual microphones is shown in Fig. 2. For the extrapolation of a virtual microphone signal, we have to check the validity of the generation method for a virtual microphone described in the previous section. For phase extrapolation, we can use the same equation as in the previous phase interpolation, but the amplitude interpolation is difficult. Extrapolation based on β-divergence sometimes outputs unrealistic amplitudes such as a negative amplitude or diverges to positive infinity. Such impossible extrapolation must be avoided. Therefore, in this method, since interaural time difference (ITD) is dominant in the frequency range below 1.5 kHz [8] [9], we use the amplitude of the real microphone closest to the virtual microphone as the amplitude of the extrapolated virtual microphone.

$$A_{\mathrm{v}} = \begin{cases} A_1 & (\alpha < 0) \\ A_2 & (\alpha > 1). \end{cases} \quad (7)$$

From the above, the extrapolated virtual microphone signal $v(\omega, t, \alpha)$ is represented in the same way as in the interpolation.

$$v(\omega, t, \alpha) = A_{\mathrm{v}} \exp(j\phi_{\mathrm{v}}). \quad (8)$$

C. Wavelength–Proportional Arrangement of Virtual Microphones (WPVM)

In this section, we introduce the wavelength–proportional arrangement of virtual microphones (WPVM) [3]. In this method, the coefficient of the position of the virtual microphone α is denoted as

$$\alpha(\omega) = \frac{\lambda k}{d} = \frac{2\pi \mathrm{c} k}{\omega d}. \quad (9)$$

where λ is the wavelength, k is the scaling of the interval between the reference microphone and the virtual microphone relative to wavelength, d is the distance between the real microphones, and c is the velocity of sound. This equation implies that the virtual microphone is placed at a position k times the wavelength corresponding to the processing frequency from x_1; therefore, the total length of the microphone array including the virtual microphone is larger at lower frequencies and smaller at higher frequencies. By setting an appropriate parameter k, spatial aliasing does not occur and a sufficient phase difference between microphones can be obtained at all frequencies. In theory, a sufficient phase difference is not obtained at k = 0.25 and spatial aliasing may occur at k = 2. A sufficient phase difference is obtained at k = 0.5 and good performance is expected. By this method, it was shown that speech enhancement performance increased in a two-microphone experiment in an underdetermined situation. According to research on WPVM, an appropriate value of k is to be 0.5 or 1 [3].

D. Triangular Beamformers in a Car

Finally, we introduce triangular beamformers in a car used for speech enhancement in the experiment. In this experiment,

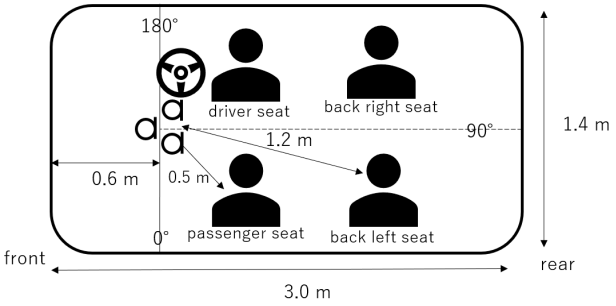

Fig. 3. Sound source and microphone layout in experiment

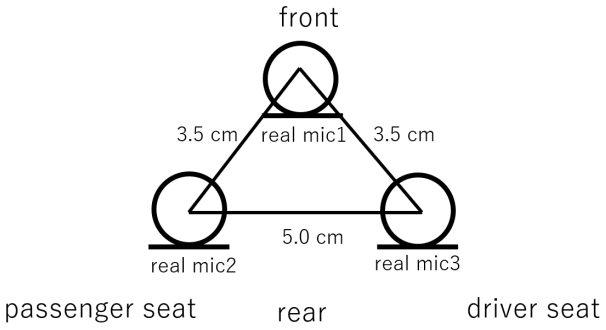

Fig. 4. Arrangement of triangular microphone array

to evaluate the performance of the virtual microphone technique, we apply an increasing number of channels by using the virtual microphone technique to a maximum SNR beamformer [10] [11]. The maximum SNR beamformer requires prior information on the covariance matrices of the target-only period and interference-only period. From the prior information of the target and interference, the maximum SNR beamformer constructs a spatial filter so that the power ratio of the target to the interference becomes maximum. In principle, the virtual microphone can be similarly applied to other microphone array signal processing techniques as well as the maximum SNR beamformer.

As a comparison, we use a minimum variance distortionless response (MVDR) beamformer without virtual microphones for speech enhancement [10] [12]. When there is no correlation between the target and the interference sound, since the variance of the observed signal is the sum of the variances of the target and interference, the MVDR beamformer reduces the effect of the interference by minimizing the variance of the observed signal.

In this paper, we use WPVM and perform extrapolation of virtual microphone signals with a triangular microphone array. In addition, we apply this technique to speech enhancement using a maximum SNR beamformer.

III. Experiments

In the experiments, we examined whether WPVM and the extrapolation of virtual microphones are effective for triangular microphones in ICC, using an observed signal obtained from four speakers and the measured impulse responses in a car. By changing k of WPVM, α used in extrapolation, the speakers

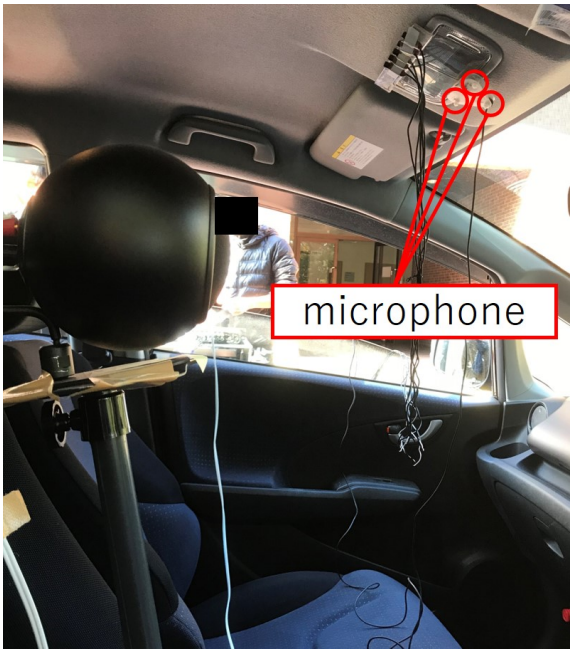

Fig. 5. Actual triangular microphone array

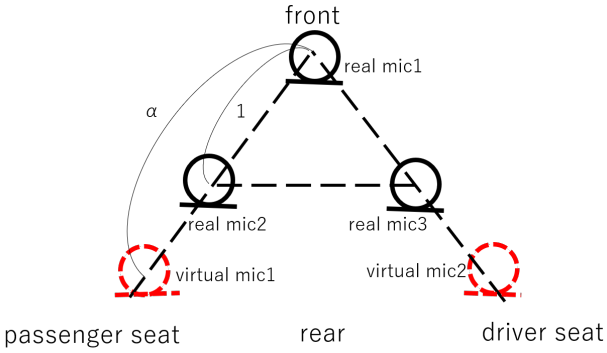

Fig. 6. Arrangement of real and virtual microphones

to be enhanced, and the direction of the virtual microphone, we examined whether the virtual microphone affects the performance. We used signal-to-distortion ratio (SDR), source-to-interference ratio (SIR), and sources-to-artifacts ratio (SAR) to evaluate the enhancement performance [13].

A. Experimental Conditions

In the experiments, we used data for a total of 10 speakers (6 male speakers and 4 female speakers), and 503 phoneme–balanced sentences contained in set B of the ATR digital speech database [14]. We selected four random individuals from this database. We generated 20 patterns of observed signals by convolving measured impulse responses into the speech signals. The number of real microphones was 3 and the number of virtual microphones was 1. There are 2 patterns of virtual microphone positions. The virtual microphones are used individually rathar than both at the sametime.

TABLE I
EXPERIMENTAL CONDITIONS

Sampling rate	8 kHz
Signal-to-noise ratio (SNR)	0 dB
FFT frame length	1024 samples
FFT shift	256 samples
Reverbration time	58 ms

We measured the impulse responses by using a time–stretched pulse (TSP) recorded in a car. In recording the TSP, we set speakers at a driving seat, a passenger seat, and two rear seats in the car and microphones at a map lamp. The layout of the sound sources and triangular microphone array is shown in Fig. 3. The arrangement of the triangular microphone array is shown in Fig. 4. An image of the actual triangular microphone array is shown in Fig. 5. Other experimental conditions are listed in Table I. In this experiment, a virtual microphone was applied to the triangular microphone array. The combinations of two real microphones used to generate the virtual microphones were {(real mic1, real mic2): virtual mic1, (real mic1, real mic3): virtual mic2}, and virtual microphones are generated towards the rear of the car. The arrangement of the virtual microphone is shown in Fig. 6. Virtual mic1 is a virtual microphone toward the passenger seat and the virtual mic2 is a virtual microphone toward the driver's seat. The combinations of two real microphones was set so that virtual microphones are close to the angle of the enhanced speech.

In addition, we set k of WPVM to $\{0.25, 0.5, 1, 2\}$ and set α to 10. We decided the values of k on the basis of the values tested in WPVM research and the value of α on the basis of previously obtained good results [3].

In the experiment, we used the MVDR beamformer [10] [12] and maximum SNR beamformer [10] [11] in only the triangular microphone array, the maximum SNR beamformer was used with a virtual microphone for the driver's or passenger's speech enhancement, and we evaluated SDR, SIR, and SAR. The maximum SNR beamformer and MVDR beamformer were given prior information about the target-only period and interference-only period. The speech given as prior information is another speech of the same speaker as the test.

SDR, SIR, and SAR were averaged over the 20 patterns corresponding to different combinations of the four speakers.

B. Results and Discussion

We evaluated the speech enhancement performance using the average of the results of 20 patterns. Table II shows the driver's speech enhancement and Table III shows the passenger's speech enhancement. The maximum SNR beamformer has better performance than the MVDR. Next, the performance was better with a virtual microphone than without it. By changing from a situation of three microphones and four sound sources (underdetermined situation) to that of four microphones and four sound sources (determined situation) using a virtual microphone, the performance was improved. From these results, the virtual microphone technique is effective in a triangular microphone array in an underdetermined situation in a car.

TABLE II
RESULTS OF DRIVER'S SPEECH ENHANCEMENT

Virtual mic	k;[1] α[2]	SDR	SIR	SAR
no virtual mic	MVDR	8.31	18.00	8.90
	max SNR	9.02	14.76	10.55
virtual mic1	k = 0.25	10.33	16.99	11.50
	k = 0.5	**10.60**	**17.64**	**11.66**
	k = 1	10.41	17.45	11.47
	k = 2	9.41	15.64	10.74
	α = 10	10.19	16.76	11.39
virtual mic2	k = 0.25	10.38	16.93	11.59
	k = 0.5	**10.47**	**17.25**	11.61
	k = 1	10.07	16.71	11.27
	k = 2	9.46	15.64	10.82
	α = 10	10.39	16.81	**11.64**

[1] Wavelength coefficient in WPVM.
[2] Coefficient of position of virtual microphone in extrapolation.

TABLE III
RESULTS OF PASSENGER'S SPEECH ENHANCEMENT

Virtual mic	k;[1] α[2]	SDR	SIR	SAR
no virtual mic	MVDR	4.86	15.22	5,45
	max SNR	8.96	15.18	10.33
virtual mic1	k = 0.25	10.29	17.10	11.42
	k = 0.5	**10.46**	**17.57**	**11.51**
	k = 1	10.15	17.32	11.20
	k = 2	9.36	16.00	10.58
	α = 10	10.20	17.01	11.34
virtual mic2	k = 0.25	9.66	16.74	10.82
	k = 0.5	**9.87**	**16.85**	**10.98**
	k = 1	9.84	16.75	10.97
	k = 2	9.35	15.93	10.97
	α = 10	9.71	16.54	10.86

Next, we found that the use of WPVM is the best when an appropriate k (k = 0.5) is chosen where we apply a virtual microphone. In contrast, when an inappropriate k, especially k = 2, was used, the results were worse than those for extrapolation with α = 10. By comparing WPVM with an appropriate k and the maximum SNR beamformer with only the triangular microphone array, the SDR was improved by up to 1.6 dB for the former. These results show that the performance is slightly better using WPVM with an appropriate k than using the extrapolation of a virtual microphone with α = 10.

The directivity pattern did not change significantrly when the driver's speech was enhanced, and SDR was improved by up to about 0.6 dB when the passenger's speech was enhanced.

Looking at the individual results, we found that the results for passenger's speech enhancement were often better with

virtual mic1, whereas the results for the driver's speech enhancement were good in some cases and the results were bad in others with virtual mic2. This result shows that the performance is not uniformly affected by the placement direction, but it is more likely that the virtual microphones are the best placed in the direction of the target.

These results are similar to those of the two-microphone experiments [3], except for the direction of placement, which indicates that the virtual microphone is effective on a surface with a triangular microphone array in ICC.

IV. Conclusion

In this paper, we applied the wavelength–proportional arrangement of virtual microphones (WPVM) to speech enhancement with a triangular microphone array in in-car communication (ICC) in an underdetermined situation. By the virtual microphone technique, it was found possible to mitigate the degradation of performance due to underdetermined conditions in three-microphone experiments. In these experiments, we generated observed signals by convolving measured impulse responses into speech in a car. We compared the results of speech enhancement by beamforming with and without a virtual microphone.

In the experiment, the virtual microphone technique was shown to be effective for improving speech enhancement performance by using a triangular microphone array in ICC and an underdetermined situation. The WPVM with an appropriate k showed especially the best enhancement performance.

V. Acknowledgment

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 19H04131, and the Strategic Core Technology Advancement Program (Supporting Industry Program).

References

[1] Rabea Landgraf, Johannes Köhler-Kaeß, Christian Lüke, Oliver Niebuhr, Gerhard Schmidt, "Can you hear me now? Reducing the Lombard effect in a driving car using an in-car communication system," in Proc. Speech Prosody, pp. 479–483, 2016.

[2] Anne Theiss, Gerhard Schmidt, Jochen Withopf, Christian Lueke, "Instrumental evaluation of in-car communication systems," in Proc. Speech Communication; 11. ITG Symposium, pp. 1–4. VDE, 2014.

[3] Ryoga Jinzai, Kouei Yamaoka, Mitsuo Matsumoto, Shoji Makino, Takeshi Yamada, "Wavelength proportional arrangement of virtual microphones based on interpolation/extrapolation for underdetermined speech enhancement," in Proc. European Signal Processing Conference (EUSIPCO), pp. 1–5, 2019.

[4] Hiroki Katahira, Nobutaka Ono, Shigeki Miyabe, Takeshi Yamada, Shoji Makino, "Nonlinear speech enhancement by virtual increase of channels and maximum SNR beamformer," EURASIP Journal on Advances in Signal Processing, vol. 2016, no. 1, pp. 1–8, 2016.

[5] Kouei Yamaoka, Shoji Makino, Nobutaka Ono, Takeshi Yamada, "Performance evaluation of nonlinear speech enhancement based on virtual increase of channels in reverberant environments," in Proc. European Signal Processing Conference (EUSIPCO), pp. 2324–2328, 2017.

[6] Ozgur Yilmaz, Scott Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Transactions on Signal Processing, vol. 52, no. 7, pp. 1830–1847, 2004.

[7] Ryoga Jinzai, Kouei Yamaoka, Mitsuo Matsumoto, Takeshi Yamada, Shoji Makino, "Microphone position realignment by extrapolation of virtual microphone," in Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 367–372, 2018.

[8] Brian Moore, "An introduction to the psychology of hearing," Academic Press, 1997.

[9] Jens Blauert, "Spatial hearing: the psychophysics of human sound localization," MIT Press, 1997.

[10] H. L. Van Trees, "Optimum array processing," John Wiley & Sons, 2002.

[11] Shoko Araki, Hiroshi Sawada, Shoji Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 41-44, 2007.

[12] Otis Lamont Frost, "An algorithm for linearly constrained adaptive array processing," Proceedings of the IEEE, vol. 60, no. 8, pp. 926–935, 1972.

[13] Emmanuel Vincent, Hiroshi Sawada, Pau Bofill, Shoji Makino, Justinian P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in Proc. International Conference on Independent Component Analysis and Signal Separation (ICA), pp. 552–559, 2007.

[14] Akira Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, Kiyohiro Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Communication, vol. 9, no. 4, pp. 357–363, 1990.