NCSP'21
RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing 2021
March 1-2, 2021

VMInNet: Interpolation of Virtual Microphones in Optimal Latent Space Explored by Autoencoder

Riki Takahashi, Li Li, Shoji Makino, Takeshi Yamada

University of Tsukuba, Japan
E-mail: {r.takahashi@mmlab.cs, lili@mmlab.cs, maki@tara, takeshi@cs}.tsukuba.ac.jp

Abstract

In this paper, we propose a new method for interpolation of virtual signals between two real microphones to improve speech enhancement performances in underdetermined situations. The virtual microphone technique is a recently proposed technique that can virtually increase the channel of observed signals by linearly interpolating the phase and nonlinearly interpolating the amplitude based on β-divergence in the short-time Fourier transform (STFT) domain. This technique has been shown to be effective in improving the speech enhancement performance of beamforming in underdetermined situations. It is reasonable to linearly interpolate the phase based on the sound propagation model and nonlinearly interpolate the amplitude to increase the information content of the observed signals. However, there is no theoretical proof that β-divergence is the optimal criterion for amplitude interpolation due to the complexity of the physical model of amplitude. In this paper, we propose using an autoencoder to search for the optimal interpolation domain in a data-driven manner. We perform amplitude interpolation in the latent space, a low dimensional representation space of observed mixture signals that is trained so that the interpolated virtual signals are optimal for conducting beamforming with high performance. The experimental results revealed that the proposed method achieved higher speech enhancement performance than the conventional methods.

1. Introduction

In recent years, with the development of automatic speech recognition (ASR) and robot hearing, the importance of speech enhancement has considerably increased. Owing to the wide usage of stereo microphone built-in small devices such as smartphones and voice recorders, speech enhancement that serves dual-channel signals is a particular requisite. Beamforming and blind source separation (BSS) are the two main methods to deal with this problem. Beamforming and BSS methods using spatial filtering [1, 2, 3] are noteworthy in the low distortion of the enhanced target speech. However, their performance degrades when there are fewer microphones than sources, i.e., underdetermined conditions. To achieve satisfactory speech enhancement performance with such devices having small microphone array, many methods have been proposed to enhance speech in underdetermined situations such as time-frequency masking [4, 5, 6], multichannel Wiener filtering [7, 8], and nonnegative matrix factorization (NMF) [9]. Although these methods are noteworthy that they can significantly improve speech intelligibility in underdetermined situations, they face a tradeoff between low signal distortion and high noise reduction performance.

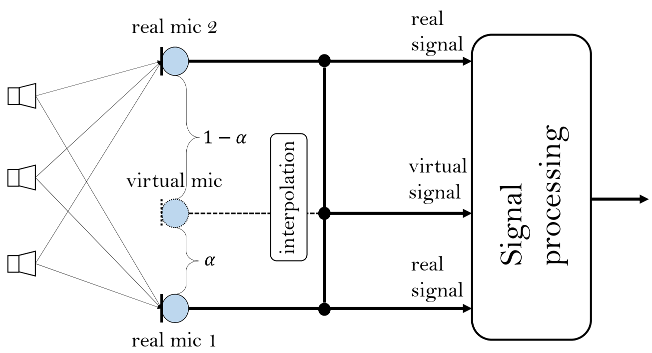


Figure 1: Microphone array signal processing with virtual microphone technique.

On the other hand, the virtual microphone technique [10] allows the well-studied determined methods to be applied to the signals recorded in underdetermined situations by virtually increase the number of channels. With two real microphone signals, the virtual microphone technique estimates the observed signal at a position where is no real microphone placed by interpolating phase and amplitude independently. Based on the W-disjoint orthogonality (W-DO) [11, 4] 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 due to the complicated acoustic environments, 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 performances using a maximum signal-to-noise (maxSNR) beamformer. Furthermore, the results in paper [10] indicated that speech enhancement performances tend to increase when the improved nonlinearity is applied to amplitude interpolation.

However, there is no theoretical proof that β-divergence is the optimal criterion for formulating the amplitude interpolation problem. Because the physical model of amplitude is complicated, it is challenging to design appropriate rules manually. To overcome this problem, this paper proposes using an autoencoder to search for the optimal interpolation domain in a data-driven manner. Specifically, we perform amplitude interpolation in the latent space and train the autoencoder so that the interpolated virtual signals become optimal for conducting beamforming. We use 2-dimensional fully convolutional neural networks (CNNs) to design the autoencoder, which allows the network to handle inputs with arbitrary length and perform interpolation by taking the whole spectrogram into account. This is different from the conventional method using β-divergence, where the interpolation is performed in a time-frequency bin-wise manner.

Related work: An attempt has already been made to employ a CNN to estimate the amplitude of virtual signals [12], where a CNN is trained to output the amplitude directly given the amplitudes of the observed signals and the interpolation position. This paper differs from this method by training an autoencoder to constrain the estimated amplitude to be a mixture signal by explicitly performing interpolation in the latent space and then passing it through the decoder, which is trained as a reconstructor of the mixture signals. This design makes the method more intuitive and the network more interpretable. To overcome the strong assumption made for phase and amplitude interpolation, a method that trains a time-domain TasNet to estimate virtual signals in the time domain has been recently proposed [13]. Although this supervised method has shown high performance in improving speech enhancement performance, training networks using real microphone signals limits the flexibility of location to place the virtual microphone. Moreover, collecting training data in various acoustic conditions are high-cost.

2. Virtual microphone technique based on β-divergence

We model the microphone signals in the STFT domain. Here, let $x_m(\omega, t),\ m = 1, 2$ be the mth real microphone signal at the 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 an observation at the point α obtained by internally dividing the line joining two real microphones in the ratio $\alpha : (1 - \alpha)$ (See Fig 1). β is the hyperparameter of β-divergence, which is used as the metric for amplitude interpolation. Hereafter, we omit ω, t, α, and β for notation simplicity.

We first introduce the interpolation of phase. We assume W-DO [4, 11] for mixed signals that each time-frequency bin is dominated by at most one sound source, which means that the observed signal in each time-frequency bin can be regarded as a single wave. Based on this assumption, the physical model of propagating waves can then be approximated as that of a plane wave. The phase $\phi_\mathrm{v} = \angle v$ of a virtual microphone signal v can then be interpolated linearly on the basis of that 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, we need to make a constraint

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

to eliminate the ambiguity of the interpolated phase.

For the interpolation of amplitude, since there are many acoustic conditions such as the distance between the sound sources and microphones, and direction of arrival (DOA) of sources, it is difficult to faithfully model the amplitude of a propagating wave. Therefore, an interpolation rule based on the β-divergence has been used instead of considering physical models [10]. The β-divergence between the amplitude of a virtual microphone $A_\mathrm{v} = |v|$ 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 $\beta = 1$. The interpolation rule of the amplitude A_v is then given as the closed-form solution of the optimization problem 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 position α:

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

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

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

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

Note that the phase can be interpolated with arbitrary real number 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 $\alpha < 0$ and $\alpha > 1$ was considered in [14].

It has shown that the virtual microphone technique based on the β-divergence is effective in improving speech enhancement performance in underdetermined situations [10]. However, there is no proof that β-divergence is the optimal criterion for amplitude interpolation since we cannot analyze the physical model of amplitude.

3. Proposed method: VMInNet

In this paper, we propose using an autoencoder to search for the optimal interpolation domain in a data-driven manner to overcome the above-mentioned problem. The flowchart of the proposed method is shown in the Fig. 2. In the proposed method, the two-channel observed signals are separated into the amplitude spectrograms A_m and phase spectrograms ϕ_m as the conventional method does. The amplitude spectrogram A_m is then embedded into latent variables $\mathbf{z}_m = E_\theta(A_m)$ by an encoder network $E_\theta(\cdot)$. We interpolate the amplitude of virtual microphone signal at the latent space as

$$\mathbf{z}_\mathrm{v} = (1 - \alpha)\mathbf{z}_1 + \alpha\mathbf{z}_2. \tag{7}$$

The latent variables $\mathbf{z}_1,\ \mathbf{z}_2,\ \mathbf{z}_\mathrm{v}$ are then converted to the amplitudes $A_m = D_\psi(\mathbf{z}_m)$ and $A_\mathrm{v} = D_\psi(\mathbf{z}_\mathrm{v})$ by a decoder

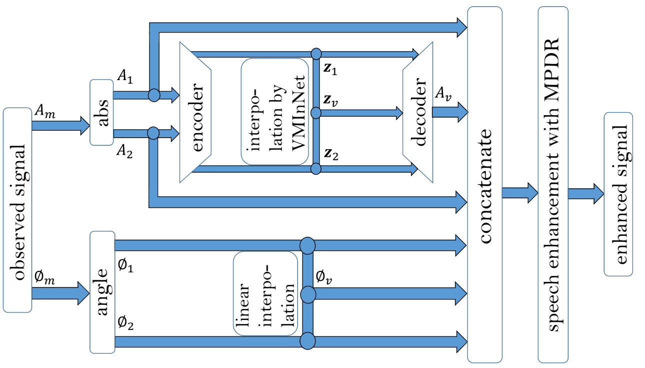
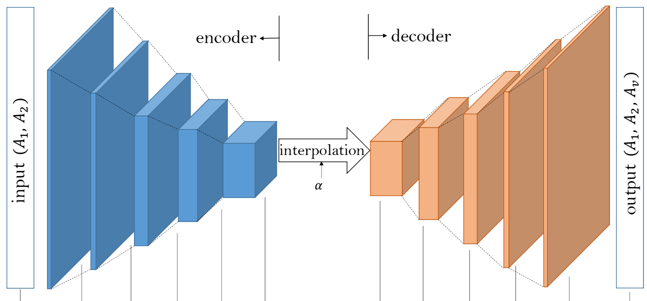


Figure 2: Flowchart of proposed method.



w	513 * 128	257 * 64	129 * 32	65 * 16	33 * 8	17 * 4
c	1	128	256	512	1024	2048
k	5 * 5	5 * 5	5 * 5	5 * 5	5 * 5	

w	17 * 4	33 * 8	65 * 16	129 * 32	257 * 64	513 * 128
c	2048	1024	512	256	128	1
k	5 * 4	5 * 4	5 * 4	5 * 4	5 * 4	

Figure 3: Network architecture of VMInNet. "w", "c", and "k" denote feature size, channel number, and kernel size, respectively.

network $D_\psi(\cdot)$. Here, θ and ψ denote trainable parameters of the encoder and decoder networks. The phase of virtual signal ϕ_{v} is linearly interpolated using (1). The virtual signal v is obtained by concatenating the estimated amplitude and phase spectrogram

$$v = A_{\mathrm{v}} \exp\{j\phi_{\mathrm{v}}\}. \tag{8}$$

We call the network "VMInNet", an abbreviation of "interpolation network for virtual microphone".

To maximize the potential of the data-driven method, where task-dependent loss function is allowed, we train the network using minimum power distortionless response (MPDR) beamformer-based loss function. Specifically, the loss function is designed to minimize the mean squared error between the target signal $s(\omega, t)$ and the output of the MPDR beamformer to enforce the generated amplitude to be optimal for constructing the MPDR beamformer. For the observed signal consisting of two real and I virtual microphone signals, we use $\boldsymbol{x}(\omega, t) = [x_1(\omega,t), v_1(\omega,t), \cdots, v_I(\omega,t), x_2(\omega,t)]^{\mathrm{T}}$ to denote the mixture signal vector. A beamformer that enhances the source of interest is given by

$$y(\omega, t) = \boldsymbol{w}^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega, t), \tag{9}$$

$$\boldsymbol{w}(\omega) = [w_1(\omega) \cdots w_M(\omega)]^{\mathrm{T}}, \tag{10}$$

where $y(\omega, t)$ is the output signal of the beamformer, $\boldsymbol{w}(\omega)$ denotes the spatial filter vector, $(\cdot)^{\mathrm{T}}$ denotes the transpose, $(\cdot)^{\mathrm{H}}$ denotes the Hermitian transpose, and $M = 2 + I$ denotes the number of channels of the observed signals. The spatial filter $\boldsymbol{w}(\omega)$ derived based on the MPDR beamformer is expressed as

$$\boldsymbol{w}(\omega) = \frac{\boldsymbol{\Phi}(\omega)^{-1}\boldsymbol{a}(\omega)}{\boldsymbol{a}^{\mathrm{H}}(\omega)\boldsymbol{\Phi}(\omega)^{-1}\boldsymbol{a}(\omega)}, \tag{11}$$

$$\boldsymbol{\Phi}(\omega) = \mathbb{E}[\boldsymbol{x}(\omega,t)\boldsymbol{x}(\omega,t)^{\mathrm{H}}]. \tag{12}$$

Here, $\boldsymbol{\Phi}(\omega)$ is a covariance matrix of the observed signals at frequency ω, and $\boldsymbol{a}(\omega)$ is the relative transfer function (RTF) of 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 encoder and decoder parameters θ and ψ are trained by minimizing the following loss function:

$$\mathcal{L}(\theta, \psi) = \sum_{m,\omega,t} |D_\psi(E_\theta(A_m(\omega,t))) - A_m(\omega,t)|^2 + \sum_{\omega,t} |\boldsymbol{w}^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega,t) - s(\omega,t)|^2. \tag{13}$$

Here, the first term of (13) is the restoration error of the autoencoder to ensure that the virtual amplitude latent variables generated in the latent space are restored to the amplitude space. The second term is the MPDR beamformer loss. This training criterion allows the network to search for the optimal interpolation space where the interpolated virtual signal is optimal for constructing an MPDR beamformer. Note that other task-dependent loss function can also be considered as long as the loss function is differentiable.

The network architecture is shown in Fig. 3. We use 2-dimensional CNNs to carefully design the network architecture, which allows the network to handle inputs with arbitrary length and perform interpolation by taking the whole spectrogram into account. After each CNN layer, we employ batch normalization to stable the training process.

4. Experiments

4.1 Experimental conditions

To evaluate the effectiveness of the proposed method, we conducted speech enhancement experiments in underdetermined situations. The training dataset consisted of 3 speakers excerpted from the Wall Street Journal (WSJ0) corpus [15] and the audio files for each speaker were 1 minute. The test dataset was comprised of 10 speakers, where the audio files

Table 1: Experimental conditions

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

Table 2: SDR, SIR, SAR [dB] achieved with each method.

Conditions	SDR	SIR	SAR
2 real mic + 1 vir mic (beta)	5.97	8.58	**10.06**
2 real mic + 1 vir mic (cnn)	5.34	**15.54**	5.92
2 real mic + 1 vir mic (prop.)	**7.04**	13.32	8.36
2 real mic	2.18	2.86	12.56
3 real mic	16.07	19.49	18.85

for each speaker were about 9 minutes. The observed mixture signals were generated by adding reverberant signals of 3 speakers together, where the reverberant signals were generated by convolving the impulse responses simulated by a room impulse response generator [16] with clean speech signals. The DOA of the target was set to 90°, and those of interferers were set to 50° and 150° with reverberation time of 120 ms. The experimental conditions are listed in Table 1. We used signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) and signal-to-artifacts ratio (SAR) [17] as evaluation criteria to quantify the speech enhancement performance. Here, we used source image of signals as reference signals.

We chose the method interpolating amplitude based on β-divergence (beta) [10], and that uses a CNN to directly estimate the amplitude of virtual signals (cnn) [12] as baseline methods. We assumed that RTF was known for all the methods. The RTF at the position of virtual microphone is estimated using the conventional β-divergence-based method. The RTF can also be estimated using a neural network, which is one direction of our future work since we would like to focus on amplitude interpolation in this work.

4.2 Results and discussion

Table 2 shows the results obtained by each method. The proposed method achieved the best score in terms of SDR, which confirmed the effectiveness of the proposed method in improving speech enhancement performance. This result indicates that interpolation in latent space is more effective than using β-divergence and directly estimating interpolated amplitudes with a CNN. We also compared the results of using phase or RTF of real microphones and VMInNet for amplitude interpolation to investigate the importance of accurately interpolating amplitude. From these results, we found that although accurate phase and RTF could improve the performance, the improvement was slight, which indicates that the accuracy of amplitude highly influences the speech enhancement performance.

5. Conclusion

In this paper, we proposed VMInNet, an autoencoder network for interpolating the amplitude of virtual microphone signals, which aims to improve speech enhancement performances in underdetermined situations. We trained the network to search for the interpolation domain, which is optimal for conducting an MPDR beamformer. The experimental results revealed that the proposed method outperformed the conventional virtual microphone techniques by achieving a SDR improvement of about 1.1 dB.

Table 3: SDR, SIR, SAR [dB] achieved with each method.

Conditions	SDR	SIR	SAR
amp (prop), phase (linear), tf (beta)	7.04	13.32	8.36
amp (prop), phase (real), tf (beta)	7.38	13.98	8.75
amp (prop), phase (linear), tf(real)	7.37	13.94	8.74
amp (prop), phase (real), tf (real)	7.41	14.02	8.77

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Numbers JPH04131 and JP19J20420.

References

[1] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[2] D. Kitamura *et al.*, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM TASLP*, vol. 24, no. 9, pp. 1622–1637, 2016.

[3] T. Higuchi *et al.*, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM TASLP.*, vol. 25, no. 4, pp. 780–793, 2017.

[4] O. Yılmaz *et al.*, "Blind separation of speech mixtures via time-frequency masking," *IEEE TSP*, vol. 52, no. 7, pp. 1830–1847, 2004.

[5] S. Rickard, "The DUET blind source separation algorithm," in *Blind Speech Separation*, S. Makino *et al.*, Eds., pp. 217–241, Springer, 2007.

[6] H. Sawada *et al.*, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE TASLP*, vol. 19, no. 3, pp. 516–527, 2010.

[7] N. Q. K. Duong *et al.*, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE TASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[8] R. V. Rompaey *et al.*, "GEVD based speech and noise correlation matrix estimation for multichannel wiener filter based noise reduction," *in Proc. EUSIPCO*, pp. 2562–2566, 2018.

[9] A. Ozerov *et al.*, "Multichannel nonnegative matrix factorization in convolutive mixtures for sudio source separation," *IEEE TASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[10] H. Katahira *et al.*, "Nonlinear speech enhancement by virtual increase of channels and maximum SNR beamformer," *EURASIP JASP*, vol. 2016, no. 1, pp. 1–8, Jan. 2016.

[11] A. Jourjine *et al.*, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures," *in Proc. ICASSP*, pp. 2985–2988, 2000.

[12] K. Yamaoka *et al.*, "Cnn-based virtual microphone signal estimation for mpdr beamforming in underdetermined situations," *Proc. EUSIPCO*, pp. 1–5, 2019.

[13] T. Ochiai *et al.*, "Neural network-based virtual microphone estimator," *arXiv preprint arXiv:2101.04315*, 2021.

[14] R. jinzai *et al.*, "Microphone position realignment by extrapolation of virtual microphone," *in Proc. APSIPA*, pp. 367–372, 2018.

[15] Garofolo *et al.*, "Csr-i (wsj0) complete ldc93s6a," Web Download. Philadelphia: Linguistic Data Consortium., 1993.

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

[17] E. Vincent *et al.*, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.