INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Variational Bayesian Multi-channel Speech Dereverberation under Noisy Environments with Probabilistic Convolutive Transfer Function

Masahito Togami[1], *Tatsuya Komatsu*[1]

[1]LINE Corporation, Tokyo, Japan
masahito.togami@linecorp.com

Abstract

In this paper, we propose a multi-channel speech dereverberation method which can reduce reverberation even when acoustic transfer functions (ATFs) are time varying under noisy environments. The microphone input signal is modeled as a convolutive mixture in a time-frequency domain so as to incorporate late reverberation whose tap length is longer than frame size of short term Fourier transform. To reduce reverberation effectively under the time-varying ATF conditions, the proposed method extends the deterministic convolutive transfer function (D-CTF) into a probabilistic convolutive transfer function (P-CTF). A variational Bayesian framework was applied to approximation of a joint posterior probability density functions of a speech source signal and the ATFs. Variational posterior probability density functions and the other parameters are iteratively updated so as to maximize an evidence lower bound (ELBO). Experimental results when the ATFs are time-varying and there is background noise showed that the proposed method can reduce reverberation more accurately than the Weighted Prediction error (WPE) and the Kalman-EM for dereverberation (KEMD).

Index Terms: speech dereverberation, time-varying acoustic transfer function, variational Bayesian linear state-space model, noise reduction

1. Introduction

Reverberation which is a reflection of sound on walls, ceilings, and floors is harmful for human listening devices or automatic speech recognition systems. To reduce reverberation contaminated in a microphone input signal, speech dereverberation techniques have been studied for a long time [1].

Theoretically speaking, complete dereverberation can be assured based on the Multi-INput multi-output Theorem (MINT) [2] when a multi-channel input signal is available and the acoustic transfer functions (ATFs) do not have any common poles. The MINT requires for estimation of the ATFs. Two stage methods which consist of a blind channel identification stage [3–6] and a multi-channel spatial inverse filtering stage have been commonly utilized. However, the MINT is highly sensitive to the estimation error of the ATFs. In addition to that, it is assumed that that the ATFs are time-invariant in the MINT, and dereverberation performance of the MINT based methods degrades when the ATFs are time-varying. In the actual environments, the ATFs easily fluctuate due to movement of a human head, fluctuation of temperature, and so on. Therefore, how to reduce reverberation stably against the estimation error of the ATFs and how to reduce reverberation even when the ATFs are time-varying are important topics.

Dereverberation techniques based on Auto-Regressive (AR) model of multi-channel microphone input signal which does not require for estimation of ATFs have been actively studied [7–13]. Weighted Prediction Error (WPE) [8] is a popular approach based on the AR model. The WPE estimates the time-varying variance of the speech source signal and the AR coefficient in an iterative manner. Although the AR model based speech dereverberation is based on multi-channel spatial inverse filtering [9], the WPE can reduce reverberation more stably than multi-channel spatial inverse filtering of the ATFs. However, in the WPE, it is also assumed that the ATFs are time-invariant. Even though online algorithms [14–17] are utilized, it is highly difficult to track fast change of the ATFs such as movement of a human head. The AR model of the noisy multi-channel microphone input signal is also problematic when there is background noise signal, because estimation accuracy of the AR coefficients degrades due to existence of background noise signal [18].

Another category of speech dereverberation is simultaneous estimation of the ATFs and parameters of a probability density function (PDF) of a speech source signal [19–22]. Based on the expectation-maximization (EM) framework [23], all of the parameters are updated to increase the likelihood function monotonically. In [20], the reverberation system is approximated by the deterministic convolutive transfer function (D-CTF) [24], and the posterior PDF of the speech source signal is estimated via a Kalman Smoother framework [25]. In [22], a variational Bayesian based approach with the D-CTF has been also proposed. However, the assumption that the ATFs are time-invariant is not adequate in the actual situation, because the ATFs easily fluctuate due to movement of a human head, and so on. Based on the D-CTF, when the ATFs are time-varying, speech dereverberation performance degrades. Another approach which estimates the ATFs and the parameters of the PDF of the speech source signal simultaneously is a variational Bayesian based approach for a time-varying acoustic channel in the time-frequency domain [21, 22, 26]. However, in this framework, it is assumed that the reverberation system is an instantaneous mixture in the time-frequency domain. When the frame size is sufficiently long, the instantaneous mixture model is adequate in the time-frequency domain, but the long frame size is not good for modeling of a speech signal which has non-stationary characteristics. Therefore, it is needed to consider a convolutive mixture in the time-frequency domain.

In this paper, instead of the conventional D-CTF, we propose a probabilistic convolutive transfer function (P-CTF) in which the ATFs are assumed to be time-varying. The proposed method optimizes the probabilistic model of the ATFs and the speech source model in the variational Bayesian framework. The proposed method can be regarded as a natural extension of the D-CTF based speech dereverberation method with the Kalman smoother [20] into a time-varying ATF scenario. We perform speech dereverberation experiments under a time-invariant ATF scenario and a time-varying ATF scenario w/ background noise and w/o background noise. Experimental results show that the proposed method can reduce reverberation more effectively than the WPE and the Kalman-EM for dere-

http://dx.doi.org/10.21437/Interspeech.2019-1220

verberation (KEMD) [20]. In addition to that, it is confirmed that estimation accuracy of the dereverberation parameters is improved under noisy environments in the proposed method.

2. Problem statement

2.1. Microphone input signal model

Let $\boldsymbol{x}_{l,k} \in \mathbb{C}^{N_m}$ be an observed multi-channel microphone input signal at the frame l and the frequency k. When l is omitted, $\boldsymbol{x}_k$ denotes all signals at all frames. We model $\boldsymbol{x}_{l,k}$ as the following convolutive transfer function (CTF):

$$\boldsymbol{x}_{l,k} = \boldsymbol{A}_{l,k}\boldsymbol{s}_{l,k} + \boldsymbol{w}_{l,k}, \tag{1}$$

where $\boldsymbol{A}_{l,k}$ is a $N_m \times L_\tau$ (L_τ is the tap-length of a ATF) time-varying multi-channel ATF matrix, $\boldsymbol{w}_{l,k}$ is a N_m dimensional background noise signal, and $\boldsymbol{s}_{l,k}$ is a L_τ dimensional delay line of the original source signal, which is defined as follows:

$$\boldsymbol{s}_{l,k} = [\begin{array}{ccc} s_{l,k} & \dots & s_{l-L_\tau+1,k} \end{array}]^T, \tag{2}$$

where T is the transpose operator of a matrix/vector. The goal of speech dereverberation and noise reduction is defined as estimation of $s_{l,k}$ from the noisy and reverberant multi-channel microphone input signal $\boldsymbol{x}_{l,k}$. However, there is a scale ambiguity between $\boldsymbol{A}_{l,k}$ and $\boldsymbol{s}_{l,k}$, because $\boldsymbol{A}_{l,k}\boldsymbol{s}_{l,k} = (\alpha\boldsymbol{A}_{l,k})(\frac{1}{\alpha}\boldsymbol{s}_{l,k})$. Therefore, in this paper, we redefine the objective of speech dereverberation as to estimate $\boldsymbol{A}_{l,\tau=0,k}s_{l,k}$, where $\boldsymbol{A}_{l,\tau,k}$ is the τ-th column of the matrix $\boldsymbol{A}_{l,k}$.

3. Proposed method

3.1. Overview

Instead of the D-CTF, the proposed method utilizes a probabilistic convolutive transfer function (P-CTF) defined in Eq. (1). The probability density function (PDF) of the ATFs is defined as a time-invariant Gaussian distribution as follows:

$$p(\boldsymbol{A}_{l,k}) = p(\text{vec}\boldsymbol{A}_{l,k}) \sim \mathcal{N}(\boldsymbol{\mu}_{va,k}, \boldsymbol{R}_{va,k}), \tag{3}$$

where $\text{vec}\boldsymbol{X} = [\begin{array}{ccc} \boldsymbol{X}_1^T & \cdots & \boldsymbol{X}_K^T \end{array}]^T$ ($\boldsymbol{X}_i$ is the i-th column of the matrix $\boldsymbol{X}$ and K is the number of the columns of $\boldsymbol{X}$). The probabilistic model of the speech source signal is defined as a time-varying Gaussian distribution as follows:

$$p(s_{l,k}) \sim \mathcal{N}(0, v_{l,k}), \tag{4}$$

where $v_{l,k}$ is a time-varying variance of the speech source signal. The noise term is modeled as a time-invariant multi-channel Gaussian distribution as follows:

$$p(\boldsymbol{w}_{l,k}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_{w,k}). \tag{5}$$

The proposed method estimates the latent variables, $\boldsymbol{A}_k$ and $\boldsymbol{s}_k$, from the microphone input signal $\boldsymbol{x}_k$. Let θ_k be the dereverberation parameter for the frequency k. The objective function for the parameter estimation is a log-likelihood function defined as follows:

$$\mathcal{LL}(\boldsymbol{x}_k, \theta_k) = \log p(\boldsymbol{x}_k|\theta_k). \tag{6}$$

Generally speaking, it is highly difficult to optimize the original log-likelihood function directly. Instead, the evidence lower bound (ELBO) is utilized as an alternate objective function, which is defined as the right term of the following equation:

$$\mathcal{LL}(\boldsymbol{x}_k, \theta_k) \geq \int_{\boldsymbol{A}_k,\boldsymbol{s}_k} q_{\boldsymbol{A}_k,\boldsymbol{s}_k} \log \frac{p_{\boldsymbol{x}_k,\boldsymbol{A}_k,\boldsymbol{s}_k|\theta_k}}{q_{\boldsymbol{A}_k,\boldsymbol{s}_k}} d\boldsymbol{A}_k d\boldsymbol{s}_k, \tag{7}$$

where a conditional PDF $p(x|y)$ is written by $p_{x|y}$ for simplicity. $q_{\boldsymbol{A}_k,\boldsymbol{s}_k}$ is a joint variational posterior PDF of $\boldsymbol{A}_k$ and $\boldsymbol{s}_k$, where $\int_{\boldsymbol{A}_k,\boldsymbol{s}_k} q_{\boldsymbol{A}_k,\boldsymbol{s}_k} d\boldsymbol{A}_k d\boldsymbol{s}_k = 1$. The right term of Eq. (7) is called as the ELBO. The ELBO can be maximized by updating $q_{\boldsymbol{A}_k,\boldsymbol{s}_k}$ and θ_k in an iterative manner. Let $\theta_k^{(t)}$ be the tentative parameter after the t-th iteration. When $\theta_k^{(t)}$ is obtained, $q_{\boldsymbol{A}_k,\boldsymbol{s}_k}$ which maximizes the ELBO can be obtained as follows:

$$q_{\boldsymbol{A}_k,\boldsymbol{s}_k} = p_{\boldsymbol{A}_k,\boldsymbol{s}_k|\boldsymbol{x}_k,\theta_k^{(t)}}. \tag{8}$$

In this case, the ELBO is equals to the Q function in the expectation-maximization (EM) framework [23], and monotonic increase of the ELBO leads to monotonic increase of the log-likelihood function, but it is also difficult to estimate the joint posterior PDF $p_{\boldsymbol{A}_k,\boldsymbol{s}_k|\boldsymbol{x}_k,\theta_k^{(t)}}$ directly. Alternatively, the proposed method approximates the joint posterior PDF as follows:

$$p_{\boldsymbol{A}_k,\boldsymbol{s}_k|\boldsymbol{x}_k,\theta_k^{(t)}} = q_{\boldsymbol{A}_k} q_{\boldsymbol{s}_k}, \tag{9}$$

where $q_{\boldsymbol{A}_k}$ is a variational posterior PDF of $\boldsymbol{A}_k$ and $q_{\boldsymbol{s}_k}$ is a variational posterior PDF of $\boldsymbol{s}_k$. In the proposed method, $q_{\boldsymbol{A}_k}$ and $q_{\boldsymbol{s}_k}$ are updated in an iterative manner so as to increase the ELBO monotonically based on the variational Bayesian framework as follows:

$$q_{\boldsymbol{A}_k}^{(t+1)} = \underset{q_{\boldsymbol{A}_k} \in \{q_{\boldsymbol{A}_k} | \int_{\boldsymbol{A}_k} q_{\boldsymbol{A}_k} d\boldsymbol{A}_k = 1\}}{\arg\max} \text{ELBO}(\theta_k^{(t)}, q_{\boldsymbol{s}_k}^{(t)}, q_{\boldsymbol{A}_k}), \tag{10}$$

$$q_{\boldsymbol{s}_k}^{(t+1)} = \underset{q_{\boldsymbol{s}_k} \in \{q_{\boldsymbol{s}_k} | \int_{\boldsymbol{s}_k} q_{\boldsymbol{s}_k} d\boldsymbol{s}_k = 1\}}{\arg\max} \text{ELBO}(\theta_k^{(t)}, q_{\boldsymbol{s}_k}, q_{\boldsymbol{A}_k}^{(t+1)}). \tag{11}$$

The other parameters are also updated so as to increase the ELBO as follows:

$$\theta_k^{(t+1)} = \underset{\theta_k}{\arg\max}\, \text{ELBO}(\theta_k, q_{\boldsymbol{s}_k}^{(t+1)}, q_{\boldsymbol{A}_k}^{(t+1)}). \tag{12}$$

The proposed method performs Eq. (10), Eq. (11), and Eq. (12) in an iterative manner.

3.2. Update of $q_{\boldsymbol{A}_k}$

Based on Eq. (10), $q_{\boldsymbol{A}_k}^{(t+1)}$ can be obtained as follows:

$$\begin{aligned} q_{\boldsymbol{A}_k}^{(t+1)} &= \frac{\exp\langle \log p_{\boldsymbol{x}_k,\boldsymbol{A}_k,\boldsymbol{s}_k|\theta_k^{(t)}} \rangle_{q_{\boldsymbol{s}_k}^{(t)}}}{\int_{\boldsymbol{s}_k} \exp\langle \log p_{\boldsymbol{x}_k,\boldsymbol{A}_k,\boldsymbol{s}_k|\theta_k^{(t)}} \rangle_{q_{\boldsymbol{s}_k}^{(t)}} d\boldsymbol{s}_k} \\ &\sim p_{\boldsymbol{A}_k|\theta_k^{(t)}} \exp\langle \log p_{\boldsymbol{x}_k|\boldsymbol{s}_k,\boldsymbol{A}_k,\theta_k^{(t)}} \rangle_{q_{\boldsymbol{s}_k}^{(t)}}. \end{aligned} \tag{13}$$

Under the independence assumption of $\boldsymbol{A}_{l,k}$ and $\boldsymbol{w}_{l,k}$ along the frame axis, $q_{\boldsymbol{A}_k}^{(t+1)}$ can be decomposed as $\prod_l q_{\boldsymbol{A}_{l,k}}^{(t+1)}$ along the frame axis. Because both $p_{\boldsymbol{A}_{l,k}|\theta_k^{(t)}}$ and $p_{\boldsymbol{x}_{l,k}|\boldsymbol{s}_k,\boldsymbol{A}_{l,k}\theta_k^{(t)}}$ are defined as Gaussian distributions, $q_{\boldsymbol{A}_{l,k}}^{(t+1)}$ is also a Gaussian distribution, which is defined as follows:

$$q_{\boldsymbol{A}_{l,k}}^{(t+1)} \sim \mathcal{N}(\boldsymbol{\mu}_{q_{\boldsymbol{A}_{l,k}}^{(t+1)}}, \boldsymbol{R}_{q_{\boldsymbol{A}_{l,k}}^{(t+1)}}). \tag{14}$$

In the next section, We show that $q_{\boldsymbol{s}_{l,k}}^{(t)}$ follows a Gaussian distribution as follows:

$$q_{\boldsymbol{s}_{l,k}}^{(t)} \sim \mathcal{N}(\boldsymbol{\mu}_{q_{\boldsymbol{s}_{l,k}}^{(t+1)}}, \boldsymbol{R}_{q_{\boldsymbol{s}_{l,k}}^{(t+1)}}). \tag{15}$$

$\boldsymbol{\mu}_{q_{\boldsymbol{A}_{l,k}}^{(t+1)}}$ can be calculated with $q_{\boldsymbol{s}_{l,k}}^{(t)}$ as follows:

$$\boldsymbol{\mu}_{q_{\boldsymbol{A}_{l,k}}^{(t+1)}} = \boldsymbol{W}_{A,l,k}(\bar{\boldsymbol{x}}_{l,k} - \bar{\boldsymbol{S}}_{l,k}\boldsymbol{\mu}_{va,k}) + \boldsymbol{\mu}_{va,k}, \tag{16}$$

where $\bar{\boldsymbol{x}}_{l,k}$ is an extended observation vector, $\bar{\boldsymbol{x}}_{l,k} = [\begin{array}{cc} \boldsymbol{x}_{l,k} & \boldsymbol{0} \end{array}]^T$, $\bar{\boldsymbol{S}}_{l,k} = \begin{pmatrix} \tilde{\boldsymbol{S}}_{l,k} \\ \boldsymbol{D}_{l,k}^H \end{pmatrix}$, $\tilde{\boldsymbol{S}}_{l,k} = \boldsymbol{\mu}_{q_{\boldsymbol{s}_{l,k}}^{(t+1)},T} \otimes \boldsymbol{I}_{N_m \times N_m}$

($\otimes$ is an operator of Kronecker product of two matrices and $\boldsymbol{I}$ is an identity matrix), $\boldsymbol{R}_{q^{(t+1),T}_{\boldsymbol{s}_{l,k}}} \otimes \boldsymbol{R}^{-1}_{w,k} = \boldsymbol{D}_{l,k}\boldsymbol{D}^H_{l,k}$, $\boldsymbol{D}_{l,k}$ can be obtained via a Cholesky decomposition, and $\boldsymbol{W}_{A,l,k}$ is a multi-channel Wiener filter which is calculated as follows:

$$\boldsymbol{W}_{A,l,k} = \boldsymbol{R}_{va,k}\bar{\boldsymbol{S}}^H_{l,k}(\bar{\boldsymbol{R}}_{w,k} + \bar{\boldsymbol{S}}_{l,k}\boldsymbol{R}_{va,k}\bar{\boldsymbol{S}}^H_{l,k})^{-1}, \quad (17)$$

$$\bar{\boldsymbol{R}}_{w,k} = \begin{pmatrix} \boldsymbol{R}_{w,k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{N_m L_\tau \times N_m L_\tau} \end{pmatrix}. \quad (18)$$

$\boldsymbol{R}_{q^{(t+1)}_{\boldsymbol{A}_{l,k}}}$ can be calculated as follows:

$$\boldsymbol{R}_{q^{(t+1)}_{\boldsymbol{A}_{l,k}}} = (\boldsymbol{I} - \boldsymbol{W}_{A,l,k}\bar{\boldsymbol{S}}_{l,k})\boldsymbol{R}_{va,k}. \quad (19)$$

In Eq. (17), $\boldsymbol{W}_{A,l,k}$ is a $L_\tau \times (1 + L_\tau)N_m$ matrix. Therefore, instead of the multi-channel Wiener filtering of the multi-microphone input signal, Eq. (17) is the multi-channel Wiener filtering for the extended microphone input signal which reflects uncertainty of estimation of the speech source signal.

3.3. Update of $q_{\boldsymbol{s}_k}$

Based on Eq. (11), $q^{(t+1)}_{\boldsymbol{s}_k}$ is given as follows:

$$q^{(t+1)}_{\boldsymbol{s}_k} \sim p_{\boldsymbol{s}_k|\theta^{(t)}} \exp\langle \log p_{\boldsymbol{x}_k|\boldsymbol{s}_k,\boldsymbol{A}_k,\theta^{(t)}} \rangle_{q^{(t+1)}_{\boldsymbol{A}_k}}, \quad (20)$$

$q^{(t+1)}_{\boldsymbol{s}_k}$ is a Gaussian distribution, because $p_{\boldsymbol{s}_k|\theta^{(t)}}$ and $p_{\boldsymbol{x}_k|\boldsymbol{s}_k,\boldsymbol{A}_k,\theta^{(t)}}$ are Gaussian distributions. $p_{\boldsymbol{s}_k|\theta^{(t)}}$ and $p_{\boldsymbol{x}_k|\boldsymbol{s}_k,\boldsymbol{A}_k,\theta^{(t)}}$ are factorized as follows:

$$p_{\boldsymbol{s}_k|\theta^{(t)}} = p_{\boldsymbol{s}_{l=0,k}|\theta^{(t)}} \prod_l p_{\boldsymbol{s}_{l,k}|\boldsymbol{s}_{l-1,k},\theta^{(t)}}, \quad (21)$$

$$p_{\boldsymbol{x}_k|\boldsymbol{s}_k,\boldsymbol{A}_k,\theta^{(t)}} = \prod_l p_{\boldsymbol{x}_{l,k}|\boldsymbol{s}_{l,k},\boldsymbol{A}_{l,k},\theta^{(t)}}. \quad (22)$$

Therefore, $\boldsymbol{s}_k$ can be modeled as the following state-transition equation:

State transition equation:

$$\boldsymbol{s}_{l,k} = \boldsymbol{G}\boldsymbol{s}_{l-1,k} + \boldsymbol{u}_{l,k}, \quad (23)$$

Observation equation:

$$\bar{\boldsymbol{x}}_{l,k} = \bar{\boldsymbol{A}}_{l,k}\boldsymbol{s}_{l,k} + \bar{\boldsymbol{w}}_{l,k}, \quad (24)$$

where

$$\begin{aligned}
\boldsymbol{G} &= \begin{pmatrix} \multicolumn{2}{c}{\boldsymbol{0}_{1\times L_\tau}} \\ \boldsymbol{I}_{L_\tau-1\times L_\tau-1} & \boldsymbol{0}_{L_\tau-1\times 1} \end{pmatrix}, & (25)\\
p(\boldsymbol{u}_{l,k}) &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_{u,l,k}), & (26)\\
\boldsymbol{R}_{u,l,k} &= \begin{pmatrix} v_{l,k} & \boldsymbol{0}_{1\times L_\tau-1} \\ \boldsymbol{0}_{L_\tau-1\times 1} & \boldsymbol{0}_{L_\tau-1\times L_\tau-1} \end{pmatrix}, & (27)\\
\bar{\boldsymbol{x}}_{l,k} &= [\ \boldsymbol{x}_{l,k} \quad \boldsymbol{0}\ ]^T, & (28)\\
\bar{\boldsymbol{A}}_{l,k} &= [\ \tilde{\boldsymbol{A}}^{(t+1)}_{l,k} \quad \boldsymbol{L}^H_{l,k}\ ]^T, & (29)\\
\boldsymbol{L}_{l,k}\boldsymbol{L}^H_{l,k} &= E[\boldsymbol{A}^H_{l,k}\boldsymbol{R}^{-1}_{w,k}\boldsymbol{A}_{l,k}]_{q^{(t+1)}_{\boldsymbol{A}_{l,k}}} & \\
&- \tilde{\boldsymbol{A}}^{(t+1),H}_{l,k}\boldsymbol{R}^{-1}_{w,k}\tilde{\boldsymbol{A}}^{(t+1)}_{l,k}, & (30)\\
\mathrm{vec}\tilde{\boldsymbol{A}}^{(t+1)}_{l,k} &= \boldsymbol{\mu}^{(t+1)}_{q_{\boldsymbol{A}_{l,k}}}, & (31)\\
\bar{\boldsymbol{w}}_{l,k} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_{w,k}), & (32)\\
\boldsymbol{R}_{w,k} &= \begin{pmatrix} \boldsymbol{R}_{w,k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}. & (33)
\end{aligned}$$

Similar to the update of $q_{\boldsymbol{A}_k}$, the observation equation is an extended observation equation which reflects uncertainty of the ATFs, and $\boldsymbol{L}_{l,k}$ is obtained via a Cholesky decomposition. Therefore, $q^{(t+1)}_{\boldsymbol{s}_k}$ can be calculated with the Kalman smoother framework [25] for the extended state-space model. By integral out of $q^{(t+1)}_{\boldsymbol{s}_k}$ along the frame axis, $q^{(t+1)}_{\boldsymbol{s}_{l,k}}$ can be estimated.

3.4. Parameter optimization

In the proposed method, the parameter θ_k is defined as $\theta_k = \{\boldsymbol{\mu}_{va,k}, \boldsymbol{R}_{va,k}, \boldsymbol{R}_{w,k}, v_k, \boldsymbol{\pi}_k, \boldsymbol{z}_k\}$. Based on Eq. (12), the parameter $\theta^{(t+1)}_k$ is updated as follows:

$$\begin{aligned}
\boldsymbol{\mu}^{(t+1)}_{va,k} &= \frac{1}{L_T}\sum_l \boldsymbol{\mu}_{q^{(t+1)}_{\boldsymbol{A}_{l,k}}}, & (34)\\
\boldsymbol{R}^{(t+1)}_{va,k} &= \frac{1}{L_T}\sum_l \boldsymbol{R}_{q^{(t+1)}_{\boldsymbol{A}_{l,k}}} & \\
&+ (\boldsymbol{\mu}_{q^{(t+1)}_{\boldsymbol{A}_{l,k}}} - \boldsymbol{\mu}^{(t+1)}_{va,k})(\boldsymbol{\mu}_{q^{(t+1)}_{\boldsymbol{A}_{l,k}}} - \boldsymbol{\mu}^{(t+1)}_{va,k})^H, & \\
\boldsymbol{R}^{(t+1)}_{w,k} &= \frac{1}{L_T}\sum_l E[\boldsymbol{w}_{l,k}\boldsymbol{w}^H_{l,k}]_{q^{(t+1)}_{\boldsymbol{A}_{l,k}}, q^{(t+1)}_{\boldsymbol{s}_{l,k}}}, & \\
\boldsymbol{R}^{(t+1)}_{w,k} &\leftarrow \text{off-diag}\boldsymbol{R}^{(t+1)}_{w,k}, & (35)
\end{aligned}$$

L_T is the number of the time-frames, off-diag is an operator which replaces value of each non-diagonal element of a matrix into 0,

$$\begin{aligned}
\boldsymbol{\pi}_k &= \boldsymbol{\mu}_{q^{(t+1)}_{\boldsymbol{s}_{l=0,k}}}, & (36)\\
\boldsymbol{z}_k &= \boldsymbol{R}_{q^{(t+1)}_{\boldsymbol{s}_{l=0,k}}}, & (37)\\
v^{(t+1)}_{l,k} &= \|\boldsymbol{\mu}_{q^{(t+1)}_{\boldsymbol{s}_{l,k}}}(1)\|^2 + \boldsymbol{R}_{q^{(t+1)}_{\boldsymbol{s}_{l,k}}}(1,1), & (38)
\end{aligned}$$

$x(1)$ is the 1st element of the vector x, and $x(1,1)$ is the 1st row and the 1st column element of the matrix x.

3.5. Estimation of output signal

After estimating parameters and the variational approximated posterior PDF, the dereverberated and noise reduced signal can be obtained as follows:

$$\boldsymbol{c}_{l,k} = \tilde{s}_{l,k}\tilde{\boldsymbol{A}}_{l,\tau=0,k}, \quad (39)$$

where $\tilde{s}_{l,k} = \boldsymbol{\mu}_{q_{\boldsymbol{s}_{l,k}}}(1)$. The dereverberated signal without noise reduction can be also obtained as follows:

$$\begin{aligned}
\boldsymbol{e}_{l,k} &= \tilde{s}_{l,k}\tilde{\boldsymbol{A}}_{l,\tau=0,k} + \tilde{\boldsymbol{w}}_{l,k} & \\
&= \boldsymbol{x}_{l,k} - \sum_{\tau=1}^{L_\tau-1} \tilde{s}_{l-\tau,k}\tilde{\boldsymbol{A}}_{l,\tau,k}. & (40)
\end{aligned}$$

4. Experiment

4.1. Experimental setup

Speech dereverberation and noise reduction performance were evaluated. The number of the microphones, N_m, was set to 2. Sampling rate was set to 16000 Hz. Framesize was 1024 pt, and frame shift was 512 pt. The number of the speech sources was set to 1. Multi-channel data was generated by convolving the measured impulse responses with the clean speech sources. The clean speech sources were extracted from TIMIT test corpus [27]. As the measured impulse responses, Multi-Channel Impulse Response Database [28] was utilized. Impulse responses of the 1st impulse response and the 2nd one from a linear microphone array with spacing "3-3-3-8-3-3-3"(cm) were utilized. The reverberation time RT_{60} was 0.61 (sec). The assumed tap-length L_τ was set to 10. The azimuth of the speech source was set to 0 degrees. The distance between microphones and

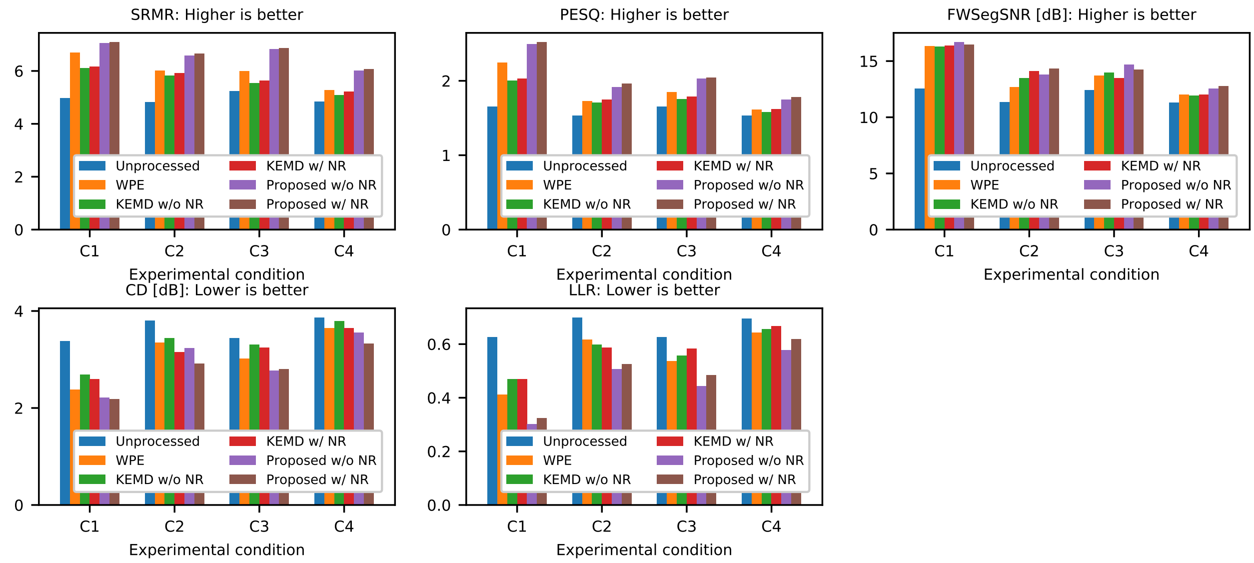

Figure 1: *Experimental results: in "C1", there is no background noise and impulse responses are time-invariant, in "C2", there is background noise and impulse responses are time-invariant, in "C3", there is no background noise and impulse responses are time-varying, and in "C4", there is background noise and impulse responses are time-varying.*

speech source location was set to 1 m and 2 m. The number of the clean speech sources that utilized in the experiment was 10. In the experiments under noisy environments, we selected "office", "cafeteria", and "meeting" noise from DEMAND dataset [29]. For each noise type, the randomly extracted noise signals were convoluved with the impulse response from each azimuth (0,15,30,45,60,75,90,270,285,300,315,330,345 degrees) and mixed so as to mimic diffuse noise. SNR was set to 20 dB. Therefore, in the noiseless environments, there were total 20 samples. In the noisy environments, there were total 60 samples. The number of the iterations were set to 100 in the proposed method. In the time-varying impulse responses case, the impulse response $a_{m,d=bL+l,i}$ (L was set to 4800 sample, m is the microphone index, i is the tap index, and d is the frame index. frame was set to 256 samples) was generated as follows:

$$\begin{aligned} a_{m,d=bL+l,i} &= (1-|\alpha_{bL+l}|)a_{m,\theta=0,i} \\ &+ \max(0,\alpha_{bL+l})a_{m,\theta=15,i} \\ &+ \max(0,-\alpha_{bL+l})a_{m,\theta=345,i}, \end{aligned} \tag{41}$$

where $a_{m,\theta,i}$ is the impulse response of the azimuth θ degrees,

$$\alpha_{bL+l} = \frac{l}{L}\beta_b + \frac{L-l}{L}\beta_{b+1}, \tag{42}$$

$$p(\beta_b) \sim \mathcal{N}(0,1). \tag{43}$$

4.2. Evaluation measures

We utilized five evaluation measures which were defined in REVERB Challenge [30], i.e., Cepstrum distance (CD) [dB], Log likelihood ratio (LLR), Frequency-weighted segmental SNR (FWSegSNR) [dB], Speech-to-reverberation modulation energy ratio (SRMR), and Perceptual Evaluation of Speech Quality (PESQ).

4.3. Comparative methods

The proposed method was compared with the WPE which is implemented in [31] and the Kalman-EM for dereverberation (KEMD) [20]. The number of the iterations were set to 100 for the KEMD. We changed the number of the iterations in the WPE, and the number of the delay frames. Eventually, three was the best for the number of the iterations in the WPE, and one was the best for the number of the delay frames. Therefore, these parameters were utilized in the WPE. In the KEMD and the proposed method, in addition to dereverberation performance with noise reduction (w/ NR), we also evaluated dereverberation performance with no noise reduction case (w/o NR) so as to evaluate only speech dereverberation performance under noisy environments. The assumed tap-length in the KEMD and the length of the AR coefficient in the WPE are the same as the assumed tap-length in the proposed method.

4.4. Experimental results

Experimental results for the four conditions are shown in Fig. 1. It is shown that the proposed method outperformed the WPE in each condition. Especially, from the comparison between the WPE and the proposed method w/o NR, it is shown that the proposed method can estimate dereverberation parameters under noisy environments more robustly than the WPE. The proposed method also outperformed the KEMD in time-invariant cases and time-varying cases. From the comparison between "proposed method w/ NR" and "proposed method w/o NR", it is shown that "proposed method w/ NR" was better than "proposed method w/o NR" except for the LLR results under noisy environments.

5. Conclusions

In this paper, we proposed a speech dereverberation and noise reduction method which is based on a probabilistic convolutive transfer function (P-CTF). A variational Bayesian based method is prposed for estimating the P-CTF and the variational probability density function of the speech source signal based on a state space model with an extended observation model which reflects uncertainty of the ATFs. Experimental results show that the proposed method estimated dereverberation parameters under noisy environments more robustly than the conventional methods. The proposed method outperformed the conventional methods in both time-invariant cases and time-varying cases.

6. References

[1] P. Naylor and N. Gaubitch, *Speech Dereverberation*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[2] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 2, pp. 145–152, Feb 1988.

[3] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Transactions on Signal Processing*, vol. 43, no. 12, pp. 2982–2993, Dec 1995.

[4] Y. Huang and J. Benesty, "Adaptive blind channel identification: Multi-channel least mean square and newton algorithms," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, May 2002, pp. II–1637–II–1640.

[5] ——, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Transactions on Signal Processing*, vol. 51, no. 1, pp. 11–24, Jan 2003.

[6] Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 882–895, Sep. 2005.

[7] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, May 2009.

[8] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, Sept 2010.

[9] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 69–84, Jan 2011.

[10] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1369–1380, July 2013.

[11] B. Li, T. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. Chin, K. Sim, R. Weiss, K. Wilson, E. Variani, C. Kim, O. Siohan, M. Weintraub, E. McDermott, R. Rose, and M. Shannon, "Acoustic modeling for google home," 2017.

[12] T. Otsuka, K. Ishiguro, T. Yoshioka, H. Sawada, and H. G. Okuno, "Multichannel sound source dereverberation and separation for arbitrary number of sources based on bayesian nonparametrics," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2218–2232, Dec 2014.

[13] H. Kagami, H. Kameoka, and M. Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 31–35.

[14] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, "Adaptive dereverberation of speech signals with speaker-position change detection," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 3733–3736.

[15] M. Togami, "Multichannel online speech dereverberation under noisy environments," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, Aug 2015, pp. 1078–1082.

[16] S. Braun and E. A. P. Habets, "Online dereverberation for dynamic scenarios using a kalman filter with an autoregressive model," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1741–1745, Dec 2016.

[17] ——, "Linear prediction-based online dereverberation and noise reduction using alternating kalman filters," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1119–1129, June 2018.

[18] M. Togami and Y. Kawaguchi, "Noise robust speech dereverberation with kalman smoother," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7447–7451.

[19] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. MIT Press, 2001, pp. 758–764. [Online]. Available: http://papers.nips.cc/paper/1908-speech-denoising-and-dereverberation-using-probabilistic-models.pdf

[20] B. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation using expectation-maximization and kalman smoothing," in *21st European Signal Processing Conference (EUSIPCO 2013)*, Sep. 2013, pp. 1–5.

[21] D. Schmid, G. Enzner, S. Malik, D. Kolossa, and R. Martin, "Variational bayesian inference for multichannel dereverberation and noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 8, pp. 1320–1335, Aug 2014.

[22] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Speech dereverberation with convolutive transfer function approximation using map and variational deconvolution approaches," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2014, pp. 50–54.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[24] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[25] A. Jazwinski, *Stochastic processes and filtering theory*, ser. Mathematics in science and engineering. New York, NY: Acad. Press, 1970, no. 64.

[26] Y. Laufer and S. Gannot, "A bayesian hierarchical model for speech enhancement with time-varying audio channel," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 225–239, Jan 2019.

[27] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.

[28] "Multi-Channel Impulse Response Database," https://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/multi-channel-impulse-response-database/.

[29] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," Jun. 2013, Supported by Inria under the Associate Team Program VERSAMUS. [Online]. Available: https://doi.org/10.5281/zenodo.1227121

[30] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *Eurasip Journal on Advances in Signal Processing*, vol. 2016, pp. 1–19, 2016.

[31] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *13. ITG Fachtagung Sprachkommunikation (ITG 2018)*, Oct 2018.