Rethinking Processing Distortions: Disentangling the Impact of Speech Enhancement Errors on Speech Recognition Performance

Tsubasa Ochiai *Member, IEEE*, Kazuma Iwamoto,
Marc Delcroix *Senior Member, IEEE*, Rintaro Ikeshita *Member, IEEE*,
Hiroshi Sato *Member, IEEE*, Shoko Araki *Fellow, IEEE*, Shigeru Katagiri *Fellow, IEEE*

***Abstract*—It is challenging to improve automatic speech recognition (ASR) performance in noisy conditions with a single-channel speech enhancement (SE) front-end. This is generally attributed to the processing distortions caused by the nonlinear processing of single-channel SE front-ends. However, the causes of such degraded ASR performance have not been fully investigated. How to design single-channel SE front-ends in a way that significantly improves ASR performance remains an open research question. In this study, we investigate a signal-level numerical metric that can explain the cause of degradation in ASR performance. To this end, we propose a novel analysis scheme based on the orthogonal projection-based decomposition of SE errors. This scheme manually modifies the ratio of the decomposed interference, noise, and artifact errors, and it enables us to directly evaluate the impact of each error type on ASR performance. Our analysis reveals the particularly detrimental effect of artifact errors on ASR performance compared to the other types of errors. This provides us with a more principled definition of processing distortions that cause the ASR performance degradation. Then, we study two practical approaches for reducing the impact of artifact errors. First, we prove that the simple observation adding (OA) post-processing (i.e., interpolating the enhanced and observed signals) can monotonically improve the signal-to-artifact ratio. Second, we propose a novel training objective, called artifact-boosted signal-to-distortion ratio (AB-SDR), which forces the model to estimate the enhanced signals with fewer artifact errors. Through experiments, we confirm that both the OA and AB-SDR approaches are effective in decreasing artifact errors caused by single-channel SE front-ends, allowing them to significantly improve ASR performance.**

***Index Terms*—single-channel speech enhancement, noise-robust speech recognition, processing distortion**

I. Introduction

Over the past decade, the performance of automatic speech recognition (ASR) systems has progressively improved [1]–[3] with the advent of deep learning techniques [4], [5]. Research interest has thus moved toward achieving robustness against acoustic interference such as background noise, reverberation, and overlapping speakers [6]. Recent studies and ASR challenges [7], [8] have reported that multichannel array processing, i.e., mask-based beamformer [9]–[12], plays a key role in improving robustness to interference, and it has been used as the de facto standard front-end for noise-robust ASR systems. However, it is not always possible to equip commercial devices with microphone arrays due to structural constraints and cost restrictions. Therefore, building noise-robust ASR systems with a single microphone (i.e., single-channel) is an essential research issue.

Recent advances in deep learning have greatly improved the performance of single-channel speech enhancement (SE), such as noise reduction [13], [14], source separation [13], [15], and target speech extraction [16]. Single-channel SE approaches can improve SE performance measures, such as signal-to-distortion ratio (SDR) [17], perceptual evaluation of speech quality (PESQ) [18], and short-time objective intelligibility (STOI) [19]. However, previous studies (e.g., [20]–[24]) have reported that such single-channel SE approaches do not necessarily contribute to improving the ASR performance for the ASR systems trained with multi-condition training data [25] (i.e., they tend to only slightly improve or even degrade word error rate (WER)).

A common view in the ASR research community is that *processing distortions* produced by single-channel nonlinear processing, like neural networks, induce such ASR performance degradation. Researchers have focused on developing approaches to mitigate the effects of such potential distortions. Such approaches include retraining the ASR back-end on enhanced speech (e.g., [21]) or joint training of the SE front-end and ASR back-end (e.g., [23], [24], [26]–[28]). However, it is not always possible to tune the ASR back-end for a specific SE front-end, due to the need to use an already deployed ASR back-end and the cost of training and maintaining an ASR system for different SE front-ends, etc. Therefore, modular ASR systems are often preferable in practice, and such systems require the design of effective standalone SE front-ends.

How to design single-channel SE front-ends that can significantly improve ASR performance remains an open research question. We believe that a key to answering this question is a better understanding of the nature of processing distortions that are detrimental to ASR. Surprisingly, there have been no detailed systematic analyses or interpretations of the processing distortions and their impact on ASR. We aim to fill this gap in this study. In particular, we investigate a signal-level numerical metric that could explain, at least partly, the cause

Tsubasa Ochiai, Marc Delcroix, Rintaro Ikeshita, Hiroshi Sato, and Shoko Araki are with NTT Corporation, Chiyoda-ku 100-8116, Japan (email: tsubasa.ochiai@ntt.com; marc.delcroix@ntt.com; rintaro.ikeshita@ntt.com; hrs.sato@ntt.com; shoko.araki@ntt.com).

Kazuma Iwamoto and Shigeru Katagiri are with Doshisha University, Kyotanabe 610-0394, Japan (kazuma.iwamoto1204@gmail.com; skatagir@mail.doshisha.ac.jp).

of ASR performance degradation.

A. Hypotheses on cause of ASR performance degradation

An SE model estimates the target clean source given the observed noisy signal. The processing distortions can be measured as the total estimation errors between the reference source signal and the enhanced signal. We hypothesize that the total estimation errors can be decomposed into *natural* and *unnatural* error components. Natural errors would consist of residual interference and noise signals, which are similar to noisy samples seen when training the ASR back-end with multi-condition data. Consequently, such natural (residual) errors may have little impact on ASR performance. In contrast, unnatural errors consist of processing artifacts caused by single-channel nonlinear processing. Such unnatural (processing) errors may have a more detrimental effect on ASR because it is difficult to train an ASR back-end that is robust to these artifacts due to their large diversity. A signal-level numerical metric is required to separate the total estimation error into natural and unnatural error components.

To define the error components, we borrow the orthogonal projection-based decomposition of the SE estimation errors, which was previously proposed as an SE performance metric for speech denoising and separation, such as the SDR of BSSEval [17]. It decomposes the SE errors into three error components, i.e., the residual *interference* error, the residual *noise* error, and the *artifact* error. The interference error component consists of a linear combination of target speech and interference signals, while the noise error component consists of a linear combination of target speech, interference, and noise signals. Accordingly, the interference and noise errors represent the signals that could be observed in the real world, i.e., natural signals. On the other hand, the artifact error component consists of a signal orthogonal to the target speech, interference, and noise signals (Figure 1-(a)); in other words, it cannot be represented by a linear combination of target speech, interference, and noise signals. Thus, the artifact errors represent the signals that are produced by nonlinear processing with single-channel SE, which can be considered unnatural (artificial) signals. Considering the above observations, we *hypothesize* that the artifact errors have a larger impact on the degradation of ASR performance than on the other residual interference and noise errors. In other words, we assume that the artifact errors could mostly explain the limited ASR performance improvement induced by single-channel SE.

B. Contributions: Experimental proof of impact of artifact errors and practical approaches to mitigate them

In this study, to confirm this hypothesis, we propose a novel analysis scheme that decomposes the SE errors by utilizing the reference signals (i.e., target speech, interfering speech, an background noise), manually modifies the ratio of the interference, noise, and artifact errors included in the estimated signals, and directly evaluates the ASR performance of the modified signals. We refer to the proposed analysis scheme as direct scaling analysis (DSA). This analysis enables us to directly investigate the impact of each SE error on the ASR performance and proves the particularly negative impact of artifact errors on ASR performance as well as the relatively limited impact of the interference and noise errors.

The results of this analysis motivate us to explore two practical approaches to mitigate the effect of the artifact errors generated by the single-channel SE for improving the ASR performance: 1) observation adding (OA) and 2) artifact-boosted signal-to-distortion ratio (AB-SDR) loss.

The OA post-processing interpolates the enhanced signal and the observed noisy signal, and it is applied in the inference stage as the post-processing of the single-channel SE. In the literature, it has often been used to reduce the auditory nonlinear distortion (i.e., musical noise [29]) of the enhanced signals generated by classical spectral subtraction [30] and binary masking [31]. However, to the best of our knowledge, OA has not been investigated as an approach to improving the ASR performance of recent single-channel SE systems. Moreover, its effect on decomposed SE errors has not been considered. In this paper, we mathematically and experimentally show that the OA post-processing has an effect on reducing the ratio of artifact errors, which contributes to improving ASR performance based on our hypothesis.

We also propose a novel training objective for an SE front-end, i.e., AB-SDR loss, which extends the conventional SDR loss to put more weight on the artifact errors than on the noise and interfering errors. Unlike the OA post-processing, AB-SDR loss is applied in the training stage as the training objective. We experimentally show that AB-SDR successfully reduces the ratio of artifact errors and improves the ASR performance of single-channel SE.

Although the two approaches try to reduce the artifact errors in different ways, we show experimentally that both of them successfully reduce artifact errors at the cost of increasing other types of errors less harmful to ASR and consequently improve ASR performance. These experiments also provide experimental evidence of our hypothesis that the (unnatural) artifact errors have a larger impact on the degradation of ASR performance compared to the other (natural) residual interference and noise errors.

The main contributions of this paper can be summarized as follows:

1) We propose a novel analysis scheme (i.e., DSA) to confirm the validity of our hypothesis that the artifact errors should be one of the reasonable signal-level numerical metrics representing the cause of ASR performance degradation, which has not been explored in the noise-robust ASR community.
2) The experimental analyses provide a novel insight into the causes of the limited ASR performance improvement induced by single-channel SE: 1) artifact errors have a larger impact on ASR than interference and noise errors and 2) ASR performances of single-channel SE could be improved by mitigating artifact errors.
3) We propose two practical approaches to mitigate the impact of the artifact errors: 1) the OA post-processing and 2) the AB-SDR training objective. The experimental results show the effectiveness of these approaches in

improving the ASR performance of single-channel SE, even for real recordings.

This paper extends our previous study [32] by 1) generalizing the derivation and proof from a single-talker to a multi-talker scenario, 2) proposing the AB-SDR training objective as a more straightforward approach to reducing the ratio of artifact errors, and 3) performing comprehensive experiments including different acoustic conditions, evaluation corpora, and SE/ASR systems.

The remainder of this paper is summarized as follows. In Section II, we briefly discuss related prior works. Section III overviews the orthogonal projection-based decomposition of the SE errors. Section IV explains the details of 1) the proposed DSA scheme, 2) the OA approach, and 3) the AB-SDR approach. In Sections V–VIII, we explain the experiments of the single-channel SE/ASR setups in noisy environments and give novel insights into the processing distortions. Finally, we conclude this paper in Section IX.

II. Related works

A. Observation adding

Concurrently with our preliminary study [32], the authors of another study [33] proposed using the OA post-processing to alleviate the processing distortions introduced by an SE front-end and reported an improvement in ASR performance. However, they did not analyze how OA post-processing could improve ASR performance.

Several studies [34]–[36] have investigated approaches to reducing the effect of processing distortion on noisy multi-speaker mixtures. They hypothesized it would be better to recognize noisy speech directly when no interference speaker was active and enhanced speech otherwise. They proposed switching the input of the ASR back-end between the enhanced and observed signals according to the interference and noise conditions. As a formalization, they adopted the weighted sum of the enhanced and observed signals, which is similar to that of the OA post-processing, where the switching weights are dynamically computed based on the overlap detector or switcher. They showed the effectiveness of such switching schemes on ASR performance. However, the effectiveness of the OA post-processing alone (i.e., simply interpolating the enhanced and observed signals with a pre-defined interpolation weight) on ASR performance has not been fully proven.

In contrast to these prior works, this study offers a hypothesis on the mathematical definition for the cause of ASR performance degradation and experimentally confirms the validity of our hypothesis with the proposed novel analysis scheme.

B. Artifact-aware training objective

The authors of a previous study [37] proposed using the SDR of BSSEval [17] as the training objective for neural network-based SE models. Because the SDR loss minimizes the total estimation errors (i.e., the sum of interference, noise, and artifact errors), it does not allow balancing the contribution of each error type within the loss.

Several studies [38], [39] have investigated adding the signal-to-artifact ratio (SAR) [17] to their training objective. The former study [38] focused on evaluation in terms of subjective listening quality but did not investigate the impact of artifact errors on ASR performance. The latter study [39] investigated the joint training framework with the ASR-level training objective and adopted the SAR as an additional regularization term. However, the effectiveness of the SAR loss alone has not been confirmed, and thus it has not explicitly shown how the SAR loss without the ASR-level loss contributes to improving ASR performance. In particular, in our preliminary experiment, we observed that the SAR-based loss may cause training instabilities probably because the SAR becomes optimal when the estimated signal matches the observed noisy signal. This may indicate that it is challenging to train an SE system without the ASR-level loss.

In contrast to these prior works, this study proposes a novel analysis scheme and, for the first time, explicitly shows that artifact errors negatively impact ASR performance. Furthermore, rather than adding the SAR-based regularization, we introduce the AB-SDR loss, which is a signal-level training objective that boosts the effect of the artifact errors in the SDR loss and can improve ASR performance without relying on the ASR-level training objective.

C. ASR-level training objectives

As described in Section I, the joint training of the SE front-end and ASR back-end (e.g., optimizing the SE front-end with the ASR-level training objective) would be a straightforward approach to mitigating the processing distortion problem, which could be done without deeply understanding its mechanism. Although such a joint training scheme successfully improves ASR performance (e.g., [23], [24], [26]–[28]), it requires modifying the SE front-end and/or the ASR back-end, which is not always an option in practice, e.g., due to the need to use an already deployed system. Moreover, its effect on SE estimation errors has not been fully explored; the previous studies [26], [27] confirmed the improvements in the ASR metric but not in SE metrics such as SDR.

III. Background

A. Overview of speech enhancement task

In this study, we focus on the single-channel SE task, where the observed signals are contaminated by the interfering speech and background noise signals. Let $\mathbf{y} \in \mathbb{R}^T$ denote a vector of a T-length time-domain waveform of the observed signals. The observed noisy signals can be modeled as:

$$\mathbf{y} = \mathbf{s} + \mathbf{i} + \mathbf{n}, \tag{1}$$

where, $\mathbf{s} \in \mathbb{R}^T$ denotes the target speaker's speech signal, $\mathbf{i} \in \mathbb{R}^T$ denotes the interfering speaker's speech signal, and $\mathbf{n} \in \mathbb{R}^T$ denotes the background noise signal. In this paper, to simplify the description, we describe the equations assuming a single interfering speaker.

The goal of the SE task is to extract the target speaker's speech signals from the observed noisy signals, i.e., to reduce the interfering speech and background noise signals. Given

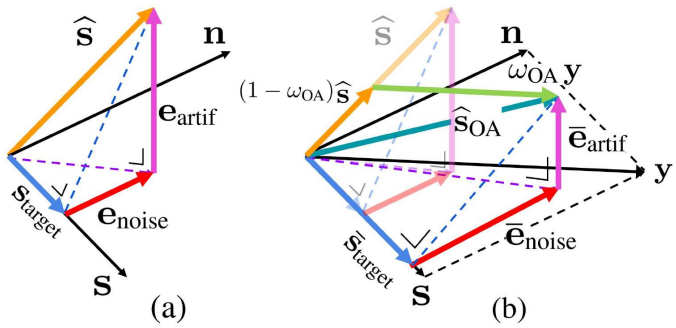

Fig. 1. Illustrations of (a) orthogonal projection-based decomposition and (b) effect of observation adding from viewpoint of orthogonal projection-based decomposition.

the observed signal $\mathbf{y}$ as an input, the SE system estimates the target signal $\widehat{\mathbf{s}} \in \mathbb{R}^T$ as:

$$\widehat{\mathbf{s}} = \mathrm{SE}(\mathbf{y}), \tag{2}$$

where $\mathrm{SE}(\cdot)$ denotes the functional representation of SE processing.

One of the major uses of an SE system is as a front-end for ASR to improve performance in noisy conditions. Given the enhanced signal as an input, the ASR system estimates the word sequence $\widehat{\mathbf{W}} \in \mathcal{V}^N$, where $\mathcal{V}$ denotes a set of symbols and N denotes the sequence length, as:

$$\widehat{\mathbf{W}} = \mathrm{ASR}(\widehat{\mathbf{s}}), \tag{3}$$

where $\mathrm{ASR}(\cdot)$ denotes the functional representation of ASR processing.

B. Orthogonal projection-based decomposition of speech enhancement errors

The enhanced signal inevitably contains estimation errors. To define SE evaluation measures such as SDR, a previous study [17] proposed decomposing the estimation errors into three error components, i.e., interference, noise, and artifact errors. Given the reference source, interference, and noise signals ($\mathbf{s}$, $\mathbf{i}$, and $\mathbf{n}$), the estimated signal $\widehat{\mathbf{s}}$ is decomposed as:

$$\widehat{\mathbf{s}} = \mathbf{s}_{\text{target}} + \mathbf{e}_{\text{interf}} + \mathbf{e}_{\text{noise}} + \mathbf{e}_{\text{artif}}, \tag{4}$$

where $\mathbf{s}_{\text{target}} \in \mathbb{R}^T$ is called a target source component, and $\mathbf{e}_{\text{interf}} \in \mathbb{R}^T$, $\mathbf{e}_{\text{noise}} \in \mathbb{R}^T$, and $\mathbf{e}_{\text{artif}} \in \mathbb{R}^T$ denote the interference, noise, and artifact error components, respectively. Figure 1-(a) illustrates the signal decomposition, where, for illustration simplicity, we assume that there is no interference signal in the observed signal, i.e., $\mathbf{y} = \mathbf{s} + \mathbf{n}$, and that the signal delay is not considered.

The above decomposition in Eq. (4) is defined using orthogonal projections. Let $\mathbf{s}^\tau$, $\mathbf{i}^\tau$, and $\mathbf{n}^\tau$ be the source, interference, and noise signals delayed by τ samples. Three orthogonal projection matrices are defined for the decomposition: 1) orthogonal projection matrix onto the subspace spanned by the source signals $\{\mathbf{s}^\tau\}_{\tau=0}^{L-1}$ (denoted by $\mathbf{P}_{\mathbf{s}} \in \mathbb{R}^{T\times T}$), 2) orthogonal projection matrix onto the subspace spanned by the source and interference signals $\{\mathbf{s}^\tau, \mathbf{i}^\tau\}_{\tau=0}^{L-1}$ (denoted by $\mathbf{P}_{\mathbf{s},\mathbf{i}} \in \mathbb{R}^{T\times T}$), and 3) orthogonal projection matrix onto the subspace spanned by the source, interference, and noise signals $\{\mathbf{s}^\tau, \mathbf{i}^\tau, \mathbf{n}^\tau\}_{\tau=0}^{L-1}$ (denoted by $\mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}} \in \mathbb{R}^{T\times T}$), where $L-1$ is the number of maximum delay allowed or, in other words, L is the number of basis vectors for the projection. These matrices can be obtained as follows:

$$\mathbf{P}_{\mathbf{s}} \coloneqq \mathbf{A}_{\mathbf{s}}(\mathbf{A}_{\mathbf{s}}^{\mathsf{T}}\mathbf{A}_{\mathbf{s}})^{-1}\mathbf{A}_{\mathbf{s}}^{\mathsf{T}}, \tag{5}$$

$$\mathbf{P}_{\mathbf{s},\mathbf{i}} \coloneqq \mathbf{A}_{\mathbf{s},\mathbf{i}}(\mathbf{A}_{\mathbf{s},\mathbf{i}}^{\mathsf{T}}\mathbf{A}_{\mathbf{s},\mathbf{i}})^{-1}\mathbf{A}_{\mathbf{s},\mathbf{i}}^{\mathsf{T}}, \tag{6}$$

$$\mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}} \coloneqq \mathbf{A}_{\mathbf{s},\mathbf{i},\mathbf{n}}(\mathbf{A}_{\mathbf{s},\mathbf{i},\mathbf{n}}^{\mathsf{T}}\mathbf{A}_{\mathbf{s},\mathbf{i},\mathbf{n}})^{-1}\mathbf{A}_{\mathbf{s},\mathbf{i},\mathbf{n}}^{\mathsf{T}}, \tag{7}$$

where

$$\mathbf{A}_{\mathbf{s}} \coloneqq [\mathbf{s}^0, \ldots, \mathbf{s}^{L-1}],$$
$$\mathbf{A}_{\mathbf{s},\mathbf{i}} \coloneqq [\mathbf{s}^0, \ldots, \mathbf{s}^{L-1}, \mathbf{i}^0, \ldots, \mathbf{i}^{L-1}],$$
$$\mathbf{A}_{\mathbf{s},\mathbf{i},\mathbf{n}} \coloneqq [\mathbf{s}^0, \ldots, \mathbf{s}^{L-1}, \mathbf{i}^0, \ldots, \mathbf{i}^{L-1}, \mathbf{n}^0, \ldots, \mathbf{n}^{L-1}].$$

Then, the decomposed components in Eq. (4) ($\mathbf{s}_{\text{target}}$, $\mathbf{e}_{\text{interf}}$, $\mathbf{e}_{\text{noise}}$, and $\mathbf{e}_{\text{artif}}$) are defined based on the projection matrices as:

$$\mathbf{s}_{\text{target}} = \mathbf{P}_{\mathbf{s}}\widehat{\mathbf{s}}, \tag{8}$$

$$\mathbf{e}_{\text{interf}} = \mathbf{P}_{\mathbf{s},\mathbf{i}}\widehat{\mathbf{s}} - \mathbf{P}_{\mathbf{s}}\widehat{\mathbf{s}}, \tag{9}$$

$$\mathbf{e}_{\text{noise}} = \mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\widehat{\mathbf{s}} - \mathbf{P}_{\mathbf{s},\mathbf{i}}\widehat{\mathbf{s}}, \tag{10}$$

$$\mathbf{e}_{\text{artif}} = \widehat{\mathbf{s}} - \mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\widehat{\mathbf{s}}. \tag{11}$$

Here, $\mathbf{e}_{\text{interf}}$ and $\mathbf{e}_{\text{noise}}$ represent the residual interference and noise error components that can be expressed as a linear combination of the reference signals (i.e., source, interference, noise signals). On the other hand, $\mathbf{e}_{\text{artif}}$ represents the error component that cannot be expressed as a linear combination of the reference signals, and thus it was deemed an artifact error that is artificially generated by the SE systems.

C. Evaluation measures

The decomposition described in Section III-B was originally proposed to derive four SE evaluation metrics [17]: 1) SDR, 2) signal-to-interference ratio (SIR), 3) signal-to-noise ratio (SNR), and 4) SAR. These metrics are defined as follows:

$$\mathrm{SDR} \coloneqq 10\log_{10}\frac{\|\mathbf{s}_{\text{target}}\|^2}{\|\mathbf{e}_{\text{interf}} + \mathbf{e}_{\text{noise}} + \mathbf{e}_{\text{artif}}\|^2}, \tag{12}$$

$$\mathrm{SIR} \coloneqq 10\log_{10}\frac{\|\mathbf{s}_{\text{target}}\|^2}{\|\mathbf{e}_{\text{interf}}\|^2}, \tag{13}$$

$$\mathrm{SNR} \coloneqq 10\log_{10}\frac{\|\mathbf{s}_{\text{target}} + \mathbf{e}_{\text{interf}}\|^2}{\|\mathbf{e}_{\text{noise}}\|^2}, \tag{14}$$

$$\mathrm{SAR} \coloneqq 10\log_{10}\frac{\|\mathbf{s}_{\text{target}} + \mathbf{e}_{\text{interf}} + \mathbf{e}_{\text{noise}}\|^2}{\|\mathbf{e}_{\text{artif}}\|^2}. \tag{15}$$

The SDR metric represents the total error (i.e., equally penalizing each error component) between reference and estimated signals, while the other metrics (i.e., SIR, SNR, SAR) individually evaluate each error component such as interference, noise, and artifact.

Algorithm 1 Processing flow of direct scaling analysis

Input: evaluation dataset D (each element $d \in D$ is a set of the observed signal, reference signals, and reference transcription, i.e., $d = \{\mathbf{y}, \mathbf{s}, \mathbf{i}, \mathbf{n}, \mathbf{W}\}$), trained SE model $\mathrm{SE}(\cdot)$, trained ASR model $\mathrm{ASR}(\cdot)$

Output: set of evaluated scores Scores

1: Scores $\leftarrow \emptyset$
2: **for all** $\{\mathbf{y}, \mathbf{s}, \mathbf{i}, \mathbf{n}, \mathbf{W}\} \in D$ **do**
3: $\quad \widehat{\mathbf{s}} = \mathrm{SE}(\mathbf{y})$
4: $\quad \mathbf{s}_{\text{target}}, \mathbf{e}_{\text{interf}}, \mathbf{e}_{\text{noise}}, \mathbf{e}_{\text{artif}} = \mathrm{OPD}(\widehat{\mathbf{s}}, \mathbf{s}, \mathbf{i}, \mathbf{n})$
5: $\quad$ **for all** $\omega_{\text{interf}} \in \{0.1, 0.2, \ldots, 1.5\}$ **do**
6: $\quad\quad$ **for all** $\omega_{\text{noise}} \in \{0.1, 0.2, \ldots, 1.5\}$ **do**
7: $\quad\quad\quad$ **for all** $\omega_{\text{artif}} \in \{0.1, 0.2, \ldots, 1.5\}$ **do**
8: $\quad\quad\quad\quad \widehat{\mathbf{s}}_{\text{DSA}} = \mathbf{s}_{\text{target}} + \omega_{\text{interf}}\ \mathbf{e}_{\text{interf}} + \omega_{\text{noise}}\ \mathbf{e}_{\text{noise}} + \omega_{\text{artif}}\ \mathbf{e}_{\text{artif}}$
9: $\quad\quad\quad\quad \widehat{\mathbf{W}}_{\text{DSA}} = \mathrm{ASR}(\widehat{\mathbf{s}}_{\text{DSA}})$
10: $\quad\quad\quad\quad S_{\text{DSA}} = \mathrm{WER}(\mathbf{W}, \widehat{\mathbf{W}}_{\text{DSA}})$
11: $\quad\quad\quad\quad \text{Scores} \leftarrow \text{Scores} \cup \{(\omega_{\text{interf}}, \omega_{\text{noise}}, \omega_{\text{artif}}, S_{\text{DSA}})\}$
12: $\quad\quad\quad$ **end for**
13: $\quad\quad$ **end for**
14: $\quad$ **end for**
15: **end for**
16: **return** Scores

IV. Proposed analyses and approaches to reducing artifact errors

A. Controlling SIR, SNR, and SAR by directly scaling error components

To reveal the impact of the different error components (i.e., interference, noise, and artifact errors) on ASR performance, we propose a novel DSA scheme.

Based on the signal decomposition in Eq. (4), we generate the modified enhanced signal $\widehat{\mathbf{s}}_{\text{DSA}} \in \mathbb{R}^T$ by directly scaling (increasing/decreasing) the error components $\mathbf{e}_{\text{interf}}, \mathbf{e}_{\text{noise}}, \mathbf{e}_{\text{artif}}$ as:

$$\widehat{\mathbf{s}}_{\text{DSA}} = \mathbf{s}_{\text{target}} + \omega_{\text{interf}}\ \mathbf{e}_{\text{interf}} + \omega_{\text{noise}}\ \mathbf{e}_{\text{noise}} + \omega_{\text{artif}}\ \mathbf{e}_{\text{artif}}, \quad (16)$$

where ω_{interf}, ω_{noise}, and ω_{artif} are the scaling parameters that control the amounts of interference, noise, and artifact error components. By changing the scaling parameters, we can obtain enhanced signals with different levels of each error component $\mathbf{e}_{\text{interf}}, \mathbf{e}_{\text{noise}}, \mathbf{e}_{\text{artif}}$ while retaining the same target source component $\mathbf{s}_{\text{target}}$.

Algorithm 1 summarizes the processing flow of DSA, where $\mathrm{OPD}(\cdot)$ denotes the functional representation of the orthogonal projection-based decomposition (i.e., Eq. (8)–(11)) and $\mathrm{WER}(\cdot)$ denotes the functional representation of the WER (i.e., Levenshtein distance) computation. Applying DSA assumes that the observed noisy signal $\mathbf{y}$, reference source, interference, and noise signals $(\mathbf{s}, \mathbf{i}, \mathbf{n})$, and reference transcription $\mathbf{W}$ are available.

Given the observed and reference signals, we first obtain the enhanced signals $\widehat{\mathbf{s}}$ by applying the SE system to the observed noisy signal $\mathbf{y}$ (line 3) and then obtain the decomposed signals $\mathbf{s}_{\text{target}}, \mathbf{e}_{\text{interf}}, \mathbf{e}_{\text{noise}}, \mathbf{e}_{\text{artif}}$ by applying the orthogonal projection-based decomposition (line 4). By changing the scaling parameters $\omega_{\text{interf}}, \omega_{\text{noise}}, \omega_{\text{artif}}$ (lines 5, 6, and 7), we obtain the modified enhanced signals $\widehat{\mathbf{s}}_{\text{DSA}}$ with various amounts of each error component (line 8). By inputting such modified enhanced signals to the ASR system (line 9) and evaluating the ASR performance (line 10), we can directly measure the impact of each error component on ASR performance.

Since DSA requires having access to the reference signals, it cannot be applied to practical use scenes, i.e., real recordings. Here, we aim to identify how each error component impacts ASR performance.

B. Improving SAR with observation adding post-processing

As a practical approach to mitigating artifact errors, we consider the OA post-processing. In this paper, we propose an interpretation of OA from the viewpoint of orthogonal projection-based decomposition and also give a mathematical proof that OA can monotonically increase the SAR under a mild condition.

1) Formulation: OA is a simple technique that interpolates the enhanced signal $\widehat{\mathbf{s}}$ with the observed signal $\mathbf{y}$ as[1]:

$$\widehat{\mathbf{s}}_{\text{OA}} = (1 - \omega_{\text{OA}})\ \widehat{\mathbf{s}} + \omega_{\text{OA}}\ \mathbf{y}, \quad (17)$$

where $\widehat{\mathbf{s}}_{\text{OA}} \in \mathbb{R}^T$ denotes the modified enhanced signal with OA, and $0 < \omega_{\text{OA}} < 1$ is an interpolation parameter that controls the balance between the enhanced and observed signals. Note that setting $\omega_{\text{OA}} = 0$ gives the original enhanced signal, and $\omega_{\text{OA}} = 1$ the observed noisy signal.

Figure 1-(b) illustrates the OA's behavior from the viewpoint of error decomposition. In the figure, $\overline{\mathbf{s}}_{\text{target}}$, $\overline{\mathbf{e}}_{\text{noise}}$, and $\overline{\mathbf{e}}_{\text{artif}}$ denote the target source, noise error, and artifact error components of the modified enhanced signal with OA, respectively. OA increases the target source and noise error components but decreases the artifact error component.

Unlike DSA, OA does not require having access to the reference signals and thus can be applied to real recordings.

2) Theoretical analysis: We show in the next proposition that OA improves the SAR of the enhanced signal $\widehat{\mathbf{s}}$ under a mild condition.

Proposition 1. The OA operation in Eq (17) improves the SAR of the original enhanced signal $\widehat{\mathbf{s}} = \mathrm{SE}(\mathbf{y})$ if it satisfies $\langle \widehat{\mathbf{s}}, \mathbf{y} \rangle > 0$.

Proof. Based on Eqs. (8)–(11), (15), and (17), the SAR improvement (denoted as SARi) from the original enhanced

[1]An alternative implementation of the OA post-processing consists of simply adding back a scaled version of the observed signal to the enhanced signal [32] as: $\widehat{\mathbf{s}}_{\text{OA}} = \widehat{\mathbf{s}} + \omega_{\text{OA}}\ \mathbf{y}$. These two implementations are equivalent except for the scale of the resultant modified signals. Note that the scale difference does not affect the orthogonal projection-based SE metrics (i.e., Eqs. (12)–(15)).

signal $\widehat{\mathbf{s}}$ to the modified enhanced signal with OA $\widehat{\mathbf{s}}_{\text{OA}}$ can be computed as:

$$\begin{aligned}
\text{SARi} &= 10\log_{10}\frac{\|\mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\widehat{\mathbf{s}}_{\text{OA}}\|^2}{\|\widehat{\mathbf{s}}_{\text{OA}}-\mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\widehat{\mathbf{s}}_{\text{OA}}\|^2} \\
&\quad - 10\log_{10}\frac{\|\mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\widehat{\mathbf{s}}\|^2}{\|\widehat{\mathbf{s}}-\mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\widehat{\mathbf{s}}\|^2}, \\
&= 10\log_{10}\frac{\|(1-\omega_{\text{OA}})\ \mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\widehat{\mathbf{s}}+\omega_{\text{OA}}\ \mathbf{y}\|^2}{(1-\omega_{\text{OA}})^2\ \|\widehat{\mathbf{s}}-\mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\widehat{\mathbf{s}}\|^2} \\
&\quad - 10\log_{10}\frac{\|\mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\widehat{\mathbf{s}}\|^2}{\|\widehat{\mathbf{s}}-\mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\widehat{\mathbf{s}}\|^2}, \\
&= 10\log_{10}\frac{\|(1-\omega_{\text{OA}})\ \mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\widehat{\mathbf{s}}+\omega_{\text{OA}}\mathbf{y}\|^2}{(1-\omega_{\text{OA}})^2\ \|\mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\widehat{\mathbf{s}}\|^2}, \\
&= 10\log_{10}\left[1+\frac{\omega_{\text{OA}}^2\|\mathbf{y}\|^2+2(1-\omega_{\text{OA}})\omega_{\text{OA}}\langle\mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\widehat{\mathbf{s}},\mathbf{y}\rangle}{(1-\omega_{\text{OA}})^2\ \|\mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\widehat{\mathbf{s}}\|^2}\right],
\end{aligned}$$

where we use $\mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\mathbf{y}=\mathbf{y}$ in the second equality. Due to $0<\omega_{\text{OA}}<1$ from OA's definition, $(1-\omega_{\text{OA}})\omega_{\text{OA}}>0$. Thus, SARi >0 holds if $\langle\mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\widehat{\mathbf{s}},\mathbf{y}\rangle>0$. This sufficient condition can be rewritten as $\langle\mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\widehat{\mathbf{s}},\mathbf{y}\rangle=\langle\widehat{\mathbf{s}},\mathbf{P}_{\mathbf{s},\mathbf{i},\mathbf{n}}\mathbf{y}\rangle=\langle\widehat{\mathbf{s}},\mathbf{y}\rangle>0$, which concludes the proof. □

The condition $\langle\widehat{\mathbf{s}},\mathbf{y}\rangle>0$ holds with 50 % probability even in the case where $\widehat{\mathbf{s}}$ is a randomly generated vector. Thus, it should be met even for SE front-ends that are not very accurate, and even more likely to be met for the recent high-quality SE systems. Assuming such SE systems, we can roughly expect $\widehat{\mathbf{s}}\approx\mathbf{s}+\varepsilon_i\mathbf{i}+\varepsilon_n\mathbf{n}+\widehat{\mathbf{e}}_{\text{artif}}$ with $\varepsilon_n\in\mathbb{R}$ and $\varepsilon_i\in\mathbb{R}$, where $|\varepsilon_n|$ and $|\varepsilon_i|$ are small. Since $\mathbf{e}_{\text{artif}}$ is orthogonal to $\mathbf{s}$, $\mathbf{i}$, and $\mathbf{n}$ (i.e., $\langle\widehat{\mathbf{e}}_{\text{artif}},\mathbf{s}\rangle=0$, $\langle\widehat{\mathbf{e}}_{\text{artif}},\mathbf{i}\rangle=0$, and $\langle\widehat{\mathbf{e}}_{\text{artif}},\mathbf{n}\rangle=0$) and $\mathbf{s}$, $\mathbf{i}$, and $\mathbf{n}$ are generally uncorrelated (i.e., $\langle\mathbf{s},\mathbf{i}\rangle\approx\mathbf{0}$, $\langle\mathbf{s},\mathbf{n}\rangle\approx\mathbf{0}$, and $\langle\mathbf{i},\mathbf{n}\rangle\approx\mathbf{0}$), we have $\langle\widehat{\mathbf{s}},\mathbf{y}\rangle\approx\langle\mathbf{s}+\varepsilon_i\mathbf{i}+\varepsilon_n\mathbf{n}+\widehat{\mathbf{e}}_{\text{artif}},\mathbf{s}+\mathbf{i}+\mathbf{n}\rangle\approx\|\mathbf{s}\|^2+\varepsilon_i\|\mathbf{i}\|^2+\varepsilon_n\|\mathbf{n}\|^2$. Since $|\varepsilon_i|$ and $|\varepsilon_n|$ can be considered sufficiently small, we can expect $\langle\widehat{\mathbf{s}},\mathbf{y}\rangle>0$ and that OA almost always improves the SAR by Proposition 1.

In addition to the above theoretical analysis, we experimentally investigated the effect of OA. In our preliminary experiment (Section VII-A1), we observed that the SAR scores of the modified enhanced signals with OA $\widehat{\mathbf{s}}_{\text{OA}}$ are better than those of the original enhanced signals $\widehat{\mathbf{s}}$ for all of the evaluated utterances. Thus, the OA can be used as simple and practical post-processing to improve the SAR of the enhanced signals.

C. Improving SAR with artifact-boosted training loss

As an alternative approach to mitigating artifact errors and consequently improving ASR performance, we can consider adopting a training objective that reduces artifact errors. Like OA (but unlike DSA), this scheme assumes that only the observed signal is available in the inference stage and thus can be applied to real recordings.

We assume that paired data of observed noisy signal and reference source, interference, and noise signals $\{\mathbf{y},\mathbf{s},\mathbf{i},\mathbf{n}\}$ are available for training the SE model. In this paper, as a novel training objective, we propose the AB-SDR loss as:

$$\mathcal{L}_{\text{AB-SDR}}=-10\log_{10}\frac{\|\mathbf{s}_{\text{target}}\|^2}{\|\mathbf{e}_{\text{interf}}+\mathbf{e}_{\text{noise}}+\alpha\mathbf{e}_{\text{artif}}\|^2}, \tag{18}$$

where $\alpha>1.0$ denotes the weighting parameter that prioritizes the artifact error component. When we set $\alpha=1.0$, the loss function corresponds to the original SDR loss [17], [37] that equally penalizes each error component (i.e., interference, noise, and artifact errors). By setting α to a larger value, we boost the effect of artifact errors in the training loss, which drives the model to reduce artifact errors in the estimated signals more than with conventional losses.

An intuitive approach to improving the SAR would be to incorporate regularization for decreasing SAR (i.e., Eq. (15)) in the training objective. However, in our preliminary experiment, we observed that SAR-based regularization may cause training instability, probably because the SAR becomes optimal when the estimated signal matches the observed noisy signal. Thus, in this study, we adopt an SDR-based loss that weighs the artifact error component as in Eq. (18).

V. Experimental conditions

To confirm our hypothesis, i.e., the artifact errors have a larger impact on the degradation of ASR performance, we conducted the analytical experiments based on the proposed DSA (Section VI). Then, we evaluated the effect of the proposed approaches to reducing the artifact errors, i.e., OA (Section VII-A) and AB-SDR (Section VII-B). Moreover, we conducted additional analyses in Section VIII: 1) analysis of the effect of SE architecture and ASR's training data, 2) evaluation with different speech and noise corpora, and 3) evaluation with the real recordings.

A. Evaluated datasets

We created three datasets of simulated single-channel speech signals under reverberant and noisy conditions. We used the Wall Street Journal (WSJ0) corpus [40] and the Corpus of Spontaneous Japanese (CSJ) corpus [41] for the speech signals, and used the CHiME-3 corpus [7] and the DEMAND corpus [42] for the noise signals. We summarize the basic information of the evaluated datasets in Table I.

For the WSJ0 speech corpus, the speech sources for the training set were selected from the WSJ0's training set "si_tr_s." Those for the development and evaluation sets were selected from the WSJ0's development set "si_dt_05" and evaluation set "si_et_05." For the CSJ speech corpus, the speech sources for the training and development sets were selected from the CSJ's training set. Those for the evaluation set were selected from the CSJ's evaluation sets "eval1," "eval2," and "eval3." Here, we divided the CSJ's training set into subsets for training and development, containing 95 % and 5 %, respectively.

For the CHiME-3 noise corpus, we divided the noise sources into three subsets for training, development, and evaluation, containing 80 %, 10%, and 10%, respectively, of the noise data of each environment (i.e., on a bus (BUS), in a cafe (CAF), a pedestrian area (PED), and at a street junction (STR)). We used the fifth-channel signals of CHiME-3's noise data for generating noisy speech signals. Similarly, for the DEMAND noise corpus, we divided noise sources into three subsets for training, development, and evaluation, containing

TABLE I
SUMMARY OF EVALUATED DATASETS.

WSJ_CHIME_ST (single-talker scenario)	# of utt.	SNR [dB]	SIR [dB]
Training set	30,000	$0 \sim 10$	–
Development set	5,000	$0 \sim 10$	–
Evaluation set	5,000	0	–
WSJ_CHIME_MT (multi-talker scenario)	**# of utt.**	**SNR [dB]**	**SIR [dB]**
Training set	30,000	$5 \sim 20$	$5 \sim 20$
Development set	5,000	$5 \sim 20$	$5 \sim 20$
Evaluation set	5,000	10	5
CSJ_DEMAND_ST (single-talker scenario)	**# of utt.**	**SNR [dB]**	**SIR [dB]**
Training set	154,147	$0 \sim 10$	–
Development set	8,112	$0 \sim 10$	–
Evaluation set	3,949	0	–
CHiME-3's real recordings (single-talker scenario)	**# of utt.**	**SNR [dB]**	**SIR [dB]**
Evaluation set	1,320	$2 \sim 7^*$	–

* The SNR value is estimated [45] because we do not have access to reference source and noise signals for real recordings.

TABLE II
SUMMARY OF CONFIGURATIONS OF SPEECH ENHANCEMENT SYSTEMS.

Configuration of temporal convolutional network	
Number of channels in bottleneck (B)	128
Number of channels in skip-connection (Sc)	128
Number of channels in convolution blocks (H)	512
Kernel size in convolution blocks (P)	3
Number of convolution blocks in each repeat (X)	8
Number of repeats (R)	3
Configuration of 1D convolution encoder/decoder	
Filter length	16 samples
Hop size	8 samples
Number of filters	512
Configuration of STFT encoder/decoder	
Frame length	1024 samples
Frame shift	256 samples
Window function	Hanning

70 %, 10%, and 20%, respectively, of the noise data of each noise environment (e.g., living room (DLIVING), university restaurant (PRESTO), meeting room (OMEETING), etc.). We used the first-channel signals of DEMAND's noise data for generating noisy speech signals.

To create the reverberant speech sources, we generated room impulse response (RIR) based on the image method [43], using the gpuRIR simulation toolkit [44]. We randomly generated the RIR configuration (e.g., room geometry, source and array positions, reverberation time, etc.) for each simulated RIR. We set the reverberation time (T60) between 0.2 s and 0.6 s.

1) WSJ_CHIME_ST and WSJ_CHIME_MT: We created single-talker and multi-talker noisy speech datasets based on the WSJ0's speech and CHiME-3's noise signals; we refer to the former single-talker dataset as "WSJ_CHIME_ST" and to the latter multi-talker dataset as "WSJ_CHIME_MT." The WSJ_CHIME_ST dataset is similar to CHiME-3's simulation dataset but with different RIRs. The original CHiME-3 simulation data contains only single-talker conditions and does not include the reference source and noise signals. To create multi-talker data, we thus generated RIRs with the image method simulating far-field conditions. We used these simulation configurations for the single-talker case as well to match the multi-talker case.

For the multi-talker scenario, the generated data consisted of simulated reverberant noisy two-speaker mixtures. In this experiment, to avoid target speaker ambiguity, we assumed that the target speaker's speech is dominant within the utterance (i.e., the target speaker's speech has a larger power than the interfering speaker's speech). Thus, the problem consists of extracting the dominant speaker[2] in the two-speaker mixture.

2) CSJ_DEMAND_ST: To verify whether the experimental findings hold for different evaluated datasets, we also created another single-talker noisy speech dataset based on CSJ's speech and DEMAND's noise signals, which we refer to as "CSJ_DEMAND_ST." "WSJ_CHIME_ST" is created using the WSJ speech (English read speech) and CHiME-3 noise (4 environments) corpora, while "CSJ_DEMAND_ST" is created using the CSJ speech (Japanese Spontaneous speech) and DEMAND noise (18 environments) corpora. It contains many more training utterances and a different language, more natural speaking styles, and more diverse noisy environments.

3) CHiME-3's real recordings: We used the simulated datasets to analyze the relations between the SE metrics (SDR, SIR, SNR, SAR [17]) and the ASR metric (WER). In addition, to confirm the effectiveness of our findings for real recordings, we used the real-recorded evaluation data of the CHiME-3 corpus "et05_real". Compared to the simulated evaluation sets, the real recordings comprise real reverberation and also cover more variations in terms of SNR condition ($2 \sim 7$ dB [45]).

B. Evaluated systems

1) Speech enhancement front-end: For the SE front-end, we adopted two types of SE models that work in different input and output domains (i.e., time-domain and short-time Fourier transform (STFT)-domain), representing major categories of SE model architectures.

We employed the temporal convolutional network (TCN)-based single-channel SE architecture [46] that is similar to the time-domain audio separation network (TasNet) [15]. The network is composed of an encoder layer, internal TCN blocks, and a decoder layer. First, the encoder layer maps the time-domain signals to an intermediate representation, and then it

[2]There also exists the source separation or target speech extraction setup, which separates all of the sources or extracts a single target source given the speaker enrollment. In this study, for model simplicity, we focused on the dominant speech extraction setup, which would be one of the practical use cases for multi-talker ASR tasks.

is further processed by the convolution blocks. Finally, the decoder layer directly remaps the intermediate representation to time-domain signals.

For the encoder/decoder mapping, we adopted two types of transformations: 1) learnable transformation with 1D convolution and 1D deconvolution and 2) fixed transformation with STFT and inverse short-time Fourier transform (iSTFT). The former learnable transformation-based model (Enhan_T) corresponds to the famous time-domain SE model, i.e., Conv-TasNet [15], and the latter fixed transformation model (Enhan_F) corresponds to its STFT-domain variant [46]. The configurations related to the encoder/decoder and TCN architecture are briefly summarized in Table II, where we follow the notations introduced in a previous study [15].

For the training loss, we compared three types of loss functions: 1) classical scale-dependent SNR loss [47], 2) original SDR loss ($\alpha = 1.0$ in Eq. (18)), and 3) proposed AB-SDR loss ($\alpha > 1.0$ in Eq. (18)). We trained all of the models using the Adam optimizer [48] with gradient clipping [49], with an initial learning rate of 0.001, a batch size of 24, and a maximum of 100 epochs. We set the number of basis vectors for the projections in the SDR and AB-SDR losses to $L = 2$ and $L = 1$ for the single-talker and multi-talker scenarios, respectively, based on the WER scores in the preliminary experiments[3]. In addition, for the AB-SDR loss, we tuned the weighting parameter α for values within $\{1.0, 1.5, \ldots, 3.5\}$, and obtained $\alpha = 1.5$ and $\alpha = 2.0$ for single-talker and multi-talker scenarios, respectively.

In this study, we implemented the SE systems based on Asteroid's Conv-TasNet implementation [15], [50].

2) Speech recognition back-end: The proposed DSA scheme requires a large number of ASR decoding experiments. Thus for the ASR back-end, we adopted a computationally efficient deep neural network-hidden Markov model (DNN-HMM) hybrid ASR model[4]. The system was trained using the lattice-free maximum mutual information (MMI) criterion [53] and decoded with a trigram language model. In this study, we implemented the ASR systems based on Kaldi's standard recipe [54] for the CHiME-4 and CSJ tasks. The details of the systems are shown in Kaldi's recipe[5,6].

For the training dataset of the acoustic models, we adopted three types of signals: 1) noisy signal (Noisy), 2) enhanced signal (Enhan), and 3) noisy and enhanced signals (Noisy+Enhan). We obtained the enhanced signals for the training dataset by inputting the noisy signals into the SE systems. For the third option, we used both noisy and enhanced signals for the training, and thus the total number of training utterances (i.e., iterations) is twice those for the first and second options.

[3]Note that setting L to a larger value in the training loss provides too much degree of freedom for the projections, which allows the trained SE model to generate meaningless signals (e.g., high SDR but quite bad WER).

[4]Parallel to this work, in another study [51], we have investigated the effectiveness of OA post-processing for the state-of-the-art end-to-end ASR model [24] with a pre-trained self-supervised learning-based feature extractor [52] trained using a huge amount of speech and noise data, and we confirmed the influence of artifact errors even for such a stronger huge ASR back-end.

[5]https://github.com/kaldi-asr/kaldi/tree/master/egs/chime4

[6]https://github.com/kaldi-asr/kaldi/tree/master/egs/csj

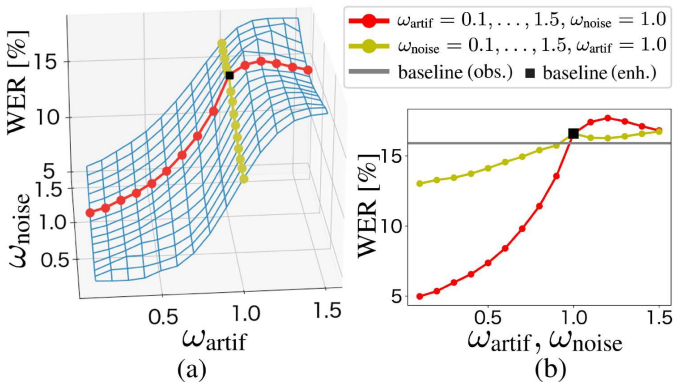

Fig. 2. Results of DSA-based evaluation (WER [%] (lower is better)) for single-talker setup (WSJ_CHIME_ST), which modifies the scale of noise (ω_{noise}) and artifact (ω_{artif}) error components.

C. Evaluation metrics

As the evaluation metrics, we used six SE and one ASR measures; 1) SDR, 2) SIR, 3) SNR, 4) SAR, 5) STOI, 6) PESQ, and 7) WER, where higher is better for SDR, SIR, SNR, SAR, STOI, and PESQ, while lower is better for WER. The number of basis vectors for calculating the orthogonal projection-based SE metrics (i.e., SDR, SIR, SNR, and SAR) is set to $L = 512$ by following the de facto standard BSSEval's implementation[7] [17]. Note that we slightly modified the BSSEval's implementation to accept reference noise signals $\mathbf{n}$ for computing the metrics under noisy conditions.

VI. Experimental analysis by directly scaling error components

A. Single-talker scenario

Figure 2 shows the results of DSA for the single-talker noisy speech dataset, WSJ_CHIME_ST. Here, for the evaluated SE system, we adopt the time-domain SE model (Enhan_T) trained with the SNR loss. Note that in the single-talker scenario, the observed signal contains background noise but no interference speakers, and thus there are only two types of SE errors, i.e., noise and artifact[8]. Figure 2-(a) is a three-dimensional plot showing the relationship between the SE errors (i.e., noise and artifact) and WER scores, which are obtained by the DSA procedure described in Section IV-A and Algorithm 1. Figure 2-(b) is the corresponding two-dimensional plot, where we vary either the scaling parameters for noise ω_{noise} or artifact ω_{artif}. Each curve in the two-dimensional plot corresponds to the curve of the same color in the three-dimensional plot.

In the figures, the gray line (baseline (obs.)) represents the baseline score of the observed signals $\mathbf{y}$, and the black square (baseline (enh.)) represents the baseline score of the original enhanced signals $\widehat{\mathbf{s}}$, i.e., $\omega_{\text{noise}} = \omega_{\text{artif}} = 1$. Similar to the results reported in the previous studies (e.g., [20]–[23]), Figure 2-(b) shows that the WER score of the original

[7]https://github.com/craffel/mir_eval

[8]From Eq. (9), $\mathbf{e}_{\text{interf}}$ becomes zero because the orthogonal projector $\mathbf{P}_{\mathbf{s},\mathbf{i}}$ is equal to $\mathbf{P}_{\mathbf{s}}$.

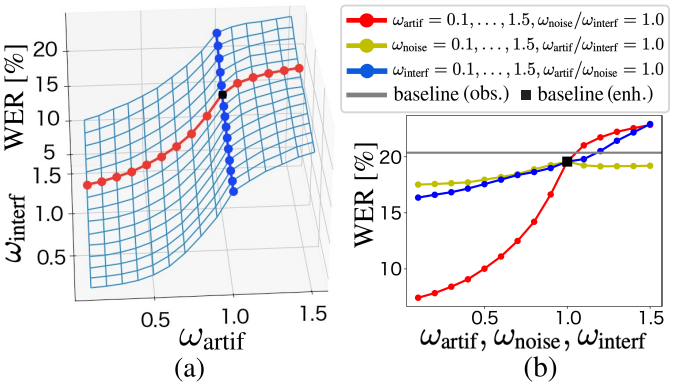

Fig. 3. Results of DSA-based evaluation (WER [%] (lower is better)) for multi-talker setup (WSJ_CHIME_MT), which modifies the scale of interference (ω_{interf}), noise (ω_{noise}), and artifact (ω_{artif}) error components.

enhanced signals (black square) is worse than that of the observed signals (gray line).

From the figures, we confirm that scaling down the artifact error component greatly improves ASR performance, while scaling down the noise error component has relatively little impact on ASR performance. These results confirm our hypothesis that, among the two types of SE errors (i.e., noise and artifact), the artifact errors have a larger impact on the degradation of ASR performance. This notable experimental finding suggests that the artifact error would be a reasonable signal-level numerical metric for explaining the cause of ASR performance degradation.

B. Multi-talker scenario

Figure 3 shows the results of DSA for the multi-talker noisy speech dataset, WSJ_CHIME_MT. Note that in the multi-talker scenario, the observed signal contains interfering speech and background noise, and thus there are three types of SE errors, i.e., interference, noise, and artifact. Figure 3-(a) and Figure 3-(b) are three-dimensional and two-dimensional plots, respectively, showing the relationship between SE errors (i.e., interference, noise, and artifact) and WER scores. Since we cannot illustrate a four-dimensional plot for the three scaling parameters, we instead show a three-dimensional plot in Figure 3-(a), where we only vary the two scaling parameters of interference ω_{interf} and artifact ω_{artif}.

From the figures, we confirm that scaling down the artifact error component greatly improves the ASR performance, while scaling down the noise or interference error component has relatively little impact on ASR performance. This trend is similar to what has been observed in the single-talker scenarios (Section VI-A). These results demonstrate that our hypothesis (i.e., artifact errors have a larger impact on the degradation of ASR performance) would hold even for a multi-talker scenario, which would involve more general and challenging acoustic conditions.

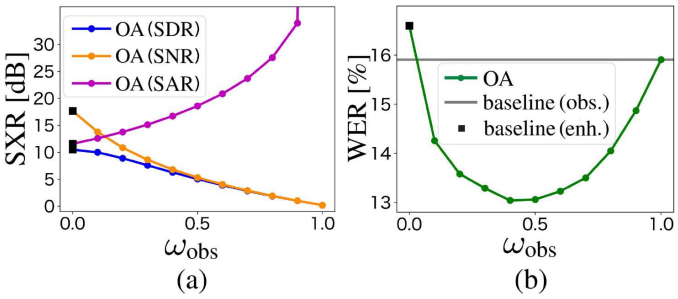

Fig. 4. Results of OA post-processing (SDR, SNR, SAR [dB] (higher is better) and WER [%] (lower is better)) for single-talker setup (WSJ_CHIME_ST). We use the compact notation SXR, in place of SDR, SNR, and SAR, to denote the ratio between two signals S and X, where X is either the distortion, noise, or artifacts.

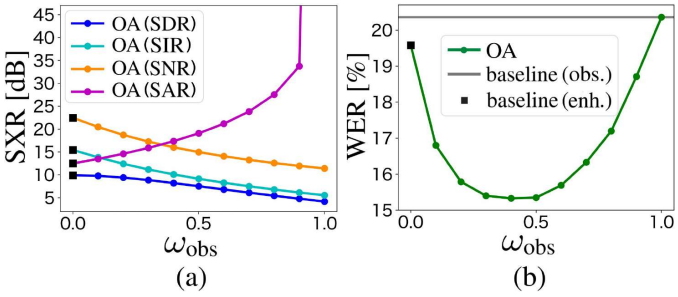

Fig. 5. Results of OA post-processing (SDR, SIR, SNR, SAR [dB] (higher is better) and WER [%] (lower is better)) for multi-talker setup (WSJ_CHIME_MT). SXR denotes values encompassing SDR, SIR, SNR, and SAR.

VII. Experimental results with practical approaches mitigating artifact errors

A. Evaluation of observation adding post-processing

1) Single-talker scenario: Figure 4 shows the results of the OA post-processing for the single-talker noisy speech dataset, WSJ_CHIME_ST. Here, for the evaluated SE system, we adopt the time-domain SE model (Enhan_T) trained with the SNR loss. Figure 4-(a) shows the SE metrics (i.e., SDR, SNR, and SAR) and Figure 4-(b) shows the ASR metric (i.e., WER) of the enhanced signals modified with OA as described in Section IV-B, where the interpolation parameter ω_{OA} varies within $\{0.0, 0.1, \ldots, 1.0\}$. In the figures, the result of $\omega_{\text{OA}} = 0$ corresponds to the baseline score of the original enhanced signals $\widehat{\mathbf{s}}$ (baseline (enh.)), and that of $\omega_{\text{OA}} = 1$ corresponds to the baseline score of the observed signals $\mathbf{y}$ (baseline (obs.)).

The figure shows that when ω_{OA} increases (i.e., increasing the proportion of the observed signal in $\widehat{\mathbf{s}}_{\text{OA}}$), the SDR and SNR decrease while the SAR monotonically increases. Moreover, we can confirm that, by adjusting ω_{OA} around 0.5 (i.e., adding the observed signal to the original enhanced signal in similar proportion), the modified enhanced signals with OA achieve better WER than both the baseline observed signals (gray line) and the original enhanced signals (black square).

2) Multi-talker scenario: Figure 5 shows the results of the OA post-processing for the multi-talker noisy speech dataset, WSJ_CHIME_MT. Figure 5-(a) and Figure 5-(b) show the SE metrics (i.e., SDR, SIR, SNR, and SAR) and the ASR

TABLE III
RESULTS OF ARTIFACT-BOOSTED TRAINING LOSS (SDR, SNR, SAR [DB], STOI, PESQ (HIGHER IS BETTER) AND WER [%] (LOWER IS BETTER)) FOR SINGLE-TALKER SETUP (WSJ_CHIME_ST).

Id	System	Loss	ω_{OA}	SDR	SNR	SAR	WER	STOI	PESQ
(1)	clean	–	–	–	–	–	4.4	–	–
(2)	noisy	–	–	0.2	0.2	–	15.9	0.73	1.21
(3)	Enhan_T	SNR	–	10.5	17.7	11.6	16.6	0.87	2.00
(4)		SDR	–	12.0	16.9	14.0	14.8	0.86	2.33
(5)		AB-SDR	–	11.3	13.1	16.7	13.0	0.84	1.94
(6)	+OA	SNR	0.4	6.3	6.8	16.7	13.0	0.83	1.53
(7)		SDR	0.4	4.1	4.4	18.8	13.1	0.77	1.25
(8)		AB-SDR	0.1	9.9	11.1	17.1	**12.8**	0.80	1.59

TABLE IV
RESULTS OF ARTIFACT-BOOSTED TRAINING LOSS (SDR, SIR, SNR, SAR [DB], STOI, PESQ (HIGHER IS BETTER) AND WER [%] (LOWER IS BETTER)) FOR MULTI-TALKER SETUP (WSJ_CHIME_MT).

Id	System	Loss	ω_{OA}	SDR	SIR	SNR	SAR	WER	STOI	PESQ
(1)	clean	–	–	–	–	–	–	3.6	–	–
(2)	noisy	–	–	4.2	5.5	11.4	–	20.4	0.77	1.39
(3)	Enhan_T	SNR	–	9.9	15.4	22.5	12.5	19.6	0.87	1.97
(4)		SDR	–	9.9	15.6	22.7	12.4	19.1	0.87	1.99
(5)		AB-SDR	–	8.5	10.6	16.6	17.1	**14.9**	0.85	1.75
(6)	+OA	SNR	0.4	8.2	10.1	16.0	17.4	15.2	0.85	1.72
(7)		SDR	0.4	8.3	10.2	16.1	17.3	**14.9**	0.85	1.72
(8)		AB-SDR	0.0[9]	8.5	10.6	16.6	17.1	14.9	0.85	1.75

metric (i.e., WER) of the enhanced signals modified with OA, respectively.

In the results of the multi-talker scenario in Figure 5, we can observe a similar trend to the results of the single-talker scenario in Figure 4: the OA post-processing 1) decreases the SDR, SIR, and SNR but monotonically increases the SAR and 2) achieves a better WER score than both the observed (gray line) and enhanced (black square) signals.

These results support our hypothesis that 1) the artifact component has a worse impact on ASR performance compared to the other error components and 2) the ASR performance of single-channel SE front-ends can be improved by decreasing the ratio of artifact errors (or increasing SAR). Moreover, the results demonstrate that OA post-processing would be an effective and practical approach to mitigating the impact of artifact errors, not only in the single-talker but also in the multi-talker scenario.

B. Evaluation of artifact-boosted training loss

1) Single-talker scenario: Table III compares evaluated SE systems trained with the conventional SNR and SDR and the proposed AB-SDR losses in terms of SE (i.e., SDR, SNR, SAR, STOI, and PESQ) and ASR (i.e., WER) performance for the single-talker noisy speech dataset, WSJ_CHIME_ST. We used the time-domain SE model (Enhan_T) in this experiment. In addition, Table III shows the performance of the evaluated SE systems applied with the OA post-processing (Enhan_T+OA). Here, we tuned the interpolation parameter ω_{OA} for each SE system by varying it within $\{0.0, 0.1, \ldots, 1.0\}$ and set the hyperparameter that achieved the best WER scores on the development set.

From the table, we observe that Enhan_T trained with AB-SDR loss (row 5) achieved a significantly higher SAR score, at the expense of decreasing the SNR score than Enhan_T trained with SNR and SDR losses (rows 3 and 4). As expected, we could confirm that the AB-SDR loss (row 5) achieved a better WER score than both the SNR and SDR losses (rows 3 and 4) and the baseline observed signals (row 2).

2) Multi-talker scenario: Table IV compares evaluated SE systems trained with the conventional SNR and SDR and the proposed AB-SDR losses in terms of SE and ASR performance for the multi-talker noisy speech dataset, WSJ_CHIME_MT. We used the time-domain SE model (Enhan_T) and also applied OA post-processing (Enhan_T+OA) in this experiment.

The table shows that the AB-SDR loss decreases the SIR and SNR scores but monotonically increases the SAR score, and it achieves a better WER score compared with both the observed signals and enhanced signals trained with SNR and SDR losses. This trend is similar to what has been observed in the single-talker scenario (in Table III).

These results further support our hypothesis (i.e., the artifact component has a worse impact on ASR performance) and demonstrate that it would hold not only for the single-talker but also for the multi-talker scenario. In addition, the results demonstrate that designing the training objective while considering the artifact errors contributes to improving ASR performance with single-channel SE front-ends, regardless of whether the single-talker or multi-talker scenarios is used.

Moreover, Table III and Table IV show that Enhan_T applied with both the AB-SDR loss and the OA post-processing (row 8) achieved the best WER score but the performance improvement was not as significant as when using OA with systems trained with SNR or SDR losses. This is probably because both approaches produce a similar effect on the resultant evaluated signals, i.e., decreasing the artifact (SAR) errors at the expense of increasing the interference (SIR) and noise (SNR) errors.

In our preliminary experiments, we investigated using the thresholded scale-invariant SAR [39], [47] as a regularization term in addition to the SDR loss. A careful examination of the results revealed that the SAR of the enhanced signal did not improve for some utterances but became very large for the other utterances. In the latter cases, the high SAR was achieved by generating enhanced signals that were very close to the observed signals $\mathbf{y}$. Consequently, although the average SAR score improved, it did not lead to improved WER score.

VIII. ADDITIONAL ANALYSES

A. Evaluation with different speech enhancement front-ends and speech recognition back-ends

In Section VII, we conducted experiments with the time-domain SE front-end and the ASR back-end trained independently of the SE front-end, i.e., trained with reverberant noisy speech. In this section, to investigate the influence of the

[9] $\omega_{\text{OA}} = 0.0$ implies that the OA post-processing did not improve the WER scores.

TABLE V
RESULTS OF OA POST-PROCESSING WITH TWO SE FRONT-ENDS AND FIVE ASR BACK-ENDS (WER [%] (LOWER IS BETTER)) ON WSJ_CHIME_ST DATASET.

SE \ ASR	Noisy	Enhan_T	Enhan_F	Noisy +Enhan_T	Noisy +Enhan_F
Noisy	15.9	30.5	23.9	15.5	14.9
Enhan_T	16.6	12.9	13.5	13.7	14.8
+ OA	13.0	12.9	13.3	**12.3**	**12.3**
Enhan_F	26.3	22.8	15.0	21.0	16.1
+ OA	14.6	18.6	15.0	14.1	13.6

SE model architecture (especially, input and output domains), we also adopt the two SE front-ends: 1) time-domain SE model (Enhan_T) and 2) STFT-domain SE model (Enhan_F). In addition, to investigate the influence of the training data for the ASR back-end, we evaluated the behavior of OA post-processing on the five ASR back-ends, which are trained with 1) the noisy signals (Noisy), 2) the enhanced signals generated by the time-domain SE model (Enhan_T), 3) the enhanced signals generated by the STFT-domain SE model (Enhan_F), 4) noisy signals and enhanced signals generated by Enhan_T, and 5) noisy signals and enhanced signals generated by Enhan_F.

Table V shows the WER scores for the combination of the evaluated SE and ASR systems. From the table, we can confirm that even when we use the weaker SE front-end (Enhan_F), the OA post-processing improves the ASR performance. These results show that mitigating the artifact errors is effective regardless of the SE model architecture.

Moreover, the table shows that by using the ASR back-end trained with the enhanced signals (e.g., Noisy+Enhan_T), both of the SE front-ends (Enhan_T and Enhan_F) obtained the ASR performance improvements compared with the ASR back-end trained independently of the SE front-end (Noisy). When the ASR back-end is trained exclusively on the matched SE front-end, OA is ineffective, as expected. However, when the back-end is trained on the combination of noisy and enhanced speech, OA improves performance. Interestingly, the latter system also achieved the best WER scores. These results suggest that ASR performance may still suffer from artifact errors not only when using the ASR back-end trained independently of the SE front-end but also when using an ASR back-end specialized for an SE front-end (i.e., trained with the enhanced signals of a specific SE front-end).

B. Evaluation with different speech and noise corpora

To verify whether the findings obtained by the above experiments hold for different speech and noise conditions, we also evaluated the behavior of OA post-processing on another single-talker noisy speech dataset, CSJ_DEMAND_ST, which is created based on the CSJ speech (Japanese Spontaneous speech) and DEMAND noise (18 environments) corpora. Figure 6-(a) shows the SE metrics (i.e., SDR, SNR, and SAR), and Figure 6-(b) shows the ASR metric (i.e., WER) of the enhanced signals modified with OA as described in Section IV-B, where the interpolation parameter ω_{OA} varies within $\{0.0, 0.1, \ldots, 1.0\}$.

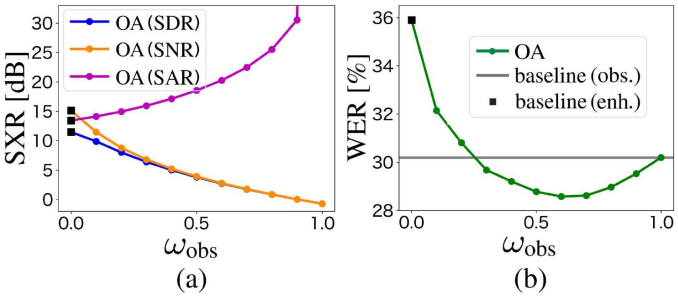

Fig. 6. Results of OA post-processing (SDR, SNR, SAR [dB] (higher is better) and WER [%] (lower is better)) on CSJ_DEMAND_ST dataset. SXR denotes values encompassing SDR, SNR, and SAR.

TABLE VI
RESULTS OF OA POST-PROCESSING AND AB-SDR TRAINING LOSS (WER [%] (LOWER IS BETTER)) FOR REAL RECORDINGS (CHIME-3'S ET05_REAL).

Id	System	Loss	ω_{OA}	WER
(1)	noisy	–	–	28.7
(2)	Enhan_T	SNR	–	31.7
(3)		SDR	–	30.8
(4)		AB-SDR	–	**22.3**
(5)	+OA	SNR	0.4	22.5
(6)		SDR	0.5	22.5
(7)		AB-SDR	0.0	22.3

The figure shows a similar trend to the results on the WSJ_CHIME_ST dataset in Section VII-A1: 1) by increasing ω_{OA}, the SAR monotonically increases at the expense of decreasing the SNR and 2) by adjusting ω_{OA} around 0.5, the modified enhanced signals with OA achieve better WER than both the baseline observed (gray line) and enhanced (black square) signals. These results further increase the experimental evidence of our hypothesis and show that it would hold for different evaluated datasets.

C. Evaluation with real recordings

To confirm the effectiveness of our findings even for real recordings, we evaluated the OA post-processing and the artifact-boosted loss training for the real-recorded evaluation data, i.e., "et05_real" of the CHiME-3 dataset. Table VI shows the WER scores of the evaluated SE systems trained with different loss functions (SNR, SDR, and AB-SDR) with/without OA post-processing (Enhan_T and Enhan_T+OA). From the table, we observe that both OA post-processing and artifact-boosted loss training, which were shown to improve the SAR (see Sections VII-A and VII-B), are also effective at improving the WER even when they are applied to real recordings, e.g., by comparing (2) with (5) and comparing (2) with (4), respectively. These results suggest that the findings based on the orthogonal projection-based analysis of the simulated dataset would hold for real recordings, i.e., mitigating the artifact errors would be a key to improving the ASR performance with the single-channel SE front-ends.

IX. Conclusion

In this paper, we analyzed the processing distortion problem induced by single-channel SE that would limit the potential improvement of ASR performance. We proposed a novel analysis scheme that applied orthogonal projection-based decomposition of SE errors and experimentally found a reasonable signal-level numerical metric, i.e., artifact errors, that could explain the cause of ASR performance degradation.

In addition, we explored two practical approaches to mitigating the effect of artifact errors for improving ASR performance, i.e., OA post-processing and AB-SDR training loss. We proved that OA can monotonically increase the SAR value under mild conditions. Furthermore, we experimentally confirmed that both approaches successfully reduced artifact errors and improved the ASR performance of single-channel SE, even for real recordings. These results further support our finding that artifact errors have a larger impact on the degradation of ASR performance than residual interference and noise errors.

The signal-level interpretation and experimental findings of the processing distortion obtained in this paper will provide novel insights into designing better SE front-ends and open novel research directions for building single-channel noise-robust ASR. Future works should include extension of this analysis to different front-end (e.g., dereverberation) and back-end (e.g., speaker recognition) tasks, as well as to multichannel scenarios.

References

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] D. Yu and L. Deng, *Automatic speech recognition*. Springer, 2016.

[3] J. Li, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[6] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-field automatic speech recognition," *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2020.

[7] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015, pp. 504–511.

[8] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Interspeech*, 2018, pp. 1561–1565.

[9] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5210–5214.

[10] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 196–200.

[11] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6697–6701.

[12] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *International Workshop on Speech Processing in Everyday Environments (CHiME-5 Workshop)*, vol. 1, 2018.

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

[14] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6875–6879.

[15] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[16] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černockỳ, and D. Yu, "Neural target speech extraction: An overview," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.

[17] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[18] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001, pp. 749–752.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[20] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 436–443.

[21] S.-J. Chen, A. S. Subramanian, H. Xu, and S. Watanabe, "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," in *Interspeech*, 2018, pp. 1571–1575.

[22] M. Fujimoto and H. Kawai, "One-pass single-channel noisy speech recognition using a combination of noisy and enhanced features," in *Interspeech*, 2019, pp. 486–490.

[23] T. Menne, R. Schluter, and H. Ney, "Investigation into joint optimization of single channel speech enhancement and acoustic modeling for robust ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6660–6664.

[24] X. Chang, T. Maekaku, Y. Fujita, and S. Watanabe, "End-to-end integration of speech recognition, speech enhancement, and self-supervised learning representation," in *Interspeech*, 2022, pp. 3819–3823.

[25] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[26] J. Woo, M. Mimura, K. Yoshii, and T. Kawahara, "End-to-end music-mixed speech recognition," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2020, pp. 800–804.

[27] T. von Neumann, C. Boeddeker, L. Drude, K. Kinoshita, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "Multi-talker ASR for an unknown number of sources: Joint training of source counting, separation and ASR," in *Interspeech*, 2020, pp. 3097–3101.

[28] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, 2016.

[29] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[30] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[31] R. Lyon, "A computational model of binaural localization and separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 8, 1983, pp. 1148–1151.

[32] K. Iwamoto, T. Ochiai, M. Delcroix, R. Ikeshita, H. Sato, S. Araki, and S. Katagiri, "How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR," in *Interspeech*, 2022, pp. 5418–5422.

[33] C. Zorila and R. Doddipatla, "Speaker reinforcement using target source extraction for robust automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6297–6301.

[34] Q. Wang, I. L. Moreno, M. Saglam, K. Wilson, A. Chiao, R. Liu, Y. He *et al.*, "Voicefilter-lite: Streaming targeted voice separation for on-device speech recognition," *Interspeech*, pp. 2677–2681, 2020.
[35] H. Sato, T. Ochiai, M. Delcroix, K. Kinoshita, T. Moriya, and N. Kamo, "Should we always separate?: Switching between enhanced and observed signals for overlapping speech recognition," in *Interspeech*, 2021, pp. 1149–1153.
[36] H. Sato, T. Ochiai, M. Delcroix, K. Kinoshita, N. Kamo, and T. Moriya, "Learning to enhance or not: Neural network-based switching of enhanced and observed signals for overlapping speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6287–6291.
[37] C. Boeddeker, W. Zhang, T. Nakatani, K. Kinoshita, T. Ochiai, M. Delcroix, N. Kamo, Y. Qian, and R. Haeb-Umbach, "Convolutive transfer function invariant SDR training criteria for multi-channel reverberant speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8428–8432.
[38] S. Venkataramani, R. Higa, and P. Smaragdis, "Performance based cost functions for end-to-end speech separation," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2018, pp. 350–355.
[39] Y. Koizumi, S. Karita, A. Narayanan, S. Panchapagesan, and M. Bacchiani, "SNRi target training for joint speech enhancement and recognition," in *Interspeech*, 2022, pp. 1173–1177.
[40] D. B. Paul and J. Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*, 1992, pp. 357–362.
[41] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *LREC*, vol. 6, 2000, pp. 1–5.
[42] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.
[43] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.
[44] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.
[45] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes," *Computer Speech & Language*, vol. 46, pp. 605–626, 2017.
[46] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7009–7013.
[47] J. L. Roux, S. Wisdom, H. Erdogan, and J. Hershey, "SDR – half-baked or well done?" in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.
[48] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.
[49] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning (ICML)*, 2013, pp. 1310–1318.
[50] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera *et al.*, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," in *Interspeech*, 2020, pp. 2637–2641.
[51] K. Iwamoto, T. Ochiai, M. Delcroix, R. Ikeshita, H. Sato, S. Araki, and S. Katagiri, "How does end-to-end speech recognition training impact speech enhancement artifacts?" *arXiv preprint arXiv:2311.11599 (to appear in ICASSP 2024)*, 2023.
[52] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2022.
[53] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang *et al.*, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Interspeech*, 2016, pp. 2751–2755.
[54] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann *et al.*, "The Kaldi speech recognition toolkit," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.