CNN-based virtual microphone signal estimation for MPDR beamforming in underdetermined situations

Kouei Yamaoka*, Li Li*, Nobutaka Ono†, Shoji Makino*, and Takeshi Yamada*
*University of Tsukuba, 1-1-1 Tennodai, Tsukuba-shi, Ibaraki 305-8577, Japan
{yamaoka@mmlab.cs, lili@mmlab.cs, maki@tara, takeshi@cs}.tsukuba.ac.jp
†Tokyo Metropolitan University, 6-6 Asahigaoka, Hino-shi, Tokyo 191-0065, Japan, onono@tmu.ac.jp

***Abstract*—In this paper, we propose a novel approach to virtually increasing the number of microphone elements between two real microphones to improve speech enhancement performance in underdetermined situations. The virtual microphone technique, with which virtual signals in the audio signal domain are estimated by linearly interpolating the phase and nonlinearly interpolating the amplitude independently on the basis of β-divergence, has been recently proposed and experimentally shown to be effective in improving speech enhancement performance. Furthermore, it has been reported that the performance tends to improve as the nonlinearity is improved. However, one drawback of this method is that the interpolation is employed in each time-frequency bin independently, in which the spectral and temporal structures of speech signals are ignored. To address this problem and improve the nonlinearity, motivated by the high capability of neural networks to model nonlinear functions and speech spectrograms, in this paper, we propose an alternative method of amplitude interpolation. In this method, we employ a convolutional neural network as an amplitude estimator that minimizes the mean squared error between the outputs of a minimum power distortionless response (MPDR) beamformer and the target speech signals. The experimental results revealed that the proposed method showed high potential for improving speech enhancement performance, which was not only superior to that of the conventional virtual microphone technique but also the performance in the corresponding determined situation.**

I. Introduction

The technique of reducing undesirable noise while enhancing the target speech in recorded mixture signals, which is referred to as speech enhancement, plays an important role in many audio signal processing applications such as automatic speech recognition [1]. Beamforming and blind source separation (BSS) [2] are commonly used methods for speech enhancement and can yield good performance as long as a sufficient number of microphones are available, namely, the number of microphones M equals or exceeds the number of sound sources N $(M \geq N)$ to suppress $N-1$ interferers by null steering. Otherwise, the performance of speech enhancement tends to decrease considerably. However, commonly used small recording devices such as voice recorders often have only two microphones, which is insufficient to meet the determined condition. To achieve satisfactory speech enhancement performance with such devices having a small microphone array, many methods have been proposed to enhance speech in underdetermined situations $(M < N)$ such as time-frequency masking [3]–[5], multichannel Wiener filtering [6], [7], and nonnegative matrix factorization (NMF) [8]–[10]. Although these methods are noteworthy in that they can significantly improve speech intelligibility in underdetermined situations, there is a tradeoff between low signal distortion and high noise reduction performance.

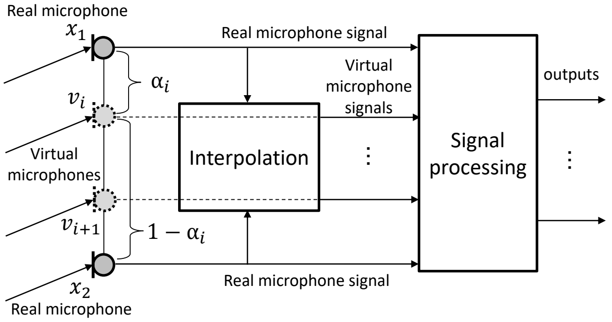

Fig. 1. Microphone array signal processing with virtual microphone technique

On the other hand, the virtual microphone technique [11] allows the well-studied methods for determined situations $(M = N)$ to be applied to signals recorded in underdetermined situations by virtually increasing the number of channels. As shown in Fig. 1, with two real microphone signals x_1 and x_2, the virtual microphone technique is used to estimate the observed signal v_i at a position where there is no real microphone placed by interpolating the phase and amplitude independently. Note that since the virtual microphone signals are generated in the audio signal domain, this technique can be applied to not only speech enhancement but also other types of signal processing [12], which is different from techniques in which signals are generated in the power domain [13]–[15] or a higher-order statistical domain [16].

In the virtual microphone technique, on the basis of the W-disjoint orthogonality (W-DO) [3], [17] assumption, phases of virtual signals can be obtained using linear interpolation by approximately modeling propagating waves as plane waves. For amplitude estimation, since modeling the amplitudes of propagating waves is difficult owing to the complicated acoustic environment, *complex logarithmic interpolation* and a generalized version, where the interpolation rules are derived as closed-form solutions of an optimization problem formulated using β-divergence, have been proposed and experimentally shown to be effective in improving speech enhancement performance using a maximum signal-to-noise ratio (MaxSNR) beamformer. Furthermore, the results reported in [11] indicated that speech enhancement performance tends to increase when the improved nonlinearity is applied to amplitude interpolation. However, one drawback of this method is that the interpolation is applied in each time-frequency bin independently, in which the spectral and temporal structures of speech signals are ignored.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

To overcome this drawback and improve the nonlinearity of amplitude generation, motivated by the high capability of neural networks to model nonlinear functions and speech spectrograms, in this paper, we propose the use of convolutional neural networks (CNNs) to learn the rules for estimating amplitudes of virtual signals. Specifically, the proposed method trains a CNN to estimate a virtual signal that can minimize the mean squared error of the outputs of a minimum power distortionless response (MPDR) beamformer and the target speech signals. Note that this is another merit of replacing the ad hoc interpolation rules with rules learned in a data-driven manner, namely, it allows a complicated model to be learned with a loss function that is easy to formulate or is task-dependent.

II. Rule-based virtual microphone technique

We model the microphone signals in the short-time Fourier transform (STFT) domain. Here, let $x_m(\omega, t)$ be the mth real microphone signal ($m = 1, 2$) at angular frequency ω in the tth time frame. The amplitudes of these signals are denoted as $A_m(\omega, t) = |x_m(\omega, t)|$ and the phases are denoted as $\phi_m(\omega, t) = \angle x_m(\omega, t)$. A virtual microphone signal $v(\omega, t, \alpha, \beta)$ is defined as the observation estimated at the point obtained by internally dividing the line joining two real microphones in the ratio $\alpha : (1 - \alpha)$ (see Fig. 1). Here, β is the hyperparameter of β-divergence, which controls the nonlinearity of amplitude interpolation. Hereafter, when there is no need to distinguish ω, t, α, or β, the signal is simply denoted as v.

To interpolate a virtual microphone signal, we need to define the distance between the real and virtual microphones on the basis of an appropriate model that is extremely complex. To simplify the model, we here consider the models for phase and amplitude interpolation separately. Additionally, interpolating the phase and amplitude separately introduces nonlinearity into virtual signal generation, which is necessary to increase the number of channels.

A. Phase interpolation based on plane wave model

We assume W-DO [3], [17] for mixed signals, that is, each time-frequency bin is dominated by at most one sound source. Then, the observed signal in each time-frequency bin can be regarded as a single wave. On the basis of this assumption, the physical model of propagating waves can then be approximated as that of a plane wave. The phase ϕ_{v} of a virtual microphone signal v can then be interpolated linearly on the basis of the model as

$$\phi_{\mathrm{v}} = (1 - \alpha)\,\phi_1 + \alpha\phi_2. \tag{1}$$

Since the observed phase has an aliasing ambiguity given by $\phi_i \pm 2n_i\pi$ with integer n_i, this interpolation requires no spatial aliasing, that is,

$$|\phi_1 - \phi_2| \leq \pi. \tag{2}$$

B. Amplitude interpolation based on β-divergence

Since there are many acoustic conditions such as the distance between the sound sources and microphones and the directions of arrival (DOAs), it is difficult to faithfully model the amplitude of a propagating wave. Thus, instead of some physical assumptions, we utilize β-divergence as an adjustable measure of distance to quantify the distance between the real and virtual microphones.

The β-divergence between the amplitude of the virtual microphone A_{v} and that of the ith real microphone A_i is defined as

$$\boldsymbol{D}_\beta(A_{\mathrm{v}}, A_i) = \begin{cases} A_{\mathrm{v}}(\log A_{\mathrm{v}} - \log A_i) + (A_i - A_{\mathrm{v}}) & (\beta = 1), \\ \dfrac{A_{\mathrm{v}}}{A_i} - \log \dfrac{A_{\mathrm{v}}}{A_i} - 1 & (\beta = 0), \\ \dfrac{A_{\mathrm{v}}^\beta}{\beta(\beta - 1)} + \dfrac{A_i^\beta}{\beta} - \dfrac{A_{\mathrm{v}} A_i^{\beta - 1}}{\beta - 1} & (\text{otherwise}), \end{cases} \tag{3}$$

where $\boldsymbol{D}_\beta(A_{\mathrm{v}}, A_i)$ is continuous at $\beta = 0$ and $\beta = 1$. Then, we derive the interpolation rule of the amplitude A_{v} that minimizes $\sigma_{\boldsymbol{D}_\beta}$, the sum of $\boldsymbol{D}_\beta(A_{\mathrm{v}}, A_i)$ weighted by the hyperparameter of the virtual microphone interpolation α, which indicates the position of the virtual microphone,

$$\sigma_{\boldsymbol{D}_\beta} = (1 - \alpha)\boldsymbol{D}_\beta(A_{\mathrm{v}}, A_1) + \alpha \boldsymbol{D}_\beta(A_{\mathrm{v}}, A_2), \tag{4}$$

$$A_{\mathrm{v}\beta} = \mathrm{argmin}_{A_{\mathrm{v}}} \sigma_{\boldsymbol{D}_\beta}. \tag{5}$$

By differentiating $\sigma_{\boldsymbol{D}_\beta}$ with respect to A_{v} and setting it to 0, the interpolated amplitude is obtained as

$$A_{\mathrm{v}\beta} = \begin{cases} \exp\left((1 - \alpha)\log A_1 + \alpha \log A_2\right) & (\beta = 1), \\ \left((1 - \alpha)\, A_1^{\beta - 1} + \alpha A_2^{\beta - 1}\right)^{\frac{1}{\beta - 1}} & (\text{otherwise}). \end{cases} \tag{6}$$

Note that $A_{\mathrm{v}\beta}$ is continuous at $\beta = 1$ and this interpolation is equivalent to *complex logarithmic interpolation* [18] with $\beta = 1$.

From the above, the virtual microphone signal v is represented as

$$v = A_{\mathrm{v}\beta} \exp\left(j\phi_{\mathrm{v}}\right). \tag{7}$$

Note that the phase can be interpolated with arbitrary real numbers α, whereas the amplitude interpolation is defined only in the domain of $0 \leq \alpha \leq 1$ when $\beta \neq 1$. The extrapolation of a virtual microphone in the domain $\alpha < 0, 1 < \alpha$ was considered in [12].

III. Proposed method: CNN-based amplitude estimation for MPDR beamformer

The effectiveness of the virtual microphone technique in improving speech enhancement performance using a MaxSNR beamformer has been confirmed in [11]. According to the experimental results in [11], the performance improvement tends to increase when a larger β is used for amplitude interpolation, which indicated the importance of the nonlinearity in generating a virtual signal. However, with the virtual microphone technique, the amplitude is estimated in each time-frequency bin independently, in which the structure of speech signals is ignored. To further improve the nonlinearity for amplitude interpolation and estimate the amplitude by taking account of all the observed time-frequency bins, in this paper, we introduce the use of a CNN as an alternative means of amplitude estimation (see Fig. 2). Furthermore, learning the rules in a data-driven manner facilitates the modeling process and allows the use of task-dependent loss functions.

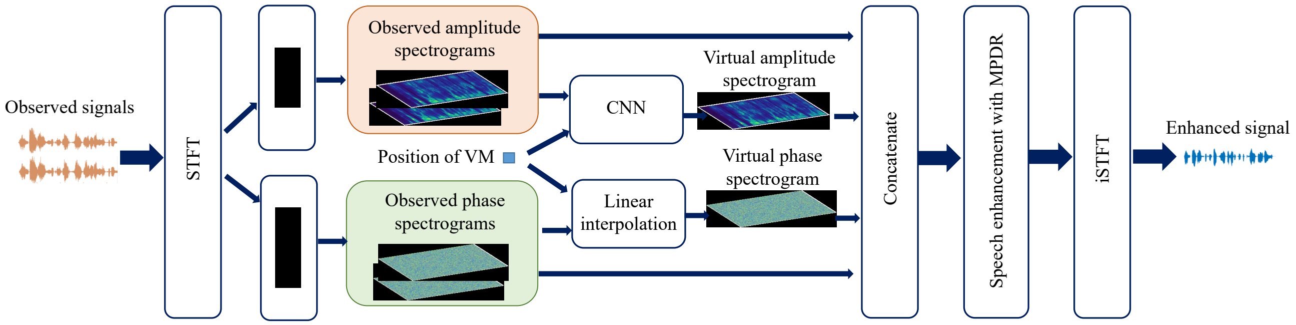
Observed signals
STFT
Observed amplitude spectrograms
Position of VM
Observed phase spectrograms
CNN
Linear interpolation
Virtual amplitude spectrogram
Virtual phase spectrogram
Concatenate
Speech enhancement with MPDR
iSTFT
Enhanced signal

Fig. 2. Flowchart of proposed method

A. *Loss function with MPDR beamformer*

In this paper, we introduce a task-dependent loss function that forces the generated amplitude to become optimal for constructing an MPDR beamformer [19] and minimizes the mean squared error between the output of the beamformer and the target signal $s(\omega, t)$. For the observed signal consisting of two real and I virtual microphone signals $\boldsymbol{x}(\omega, t) = [x_1(\omega, t), v_1(\omega, t), \cdots, v_I(\omega, t), x_2(\omega, t)]^{\mathrm{T}}$, an MPDR beamformer enhances the source of interest by

$$y(\omega, t) = \boldsymbol{w}^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega, t), \tag{8}$$

$$\boldsymbol{w}(\omega) = [w_1(\omega) \cdots w_M(\omega)]^{\mathrm{T}}, \tag{9}$$

where $y(\omega, t)$ denotes the output signal of the beamformer, $\boldsymbol{w}(\omega)$ the spatial filter vector, $(\cdot)^{\mathrm{T}}$ the transpose, and $(\cdot)^{\mathrm{H}}$ the Hermitian transpose. The spatial filter $\boldsymbol{w}(\omega)$ is given as

$$\boldsymbol{w}(\omega) = \frac{\boldsymbol{\Phi}(\omega)^{-1}\boldsymbol{a}(\omega)}{\boldsymbol{a}^{\mathrm{H}}(\omega)\boldsymbol{\Phi}(\omega)^{-1}\boldsymbol{a}(\omega)}, \tag{10}$$

$$\boldsymbol{\Phi}(\omega) = \mathbb{E}[\boldsymbol{x}(\omega, t)\boldsymbol{x}(\omega, t)^{\mathrm{H}}], \tag{11}$$

which is a well-known closed-form solution of the optimization problem

$$\mathcal{J} = \sum_{\omega} \left\{ \mathbb{E}\left[|\boldsymbol{w}^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega, t)|^2\right] + 2\mathrm{Re}[\lambda^*(\boldsymbol{w}^{\mathrm{H}}(\omega)\boldsymbol{a}(\omega) - 1)] \right\}, \tag{12}$$

which is derived using the method of Lagrange multipliers. Here, $\mathbb{E}[\cdot]$ is the expectation operator, $\mathrm{Re}[\cdot]$ takes the real part of the input argument, λ^* is the complex-valued Lagrange multiplier, and $\boldsymbol{a}(\omega)$ is the relative transfer function (RTF) of the target, which is defined as the ratio of the acoustic transfer functions $\boldsymbol{h}(\omega) = [h_1(\omega) \cdots h_M(\omega)]^{\mathrm{T}}$ from the target source to the microphone array, i.e., $\boldsymbol{a}(\omega) = \left[1 \;\; \frac{h_2(\omega)}{h_1(\omega)} \;\; \cdots \;\; \frac{h_M(\omega)}{h_1(\omega)}\right]^{\mathrm{T}}$. The training loss function of the CNN can then be written as

$$\mathcal{J}_{\mathrm{c}} = \sum_{\omega} \mathbb{E}\left[|\boldsymbol{w}^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega, t) - s(\omega, t)|^2\right]. \tag{13}$$

B. *Network architecture design*

The network architecture employed in the proposed method is designed by considering the following two aspects: 1) amplitude spectrograms of speech signals show region dependence, i.e., they have different frequency structures in voiced and unvoiced segments; 2) the entire amplitude spectrogram should be estimated by taking all the observed time-frequency bins as a cue, which means a large receptive field is required.

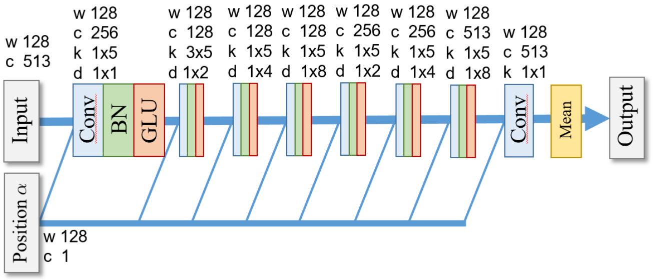

Fig. 3. Network architectures used for amplitude estimation. The inputs and outputs are 1D data, where the frequency dimension of spectrograms is regarded as the channel dimension. “w”, “c”, “k” and “d” denote the width, channel number, kernel size, and dilation factor, respectively. “Conv”, “BN” and “GLU” denote 1D convolution, batch normalization, and gated linear unit, respectively.

We use gated CNNs [20], which were originally introduced to model word sequences for language modeling and were shown to outperform long short-term memory (LSTM)-based language models trained in a similar setting. Similar to LSTMs, the gating mechanism of gated CNNs allows a network to learn what information should be propagated through the hierarchy of layers. There are some attempts have been made to adopt CNNs into beamforming techniques. In [21], time-frequency masks for the target and noise signals are estimated with a CNN so that the power spectral density matrices for conducting beamforming can be obtained using the masked signals.

In the gated CNN, by using $\mathbf{H}_{l-1}$ to denote the output of the $(l-1)$th layer, the output of the lth layer $\mathbf{H}_l$ of a gated CNN can be written as

$$\mathbf{H}_l = (\mathbf{H}_{l-1} * \mathbf{W}_l^{\mathrm{f}} + b_l^{\mathrm{f}}) \otimes \sigma(\mathbf{H}_{l-1} * \mathbf{W}_l^{\mathrm{g}} + b_l^{\mathrm{g}}), \tag{14}$$

where $\mathbf{W}_l^{\mathrm{f}}$ and $\mathbf{W}_l^{\mathrm{g}}$ are the weight parameters and b_l^{f} and b_l^{g} are the bias parameters of the lth layer, respectively, $\otimes$ denotes element-wise multiplication and σ is the sigmoid function. The main difference between a gated CNN and a regular CNN layer is that a gated linear unit (GLU), namely, the second term of (14), is used as a nonlinear activation function instead of tanh activation or regular rectified linear units

(ReLUs) [22]. Similar to LSTMs, GLUs are data-driven gates, which control the information passed on in the hierarchy. This particular mechanism allows us to capture long-range context dependences efficiently without encountering the vanishing gradient problem. We also employ a one-dimensional (1D) CNN and dilated convolution to capture long-term contextual dependences. With 1D convolution models, the frequency dimension is regarded as the channel dimension and an input spectrogram is convolved with a $(1, k_T)$ filter so that the features extracted in the low-level layers can take into account all the frequency bins, where k_T is the filter width in the time dimension. Dilated convolution [23] is another effective approach whereby CNNs can capture wider receptive fields with fewer layers by convolving a larger filter derived from the original filter with dilating zeros, namely, the original filter is applied by skipping certain elements in the input. Note that the network is designed to be fully convolutional so that inputs having arbitrary lengths can be handled. The details of the network used in experiments are shown in Fig. 3.

IV. Evaluation experiments

To evaluate the effect of the proposed method, we conducted experiments designed to compare the speech enhancement performance with the MPDR beamformer among the cases of using two real microphones, two real microphones and one virtual microphone signal estimated by the proposed method or the conventional method, and three real microphones.

A. Experimental conditions

We prepared two datasets for the closed test (dataset 1) and open test (dataset 2). Dataset 1 consisted of three speakers and the audio files for each speaker were one minute long, which were used to generate mixture signals. Dataset 2 comprised 10 speakers excerpted from the Wall Street Journal (WSJ0) corpus and the audio files for each speaker were about 18 minutes long, where randomly selected half of them (9 minutes) was used as the training set and the others (9 minutes) served as the test set. The mixture signals, consisting of a target speech and two interferers, were generated by adding each audio signal of the target speaker to all the paired combinations from the other nine speakers. Therefore, the training dataset included four hours of data in total. For both datasets, the observed signals were convolutive mixtures of impulse responses simulated by a room impulse response generator [24]. The experimental conditions are listed in Table I. The DOA of the target speech was set to 90°, and those of the interferers were set to 50° and 150° with a reverberation time of 120 ms.

We used the exact RTF for all the methods. The RTF at the position of a virtual microphone is estimated using the conventional virtual microphone technique. For the proposed method, the RTF can also be estimated using a neural network, which is one direction of future work. We used signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifacts ratio (SAR) as the objective criteria for quantifying the speech enhancement performance. A concise representation of the results was obtained by averaging these criteria over the test dataset. Here, the reference signal was the source image, i.e., the noise-free reverberant speech signal.

TABLE I
Experimental conditions

Number of real microphones M	2 or 3
Number of sound sources N	3
Distance between microphones	4 cm ($M = 2$), 2 cm ($M = 3$)
Reverberation time	120 ms
Sampling rate	8 kHz
Input SNR	0 dB
Window length / shift	1024 / 512 samples

TABLE II
Experimental results

Closed Test			
conditions	SDR	SIR	SAR
2 real mic	1.7960	2.3456	13.4448
2 real mic + 1 vir mic (conv.)	5.9321	8.4432	10.2255
2 real mic + 1 vir mic (prop.)	11.6489	19.6029	12.4629
3 real mic	8.1513	12.9107	10.1648
Open Test			
conditions	SDR	SIR	SAR
2 real mic	2.1939	2.8624	12.6560
2 real mic + 1 vir mic (conv.)	5.9698	8.5820	10.0593
2 real mic + 1 vir mic (prop.)	5.3579	15.5531	5.9356
3 real mic	16.1139	19.5216	18.8855

B. Results and discussion

The SDR, SIR, and SAR are shown in Table II for each dataset. Since the MPDR beamformer with two real microphones can suppress only one interferer in each frequency bin, it is natural that the beamformer failed to enhance speech in underdetermined situations. By using an additional virtual microphone signal estimated by the conventional virtual microphone technique, an improvement of about 4 dB was obtained in terms of SDR, which showed the effectiveness of virtual signals in speech enhancement.

In the closed test, the proposed method achieved a major improvement in all the criteria and even outperformed the method that used three observed signals (i.e., determined situations). The improvements were about 10 dB and 17 dB in terms of SDR and SIR, respectively. These results are interesting in that they reveal not only the high potential of the proposed method to improve the performance of the conventional method, but also the fact that the optimal amplitude for constructing an MPDR beamformer is not equivalent to the observed one. Fig. 4 shows an example of an amplitude spectrogram estimated by the proposed method and Figs. 5 and 6 show the enhanced speech signals. However, the performance of the proposed method decreased significantly in the open test, in which SAR was lower than that in the underdetermined situations (two real microphones). Since the estimated virtual signals have nonlinear noise, the output signal computed by linear transformation of a concatenated input signal also contains some artifacts. One possible reason is that the model trained with dataset 2 was not generalized well enough to generate amplitudes for all the unseen data, which generated virtual signals including high-level artifacts. To improve the generalization of the model, a larger dataset and an appropriate network architecture are needed, which is another direction of future work.

From the above, we concluded that the proposed method has the potential to improve the conventional virtual microphone technique and even outperform the MPDR beamformer for determined situations in terms of speech enhancement perfor-

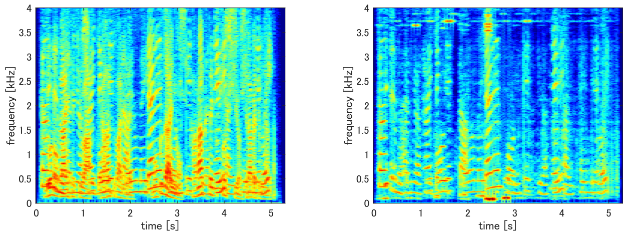

Fig. 4. Amplitude spectrogram of one of the observed two-channel mixture signals (left) and one estimated using proposed method in closed test (right).

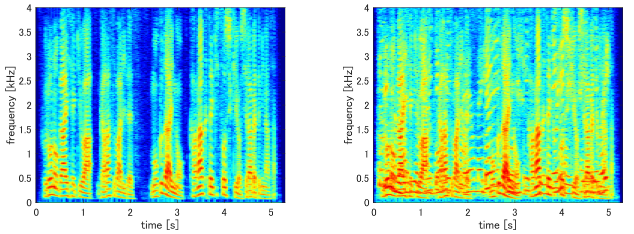

Fig. 5. Target speech signal (left) and signal enhanced using two real microphones (right).

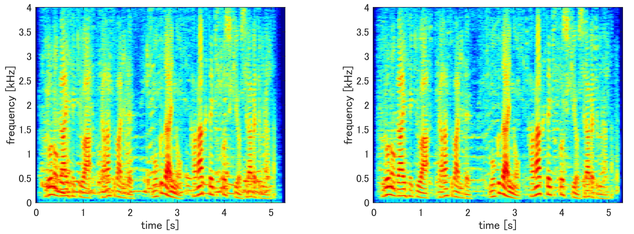

Fig. 6. Speech signals enhanced using two real microphones and one virtual microphone signal estimated by conventional (left) and proposed (right) methods.

mance when the model is trained to be well generalized.

V. Conclusions

In this paper, we proposed an alternative method in which a CNN is used as an estimator of the amplitude of a virtual signal, which aims to improve speech enhancement performance in underdetermined situations. The CNN is trained with a task-dependent loss function that minimizes the mean squared error between the output of an MPDR beamformer and the target speech signal so that the estimated amplitudes could be optimal for constructing an MPDR beamformer. An experiment conducted that involved a closed test showed that the proposed method can potentially improve the conventional method of speech enhancement in underdetermined situations and even outperform other methods for determined situations by using the generated amplitudes specialized for the MPDR beamformer instead of the observed one.

Acknowledgments

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Numbers JP16H01735, JPH04131, and JP19J20420.

References

[1] Takuya Higuchi, Nobutaka Ito, Shoko Araki, Takuya Yoshioka, Marc Delcroix, and Tomohiro Nakatani, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 4, pp. 780–793, 2017.

[2] Shoji Makino, Te-Won Lee, and Hiroshi Sawada, *Blind Speech Separation*, Springer, 2007.

[3] Özgür Yılmaz and Scott Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[4] Scott Rickard, "The DUET blind source separation algorithm," in *Blind Speech Separation*, Shoji Makino, Te-Won Lee, and Hiroshi Sawada, Eds., pp. 217–241. Springer, 2007.

[5] Hiroshi Sawada, Shoko Araki, and Shoji Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 516–527, 2010.

[6] Ngoc Q. K. Duong, Emmanuel Vincent, and Rémi Gribonval, "Underdetermined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[7] Robbe Van Rompaey and Marc Moonen, "GEVD based speech and noise correlation matrix estimation for multichannel wiener filter based noise reduction," *in Proc. EUSIPCO*, pp. 2562–2566, 2018.

[8] Alexey Ozerov and Cedric Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. ASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[9] Masahiro Nakano, Hirokazu Kameoka, Jonathan Le Roux, Yu Kitano, Nobutaka Ono, and Shigeki Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," *in Proc. IEEE MLSP*, pp. 283–288, 2010.

[10] Cédric Févotte and Jérôme Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

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

[12] Ryoga Jinzai, Kouei Yamaoka, Mitsuo Matsumoto, Takeshi Yamada, and Shoji Makino, "Microphone position realignment by extrapolation of virtual microphone," *in Proc. APSIPA*, pp. 367–372, 2018.

[13] Hiroshi Saruwatari, Shoji Kajita, Kazuya Takeda, and Fumitada Itakura, "Speech enhancement using nonlinear microphone array based on complementary beamforming," *IEICE Trans. Fundamentals*, vol. E82-A(8), pp. 1501–1510, 1999.

[14] Shigeki Miyabe, Biing Hwang Juang, Hiroshi Saruwatari, and Kiyohiro Shikano, "Analytical solution of nonlinear microphone array based on complementary beamforming," *in Proc. IWAENC*, pp. 1–4, 2008.

[15] Yusuke Hioka and Terence Betlehem, "Under-determined source separation based on power spectral density estimated using cylindrical mode beamforming," *in Proc. WASPAA*, pp. 1–4, 2013.

[16] Pascal Chevalier, Anne Ferréol, and Laurent Albera, "High-resolution direction finding from higher order statistics: The 2q-MUSIC algorithm," *IEEE Trans. Signal Process.*, vol. 53, no. 4, pp. 2986–2997, 2006.

[17] Alexander Jourjine, Scott Rickard, and Özgür Yılmaz, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures," *in Proc. ICASSP*, pp. 2985–2988, 2000.

[18] Hiroki Katahira, Nobutaka Ono, Shigeki Miyabe, Takeshi Yamada, and Shoji Makino, "Virtually increasing microphone array elements by interpolation in complex-logarithmic domain," *in Proc. EUSIPCO*, pp. 1–5, 2013.

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

[20] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier, "Language modeling with gated convolutional networks," *in Proc. ICML*, pp. 933–941, 2017.

[21] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbash, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *ELSEVIER*, vol. 46, pp. 374–385, 2017.

[22] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," *in Proc. ICML*, pp. 807–814, 2010.

[23] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[24] E. A. P. Habets, "Room impulse response (RIR) generator," Available at: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator, 2008.