Run-Time Adaptation of Neural Beamforming for Robust Speech Dereverberation and Denoising

Yoto Fujita[1], Aditya Arie Nugraha[2], Diego Di Carlo[2],
Yoshiaki Bando[3,2], Mathieu Fontaine[4,2], and Kazuyoshi Yoshii[5,2]

[1]Graduate School of Informatics, Kyoto University, Japan
[2]Center for Advanced Intelligence Project (AIP), RIKEN, Japan
[3]AIRC, National Institute of Advanced Industrial Science and Technology (AIST), Japan
[4]LTCI, Télécom Paris, France
[5]Graduate School of Engineering, Kyoto University, Japan

***Abstract*—This paper describes speech enhancement for real-time automatic speech recognition (ASR) in real environments. A standard approach to this task is to use neural beamforming that can work efficiently in an online manner. It estimates the masks of clean dry speech from a noisy echoic mixture spectrogram with a deep neural network (DNN) and then computes a enhancement filter used for beamforming. The performance of such a supervised approach, however, is drastically degraded under mismatched conditions. This calls for run-time adaptation of the DNN. Although the ground-truth speech spectrogram required for adaptation is not available at run time, blind dereverberation and separation methods such as weighted prediction error (WPE) and fast multichannel nonnegative matrix factorization (FastMNMF) can be used for generating pseudo ground-truth data from a mixture. Based on this idea, a prior work proposed a dual-process system based on a cascade of WPE and minimum variance distortionless response (MVDR) beamforming asynchronously fine-tuned by block-online FastMNMF. To integrate the dereverberation capability into neural beamforming and make it fine-tunable at run time, we propose to use weighted power minimization distortionless response (WPD) beamforming, a unified version of WPE and minimum power distortionless response (MPDR), whose joint dereverberation and denoising filter is estimated using a DNN. We evaluated the impact of run-time adaptation under various conditions with different numbers of speakers, reverberation times, and signal-to-noise ratios (SNRs).**

***Index Terms*—speech enhancement, dereverberation, neural beamforming, blind source separation,**

I. Introduction

Robust speech enhancement is a key technique in practical automatic speech recognition (ASR) systems that work in real time in real environments. For this purpose, one may use blind source separation (BSS) methods such as multichannel nonnegative matrix factorization (MNMF) [1], [2], independent low-rank matrix analysis (ILRMA) [3], and FastMNMF [4], [5]. Among these, FastMNMF is a state-of-the-art method that has been shown to outperform MNMF and ILRMA [5]. These methods are based on unsupervised learning (maximum likelihood estimation) of probabilistic models of mixture signals and are thus essentially free from the condition mismatch problem of supervised learning methods. However, these methods are hard to use for real-time systems due to the computationally demanding iterative optimization required at run time.

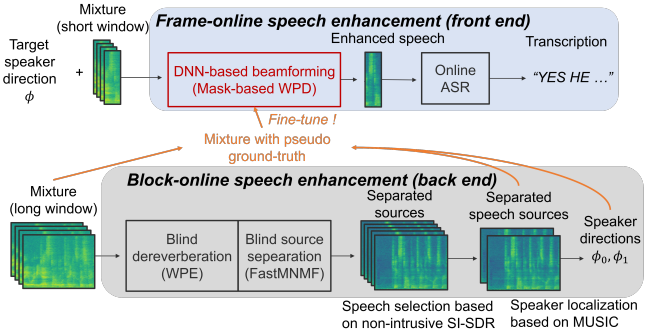

Fig. 1. The overview of the proposed joint adaptation of dereverberation and beamforming based on blind dereverberation by WPE and blind source separation by FastMNMF.

In recent years, deep neural networks (DNNs) have widely been used for speech enhancement. This approach, for example, performs direct mapping from a noisy mixture into multiple speech sources [6], mixture-conditioned deep speech generation [7], [8] for single-channel data, and neural beamforming with a DNN-based mask estimator [9] for multichannel data. In general, while supervised training of a DNN is computationally demanding, inference with the DNN works fast even on edge devices [10]. Considering the low-latency and high-performance nature and the potential contribution to the ASR [11], we focus on DNN-based beamforming (a.k.a. neural beamforming) as a front end of real-time distant ASR.

In neural beamforming, a DNN is used for estimating speech masks in the time-frequency domain. To separate clean speech from noisy speech mixture, the spatial covariance matrices (SCMs) of speech and noise are computed from the estimated masks and a enhancement filter used for beamforming is computed from the SCMs. The DNN is trained in a supervised manner by using pairs of noisy mixture and clean speech, often with the directions of the target speakers [10], [12]. However, it is not realistic to collect training data covering diverse acoustic environments that the model is potentially applied to. This makes the model less robust to unseen acoustic environments.

A promising solution to this problem is a run-time adaptation of neural beamforming [10], [13]. The biggest challenge in this task is that no ground-truth data (clean speech) are available unlike in standard offline benchmarks. To solve this problem, one can fine-tune a DNN-based mask estimator using *pseudo* ground-truth data given by FastMNMF. In a dual-process system [10], a light-weight minimum variance distortionless response (MVDR) beamforming (front end) is used for streaming speech enhancement, where the mask estimator is fine-tuned with the target speech separated by asynchronously-running FastMNMF (back end), often with the direction of the target speaker. It has been shown that the ASR performance tends to improve along with the amount of fine-tuning data (e.g., multi-party conversation data) [10]. This system also uses weighted prediction error (WPE) [14], [15], a popular blind dereverberation method, before MVDR beamforming and FastMNMF for improved ASR. However, since the adaptation is only applied to the mask estimator for beamforming, the adaptation capability of this system is thus limited to MVDR beamforming only.

In this paper, we propose run-time adaptation of neural beamforming for joint speech dereverberation and denoising. Since WPE works stably in various environments thanks to the unsupervised nature [10], we aim to draw its full potential with its neural extension. Specifically, we use weighted power minimization distortionless response (WPD) beamforming [16], [17], a unified version of WPE and a minimum power distortionless response (MPDR), where a DNN-based mask estimator is used to estimate a unified dereverberation and denoising filter. Both the speech dereverberation and denoising functions of the system can be adapted to a test environment while considering the mutual dependency of both tasks. We comprehensively investigate acoustic conditions in which the adaptation effectively contributes to the improvement of speech enhancement and ASR.

II. Related Work

This section reviews speech dereverberation based on WPE [14], [15], speech enhancement based on MPDR beamforming, joint dereverberation and denoising based on WPD beamforming [16], [17], and BSS based on FastMNMF [5].

A. Dereverberation

WPE is a well-known blind dereverberation method based on an autoregressive model of late reverberation. Let $\mathbf{x}_{ft} \in \mathbb{C}^M$ be the short-time Fourier transform (STFT) spectrum of an observed mixture captured by an M-channel microphone array at frequency $f \in [1, F]$ and time frame $t \in [1, T]$, where F is the number of frequency bins and T is the number of frames. We assume $\mathbf{x}_{ft}$ can be decomposed as follows:

$$\mathbf{x}_{ft} = \mathbf{d}_{ft} + \mathbf{r}_{ft}, \tag{1}$$

where $\mathbf{d}_{ft} \in \mathbb{C}^M$ is direct signals with early reflections, $\mathbf{r}_{ft} \in \mathbb{C}^M$ is the spectrum of late reverberation. The late reverberation is assumed to be the weighted sum of past observations as follows:

$$\mathbf{r}_{ft} = \sum_{\tau=b}^{L} \mathbf{W}_{f\tau}^{\mathsf{H}} \mathbf{x}_{f,t-\tau}, \tag{2}$$

where $\mathbf{W}_{f\tau} \in \mathbb{C}^{M\times M}$ is a mixing filter for delay τ, L is a tap length, and b is a prediction delay representing the boundary between the early reflections and late reverberation. The target $\mathbf{d}_{ft}$ is thus given by

$$\mathbf{d}_{ft} = \mathbf{x}_{ft} - \sum_{\tau=b}^{L} \mathbf{W}_{f\tau}^{\mathsf{H}} \mathbf{x}_{f,t-\tau}. \tag{3}$$

The filter is estimated by minimizing the weighted power of the estimated direct signal as follows:

$$\widehat{\mathbf{W}}_f = \underset{\mathbf{W}_f}{\operatorname{argmin}}\, \mathbb{E}_t \left[\frac{|\mathbf{x}_{ft} - \sum_{\tau=b}^{L} \mathbf{W}_{f\tau}^{\mathsf{H}} \mathbf{x}_{f,t-\tau}|^2}{\sigma_{ft}^2} \right], \tag{4}$$

where σ_{ft}^2 represents the time-varying power spectral density (PSD) of the target speech. The PSD can be obtained through iterative updates of the estimated target speech and its power [18], or by source mask estimation using a DNN [19].

B. Speech Enhancement

The multichannel signal model in (1) can be rewritten by decomposing a target speech signal as the product of a steering vector $\mathbf{a}_f \in \mathbb{C}^M$ and a source $s_{ft} \in \mathbb{C}$, while considering as noise other components including non-target components, early reflections, and late reverberations as follows:

$$\mathbf{x}_{ft} = \mathbf{a}_f s_{ft} + \mathbf{n}_{ft}, \tag{5}$$

where $\mathbf{n}_{ft}$ is the spectrum of noise. The target speech $\widehat{d}_{ft}$ is estimated by applying an enhancement filter $\mathbf{w}_{f0} \in \mathbb{C}^M$ to the mixture $\mathbf{x}_{ft}$ as follows:

$$\widehat{d}_{ft} = \mathbf{w}_{f0}^{\mathsf{H}} \mathbf{x}_{ft}. \tag{6}$$

In MPDR beamforming [20], the filter is estimated by minimizing the power of the observed mixture $\mathbf{x}_{ft}$, while maintaining a distortionless response in the direction of the steering vector $\mathbf{a}_f$ as follows:

$$\widehat{\mathbf{w}}_{f0}^{\text{MPDR}} = \underset{\mathbf{w}_{f0}}{\operatorname{argmin}}\, \mathbb{E}_t \left[|\mathbf{w}_{f0}^{\mathsf{H}} \mathbf{x}_{ft}|^2 \right] \quad \text{s.t.} \quad \mathbf{w}_{f0}^{\mathsf{H}} \mathbf{a}_f = 1. \tag{7}$$

The closed-form solution of the optimal filter is given by

$$\widehat{\mathbf{w}}_{f0}^{\text{MPDR}} = \frac{\mathbf{K}_f^{-1} \mathbf{a}_f}{\mathbf{a}^{\mathsf{H}} \mathbf{K}_f^{-1} \mathbf{a}_f}, \tag{8}$$

where $\mathbf{K}_f = \mathbb{E}_t[\mathbf{x}_{ft} \mathbf{x}_{ft}^{\mathsf{H}}]$ is the SCM of the mixture.

C. Joint Speech Dereverberation and Denoising

WPD beamforming is formulated by integrating WPE and MPDR beamforming for jointly dereverberation and enhancement. Specifiacally, using (3) and (6), the signal obtained with WPD beamforming is given by

$$\widehat{d}_{ft} = \mathbf{w}_{f0}^{\mathsf{H}} \left(\mathbf{x}_t + \sum_{\tau=b}^{L} \mathbf{W}_{f\tau}^{\mathsf{H}} \mathbf{x}_{f,t-\tau} \right) = \overline{\mathbf{w}}_f^{\mathsf{H}} \overline{\mathbf{x}}_{ft}, \tag{9}$$

where $\overline{\mathbf{w}}_f \in \mathbb{C}^{(L-b+1)M}$ is an integrated filter consisting of $\{\mathbf{w}_{ft}\}_{t=0,b,\ldots,L}$ and $\overline{\mathbf{x}}_{ft} \in \mathbb{C}^{(L-b+1)M}$ is the concatenation of the current and past observations $\{\mathbf{x}_{f,t-\tau}\}_{\tau=0,b,\ldots,L}$.

Using (4) and (7), the filter $\overline{\mathbf{w}}_f$ is estimated as

$$\widehat{\mathbf{w}}_f^{\text{WPD}} = \underset{\overline{\mathbf{w}}_f}{\operatorname{argmin}}\, \mathbb{E}_t \left[\frac{|\overline{\mathbf{w}}_f^{\mathsf{H}} \overline{\mathbf{x}}_{ft}|^2}{\sigma_{ft}^2} \right] \text{ s.t. } \mathbf{w}_{f0}^{\mathsf{H}} \mathbf{a}_f = 1. \quad (10)$$

The closed-form solution of the optimal filter is given by

$$\widehat{\mathbf{w}}_f^{\text{WPD}} = \frac{\overline{\mathbf{K}}_f^{-1} \overline{\mathbf{a}}_f}{\overline{\mathbf{a}}_f^{\mathsf{H}} \overline{\mathbf{K}}_f^{-1} \overline{\mathbf{a}}_f}, \quad (11)$$

where $\overline{\mathbf{K}}_f = \mathbb{E}_t[\overline{\mathbf{x}}_{ft} \overline{\mathbf{x}}_{ft}^{\mathsf{H}} \sigma_{ft}^{-2}]$ is the SCM of the mixture compensated by the PSD of the target speech, and $\overline{\mathbf{a}}_f \in \mathbb{C}^{(L-b+1)M}$ is the concatenation of the steering vector $\mathbf{a}_f$ and a zero vector $\mathbf{0} \in \mathbb{R}^{(L-b)M}$.

Using the SCM of the target speech $\overline{\mathbf{R}}_f = \overline{\mathbf{a}}_f \overline{\mathbf{a}}_f^{\mathsf{H}} |s_{ft}|^2$, (11) can be reformulated as:

$$\widehat{\mathbf{w}}_f^{\text{WPD}} = \frac{\overline{\mathbf{K}}_f^{-1} \overline{\mathbf{R}}_f}{\operatorname{tr}(\overline{\mathbf{K}}_f^{-1} \overline{\mathbf{R}}_f)} \overline{\mathbf{u}}_q, \quad (12)$$

where $\overline{\mathbf{u}}_q \in \mathbb{R}^{(L-b+1)M}$ is a one-hot vector whose q-th element (reference channel) takes one and zero otherwise.

In mask-based WPD beamforming, the PSD σ_{ft}^2 and SCM $\overline{\mathbf{R}}_f$ of the target speech are computed with $\widehat{\mathbf{s}}_{ft} = \omega_{ft} \mathbf{x}_{ft}$, where a time-frequency (TF) speech mask $\omega_{ft} \in [0, 1]$ is estimated by a DNN [21], [22].

D. Blind Source Separation

The general goal of BSS is to separate a mixture spectrogram $\{\mathbf{x}_{ft}\}_{f,t=1}^{F,T}$ into N source spectrograms $\{\{\mathbf{x}_{nft}\}_{f,t=1}^{F,T}\}_{n=1}^{N}$, where $\mathbf{x}_{ft}, \mathbf{x}_{nft} \in \mathbb{C}^M$. In modern BSS methods, each source $\mathbf{x}_{nft}$ is typically assumed to follow an M-variate circularly-symmetric complex Gaussian distribution as follows:

$$\mathbf{x}_{nft} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \lambda_{nft} \mathbf{G}_{nf}), \quad (13)$$

where λ_{nft} and $\mathbf{G}_{nft}$ are the PSD and SCM of the source n. Assuming the additivity of complex spectrograms and using the reproductive property of the Gaussian distribution, the mixture $\mathbf{x}_{ft}$ is given by

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}, \sum_{n=1}^{N} \lambda_{nft} \mathbf{G}_{nf} \right). \quad (14)$$

In MNMF [1], [2] and its variants including FastMNMF [5], the PSDs $\{\lambda_{nft}\}_{f,t=1}^{F,T}$ are factorized with NMF as follows:

$$\lambda_{nft} = \sum_{k=1}^{K} u_{nkf} v_{nkt}, \quad (15)$$

where $\mathbf{u}_{nk} \in \mathbb{R}_+^F$ is a basis vector, $\mathbf{v}_{nk} \in \mathbb{R}_+^T$ is an activation vector, and K is the number of bases. In FastMNMF, the SCM $\mathbf{G}_{nf}$ is also factorized as follows:

$$\mathbf{G}_{nf} = \mathbf{Q}_f^{-1} \operatorname{Diag}(\tilde{\mathbf{g}}_n) \mathbf{Q}_f^{-\mathsf{H}}, \quad (16)$$

where $\mathbf{Q}_f \in \mathbb{C}^{M \times M}$ is a time-invariant diagonalization matrix and $\tilde{\mathbf{g}}_n \in \mathbb{R}_+^M$ is a frequency-invariant nonnegative vector.

The model parameters $\{\mathbf{u}_{nk}\}_{n,k=1}^{N,K}$, $\{\mathbf{v}_{nk}\}_{n,k=1}^{N,K}$, $\{\mathbf{Q}_f\}_{f=1}^{F}$, and $\{\tilde{\mathbf{g}}_n\}_{n=1}^{N}$ are estimated with an iterative optimization algorithm such that the likelihood of the parameters for the mixture given by (14) is maximized [5]. Given the optimal parameters, the separation filter $\mathbf{w}_{nft}^{\text{BSS}}$ is given by

$$\widehat{\mathbf{w}}_{nft}^{\text{BSS}} = \mathbf{Q}_f^{\mathsf{H}} \operatorname{Diag} \left(\frac{\lambda_{nft} \tilde{\mathbf{g}}_n}{\sum_{n'} \lambda_{n'ft} \tilde{\mathbf{g}}_{n'}} \right) \mathbf{Q}_f^{-\mathsf{H}} \mathbf{u}_q, \quad (17)$$

where $\mathbf{u}_q \in \{0, 1\}^M$ is a one-hot vector whose q-th element (reference channel) takes one and zero otherwise.

III. Proposed method

This section describes the proposed adaptive joint dereverberation and denoising system based on a dual-process architecture consisting of mask-based WPD beamforming with FastMNMF-guided fine-tuning. This system uses the WPD beamforming to perform low-latency speech dereverberation and denoising, resulting in a single-channel speech signal useful for the ASR system. To be adaptive to dynamic environments, its DNN-based mask estimator is fine-tuned at run time using speech signals dereverberated and separated with high-latency yet environment-robust WPE and FastMNMF.

A. Joint Neural Speech Dereverberation and Denoising

Given a mixture spectrogram $\mathbf{X} \triangleq \{\mathbf{x}_{ft} \in \mathbb{C}^M\}_{f=1,t=1}^{F,T}$ with target speaker DOAs $\boldsymbol{\phi} \triangleq \{\phi_t \in [0, 2\pi]\}_{t=1}^{T}$, a DNN $\mathcal{F}_{\boldsymbol{\Theta}}$ parameterized by $\boldsymbol{\Theta}$ is used for estimating TF masks $\boldsymbol{\omega} \triangleq \{\omega_{ft}\}_{f=1,t=1}^{F,T}$ as follows:

$$\boldsymbol{\omega} = \mathcal{F}_{\boldsymbol{\Theta}}(\mathbf{X}, \boldsymbol{\phi}). \quad (18)$$

The PSD $\widehat{\sigma}_{ft}^2$ and SCM $\widehat{\mathbf{R}}_f$ of the target speech can be computed using the mask estimate as follows:

$$\widehat{\sigma}_{ft}^2 = \frac{1}{M} \sum_{m=1}^{M} |\omega_{ft} x_{ftm}|^2, \quad (19)$$

$$\widehat{\mathbf{R}}_f = \frac{1}{T} \sum_{t=1}^{T} \widehat{\mathbf{s}}_{ft} \widehat{\mathbf{s}}_{ft}, \quad (20)$$

where $\widehat{\mathbf{s}}_{ft} = [\omega_{ft} \mathbf{x}_{ft}^{\top}, \omega_{f,t-b} \mathbf{x}_{f,t-b}^{\top}, \ldots, \omega_{f,t-L} \mathbf{x}_{f,t-L}^{\top}]^{\top} \in \mathbb{C}^{(L-b+1)M}$. The WPD filter $\widehat{\mathbf{w}}_f^{\text{WPD}}$ is then computed from $\widehat{\mathbf{K}}_f = \sum_t \overline{\mathbf{x}}_{ft} \overline{\mathbf{x}}_{ft}^{\mathsf{H}} \widehat{\sigma}_{ft}^{-2}$ and $\widehat{\mathbf{R}}_f$ as in (12). Finally, the target speech signal $\widehat{d}_{ft}$ corresponding to the DOA ϕ_t is obtained by applying $\widehat{\mathbf{w}}_f^{\text{WPD}}$ to $\overline{\mathbf{x}}_{ft}$ as in (9).

B. Pretraining of Mask Estimator

The DNN-based mask estimator $\mathcal{F}_{\boldsymbol{\Theta}}$ is pretrained using triples consisting of an M-channel mixture, a reference speech, and a target DOA. It is optimized to minimize the negative signal-to-distortion ratio (SDR) between the estimated time-domain speech signal $\widehat{\mathbf{d}} \in \mathbb{R}^S$ and the reference time-domain speech signal $\mathbf{d}^{\text{ref}} \in \mathbb{R}^S$ given by

$$\mathcal{L} = -10 \log_{10} \frac{\widehat{\mathbf{d}}^{\top} \widehat{\mathbf{d}}}{(\widehat{\mathbf{d}} - \mathbf{d}^{\text{ref}})^{\top} (\widehat{\mathbf{d}} - \mathbf{d}^{\text{ref}})}, \quad (21)$$

where S denotes the number of samples. The estimated time-domain speech signal $\widehat{\mathbf{d}}$ is obtained by applying the inverse STFT to the estimated speech signal in the STFT domain $\{\widehat{d}_{ft}\}_{f=1,t=1}^{F,T}$.

C. Run-Time Adaptation of Mask Estimator

To make the mask estimator $\mathcal{F}_{\mathbf{\Theta}}$ adaptive to environmental changes, we fine-tune it at run time using triples of the observed mixture $\{\mathbf{x}_{ft}\}_{f=1,t=1}^{F',T'}$, the pseudo ground-truth speech signal $\mathbf{d}_p^{\text{ref}} \in \mathbb{R}^S$, and the pseudo ground-truth DOA ϕ_p to minimize the negative SDR loss in (21), where F' and T' are the number of frequency bins and that of frames of the mixture used for fine-tuning, respectively.

To obtain the pseudo ground-truth speech signal, we first dereverberate the mixture $\mathbf{x}_{ft}$ using WPE as follows:

$$\mathbf{x}_{ft}^{\text{dry}} = \mathbf{x}_{ft} - \sum_{\tau=b}^{L} \mathbf{W}_{f\tau}^{\mathsf{H}} \mathbf{x}_{f,t-\tau}. \tag{22}$$

where $\mathbf{x}_{ft}^{\text{dry}}$ is the dereverberated mixture and $\mathbf{W}_{f\tau}$ is a filter for delay τ. The filter is estimated as (4) with the PSD σ_{ft}^2 obtained through iterative updates of the estimated dereverberated signal $\mathbf{x}_{ft}^{\text{dry}}$ and its power σ_{ft}^2 [18]. Then, we separate the sources $\{\{x_{nft}\}_{f,t=1}^{F',T'}\}_{n=1}^{N}$ from the dereverberated mixture using FastMNMF as follows:

$$x_{nft} = \left(\widehat{\mathbf{w}}_{nft}^{\text{BSS}}\right)^{\mathsf{H}} \mathbf{x}_{ft}^{\text{dry}}, \tag{23}$$

where the separation filter $\mathbf{w}_{nft}^{\text{BSS}}$ is calculated as in (17). We measure the signal quality of each separated source signal using the reference-less non-intrusive scale-invariant SDR [23], [24]. We take $N' (\leq N)$ separated signals that satisfy a predefined threshold α and consider these as pseudo ground-truth speech signals $\{\mathbf{d}_{p,n}^{\text{ref}}\}_{n=1}^{N'}$. Finally, the corresponding N' pseudo ground-truth DOAs $\{\phi_{p,n}\}_{n=1}^{N'}$ are estimated based on multiple signal classification (MUSIC) [25].

IV. Evaluation

We report a comprehensive evaluation conducted for assessing the performance of our adaptive system in various simulated acoustic environments.

A. Dataset

For the mask estimator $\mathcal{F}_{\mathbf{\Theta}}$, we made a *training dataset* comprising 36,000 triples and a *validation dataset* comprising 3,600 triples. Each triple consisted of a 2-second 7-channel mixture signal, a 2-second 7-channel target speech signal, and the corresponding target DOA. Each mixture was composed of two speech signals by different speakers, who were stationary during the recording, and a diffuse noise signal. The speech signals were randomly taken from the training set (for the *training dataset*) and the development set (for the *validation dataset*) of Librispeech [26], while the diffuse or moving noise signals were randomly taken from the DEMAND dataset [27].

Both mixture and target speech signals were simulated using Pyroomacoustics [28] by considering a 7-channel circular microphone array, configured to match the geometry constraints of the DEMAND dataset, within a 2-dimensional room. The room length and width were randomly sampled from uniform distributions $\mathcal{U}(7.6\text{m}, 8.4\text{m})$ and $\mathcal{U}(5.6\text{m}, 6.4\text{m})$, respectively. The 2-dimensional coordinates of the array center were sampled from $\mathcal{U}(3.6\text{m}, 4.4\text{m})$ and $\mathcal{U}(2.6\text{m}, 3.4\text{m})$, respectively. The distance between the array center and each speaker was sampled from $\mathcal{U}(1\text{m}, 2\text{m})$. The reverberation time (RT60) for the mixtures was varied between 0.25 and 0.7 seconds. The target speech signals, which were supposed to include the direct path and early component, were simulated in the same rooms as the mixtures, but with the RT60 fixed at 0.25 seconds. Indoor noise signals from the DEMAND dataset, excluding the environment "PSTATION" that was used for the test set (see below), were randomly selected and added to the simulated mixtures, with an SNR between -5.0 and 5.0 dB.

To evaluate the effectiveness of our adaptation method in various acoustic environments, we prepared a *test dataset* consisting of seven distinct simulation settings, i.e., one "default" setting and six variations. The default setting considered mixtures of two speakers with an RT60 of 0.5 seconds, an SNR of 30.0 dB, and other parameters were randomly sampled as in the pretraining dataset. For the other six settings, we independently varied the number of stationary speakers (3 and 4), the RT60s (0.8 and 1.2 seconds), and the SNR (5.0 and -5.0 dB). Speech signals were taken from the test set of Librispeech, with diffuse or moving noise signals from the environment "PSTATION" of the DEMAND dataset. For each setting, we generated 30 triples, each consisting of an approximately 8-minute mixture, the target speech signal, the target DOA, and corresponding transcriptions. The first four minutes of each recording were used for fine-tuning the mask estimator, with the remaining duration reserved for evaluation.

B. Experimental Settings

Both the front and back ends operated in the STFT domain. The STFT coefficients were computed using a window size of 1024 ($F = 513$) with a hop length of 256.

For the mask-based WPD beamforming, the prediction delay and the tap size of the convolutional filter were set to $b = 3$ and $L = 8$, respectively. The TF mask was estimated using a DNN as in [10]. The DNN was composed of a preprocessing network, a direction attractor network (DAN), and a bidirectional long short-term memory (BLSTM) network. The preprocessing network took as input the concatenation of the log magnitude of the mixture, the inter-channel phase difference, and the beamforming output by delay-sum beamforming, while the DAN took the target DOA as input. Given the outputs of these two networks, the BLSTM then estimates a TF mask. This mask estimator was pretrained on the *training dataset* using the AdamW optimizer with a learning rate of 10^{-4} and a batch size of 4. The model was trained for 20 epochs, and the model with the lowest validation negative SDR loss was selected for evaluation.

For the joint adaptation of mask-based WPD beamforming, the prediction delay, the tap size, and the number of iterations for parameter updates in WPE were set to $b = 3$ and $L = 13$, and 3, respectively. The number of sources, the number of bases, and the number of iterations for parameter updates in FastMNMF were set to $N = 5$, $K = 16$, and 200, respectively. The threshold for non-intrusive SI-SDR was set to $\alpha = 10.0$. The window size of the mixture given as the input to FastMNMF was set to 30 seconds. The learning rate was set to 4×10^{-5},

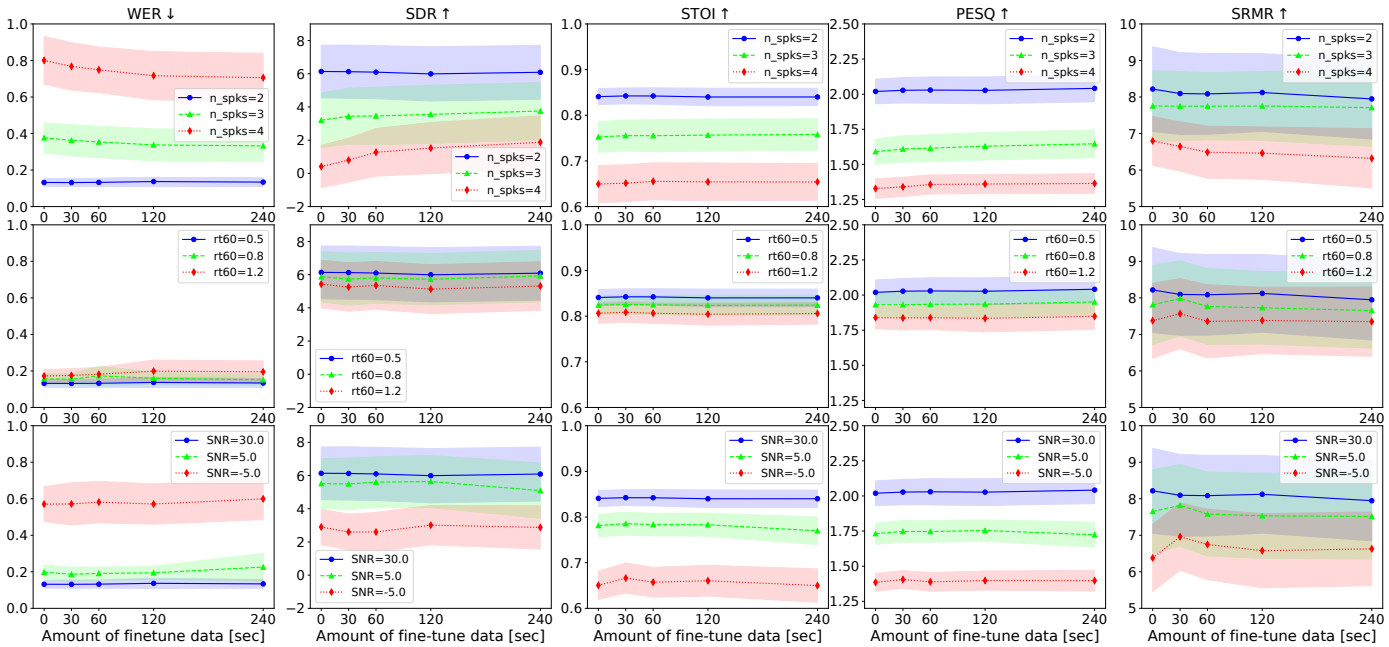

Fig. 2. The evaluation result of the joint adaptation of dereverberation and denoising by mask-based WPD beamforming based on blind dereverberation by WPE and BSS by FastMNMF. Blue lines refer to the "default" setting, constant across different plots. The shaded regions surrounding each line indicate the 95% confidence intervals.

and the batch size was 4 for fine-tuning. To stabilize the fine-tuning, we added the same amount of the pretraining data to the fine-tuning data so a batch may contain these two types of data.

The evaluation compared different amounts of fine-tuning data (30, 60, 120, and 240 seconds) against several key metrics. These metrics include word error rate (WER), signal-to-distortion ratio (SDR), short-time objective intelligibility (STOI), perceptual evaluation of speech quality (PESQ), and speech-to-reverberation modulation energy ratio (SRMR). We used a Transformer-based encoder-decoder ASR model from the SpeechBrain toolkit [29] to measure WER. This ASR model was trained on the Librispeech dataset. For all metrics except WER, higher values are better.

C. Experimental Results

Figure 2 shows the evaluation results of the joint adaptation of dereverberation and denoising using mask-based WPD beamforming with different durations of fine-tuning data and various simulation settings. The upper left plot in the figure indicates that our adaptation method improved WER, SDR, STOI, and PESQ across different numbers of stationary speakers, which proves the effectiveness of our adaptation. These improvements seem to benefit from the robust separation capability of FastMNMF for stationary sources.

However, when we used a large ammount of fine-tuning data, the WER was slightly degraded as the RT60 increased or the SNR decreased. This would be because the pretrained mask-based WPD beamforming already has a strong capability of dereverberation and FastMNMF is less robust to a mixture with moving sources. In contrast, when we used a small ammount of fine-tuning data, the ASR performance hardly changed and the other metrics, STOI, PESQ, and SRMR improved in noisy or reverberant conditions. This suggests that 30 seconds of fine-tuning data is optimal for practical use.

V. Conclusion

This paper proposes a run-time adaptation method for the joint neural dereverberation and denoising with mask-based WPD beamforming using fine-tuning data obtained using WPE and FastMNMF. Evaluations showed robust improvements of the ASR performance across different numbers of stationary speakers, RT60s, and SNRs when we used a small ammount of fine-tuning data. For future work, BSS methods that are capable of dealing with more various acoustic conditions should be investigated to improve the ASR performance even in noisy conditions with moving sources.

Acknowledgment

This work was supported by JST PRESTO no. JPMJPR20CB, JSPS KAKENHI nos. JP20H00602, JP21H03572, JP23K16912, and JP23K16913, ANR Project SAROUMANE (ANR-22-CE23-0011), and Hi! Paris Project MASTER-AI.

References

[1] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 550–563, 2010.

[2] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 21, no. 5, pp. 971–982, 2013.

[3] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[4] N. Ito and T. Nakatani, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 371–375.

[5] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2610–2625, 2020.

[6] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[7] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[8] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7402–7406.

[9] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Comput. Speech Lang.*, vol. 46, pp. 374–385, 2017.

[10] K. Sekiguchi, A. A. Nugraha, Y. Du, Y. Bando, M. Fontaine, and K. Yoshii, "Direction-aware adaptive online neural speech enhancement with an augmented reality headset in real noisy conversational environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 9266–9273.

[11] CHiME-4 Challenge Organizers, *CHiME-4 results*, [Online] Available: https://www.chimechallenge.org/challenges/chime4/results, [Accessed Jul. 6, 2024], 2016.

[12] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *Proc. IEEE Spoken Language Technol. Workshop*, 2018, pp. 558–565.

[13] T. Aizawa, Y. Bando, K. Itoyama, K. Nishida, K. Nakadai, and M. Onishi, "Unsupervised domain adaptation of universal source separation based on neural full-rank spatial covariance analysis," in *Proc. IEEE Int. Workshop Mach. Learn. Signal. Process.*, 2023, pp. 1–6.

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

[15] T. Yoshioka and T. Nakatani, "Generalization of multichannel linear prediction methods for blind MIMO impulse response shortening," *IEEE Audio, Speech, Language Process.*, vol. 20, no. 10, pp. 2707–2720, 2012.

[16] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 903–907, 2019.

[17] T. Nakatani and K. Kinoshita, "Simultaneous denoising and dereverberation for low-latency applications using frame-by-frame online unified convolutional beamformer," in *Proc. INTERSPEECH*, 2019, pp. 111–115.

[18] L. Drude, C. Boeddeker, J. Heymann, *et al.*, "Integrating neural network based beamforming and weighted prediction error dereverberation.," in *Proc. INTERSPEECH*, 2018, pp. 3043–3047.

[19] M. Togami, "Joint training of deep neural networks for multi-channel dereverberation and speech source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 3032–3036.

[20] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, 2018.

[21] T. Nakatani, R. Takahashi, T. Ochiai, *et al.*, "DNN-supported mask-based convolutional beamforming for simultaneous denoising, dereverberation, and source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6399–6403.

[22] W. Zhang, A. S. Subramanian, X. Chang, S. Watanabe, and Y. Qian, "End-to-end far-field speech recognition with unified dereverberation and beamforming," *Proc. INTERSPEECH*, pp. 324–328, 2020.

[23] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 626–630.

[24] A. Kumar, K. Tan, Z. Ni, *et al.*, "Torchaudio-Squim: Reference-less speech quality and intelligibility measures in torchaudio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[25] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.

[27] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," *Proc. Mtgs. Acoust.*, vol. 19, no. 1, pp. 1–6, 2013.

[28] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 351–355.

[29] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, "SpeechBrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.