Computer Speech and Language 95 (2026) 101813

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

An end-to-end integration of speech separation and recognition with self-supervised learning representation

Yoshiki Masuyama [a,b,*], Xuankai Chang [c], Wangyou Zhang [d], Samuele Cornell [c], Zhong-Qiu Wang [e], Nobutaka Ono [b], Yanmin Qian [d], Shinji Watanabe [c]

[a] Mitsubishi Electric Research Laboratories (MERL), USA
[b] Department of Computer Science, Tokyo Metropolitan University, Japan
[c] Language Technologies Institute, Carnegie Mellon University, USA
[d] Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[e] Department of Computer Science and Engineering, Southern University of Science and Technology, China

ARTICLE INFO

Keywords:
Speech separation
Automatic speech recognition
Self-supervised learning
Joint training
Multi-task learning

ABSTRACT

Multi-speaker automatic speech recognition (ASR) has gained growing attention in a wide range of applications, including conversation analysis and human–computer interaction. Speech separation and enhancement (SSE) and single-speaker ASR have witnessed remarkable performance improvements with the rapid advances in deep learning. Complex spectral mapping predicts the short-time Fourier transform (STFT) coefficients of each speaker and has achieved promising results in several SSE benchmarks. Meanwhile, self-supervised learning representation (SSLR) has demonstrated its significant advantage in single-speaker ASR. In this work, we push forward the performance of multi-speaker ASR under noisy reverberant conditions by integrating powerful SSE, SSL, and ASR models in an end-to-end manner. We systematically investigate both monaural and multi-channel SSE methods and various feature representations. Our experiments demonstrate the advantages of recently proposed complex spectral mapping and SSLRs in multi-speaker ASR. The experimental results also confirm that end-to-end fine-tuning with an ASR criterion is important to achieve state-of-the-art word error rates (WERs) even with powerful pre-trained models. Moreover, we show the performance trade-off between SSE and ASE and mitigate it with a multi-task learning framework with both SSE and ASR criteria.

1. Introduction

Recent advancements in deep learning have significantly improved the performance of single-speaker automatic speech recognition (ASR) (Hinton et al., 2012; Li et al., 2017; Chiu et al., 2018; Radford et al., 2023). The end-to-end (E2E) framework simplifies the system and has demonstrated promising results. In the literature, various sequence-to-sequence machine learning techniques have been developed, such as the connectionist temporal classification (CTC) (Graves et al., 2006; Miao et al., 2015), the attention-based encoder–decoder (AED) (Chorowski et al., 2015; Chan et al., 2016), and the recurrent neural network transducer (Graves, 2012). Neural network architectures have also been explored, including Transformer (Karita et al., 2019) and Conformer (Gulati et al., 2020). Despite these advancements, there remains a large performance gap between single and multi-speaker conditions, especially in noisy and reverberant environments. In such "cocktail party" scenarios, speech separation and enhancement (SSE) is a

* Corresponding author at: Mitsubishi Electric Research Laboratories (MERL), USA.
E-mail address: yoshiki.masuyama@ieee.org (Y. Masuyama).

https://doi.org/10.1016/j.csl.2025.101813
Received 6 November 2024; Received in revised form 23 April 2025; Accepted 25 April 2025
Available online 14 May 2025

Table 1
Comparison of each module used in the proposed SIMO/MIMO-IRIS and relevant multi-speaker ASR systems with E2E training of SSE and ASR models. The system with † combines pre-trained models without fine-tuning (von Neumann et al., 2024). The system with ‡ uses the transducer for ASR (Kanda et al., 2023) while the others are based on joint CTC/AED (Watanabe et al., 2017).

	SSE	Feature representation	ASR
Monaural			
Settle et al. (2018)	Time–frequency masking	log mel-filterbank	BLSTM
von Neumann et al. (2020b)	ConvTasNet	log mel-filterbank	BLSTM
Shi et al. (2020)	Conditional TasNet	log mel-filterbank	Transformer
von Neumann et al. (2024)†	TF-GridNet	WavLM	Conformer
SIMO-IRIS (Proposed)	TF-GridNet	WavLM	Conformer
Multi-channel			
Chang et al. (2019)	Mask-based MVDR	log mel-filterbank	BLSTM
Chang et al. (2020)	Mask-based MVDR	log mel-filterbank	Transformer
Zhang et al. (2022)	Mask-based WPE+MVDR	log mel-filterbank	Transformer
Kanda et al. (2023)‡	VarArray+MVDR	log mel-filterbank	Transformer
MIMO-IRIS (Proposed)	TF-GridNet	WavLM	Conformer

key component to tackle a multi-speaker recording with an ASR system (Carletta et al., 2005; Barker et al., 2018; Watanabe et al., 2020; Cornell et al., 2023b; Liang et al., 2023).

At the same time, in the last decade, SSE has remarkably progressed with the adoption of supervised deep learning approaches, most notably the permutation invariant training (PIT) (Yu et al., 2017b). PIT allows to train neural networks so that the predicted time–frequency masks are close to the ideal masks in the short-time Fourier transform (STFT) domain Yu et al. (2017b), Wang and Chen (2018). The encoder-masker-decoder approach, e.g., ConvTasNet (Luo and Mesgarani, 2019), typically leads to better performance by applying the masks in trainable latent space (Luo and Mesgarani, 2019; Luo et al., 2020; Subakan et al., 2021). Another approach is complex spectral mapping, where neural networks directly predict the real and imaginary parts of the STFT coefficients of each speaker. Recently, this approach outperforms the aforementioned approaches (Wang et al., 2023b,a; Tan et al., 2022; Cornell et al., 2023a; Pan et al., 2023). These SSE models are, however, trained to minimize signal-level differences between the target and separated signals (Luo and Mesgarani, 2019; Luo et al., 2020). As such, the overall recognition performance may be sub-optimal or even degraded, if these SSE models are used as a front-end of ASR applications.

The SSE process typically introduces artifacts that degrade recognition performance when the SSE and ASR systems are pre-trained separately (Koizumi et al., 2022; Iwamoto et al., 2022). To mitigate this issue, E2E integration of SSE and ASR via their joint training is an active research direction (Seltzer et al., 2004; Li et al., 2016; Heymann et al., 2017; Ochiai et al., 2017; Minhua et al., 2019; Chang et al., 2019, 2020; von Neumann et al., 2020b; Zhang et al., 2022). For instance, early research showed promising results in the context of single-speaker robust ASR by integrating a neural beamformer and a joint CTC/AED model (Ochiai et al., 2017). This integration has been extended to multi-speaker settings, including MIMO-Speech (Chang et al., 2019). MIMO-Speech explicitly separates the observed mixture with beamforming, while the entire system is trained only by the multi-speaker ASR criterion. Such an approach preserves the modularity of the entire system and is able to output intermediate separated signals in contrast to fully E2E black-box approaches (Yu et al., 2017a; Seki et al., 2018; Meng et al., 2023; Kanda et al., 2020; Sklyar et al., 2021). This approach has been successfully developed in both monaural (Settle et al., 2018; Chang et al., 2019; Shi et al., 2020) and multi-channel (Chang et al., 2019, 2020; Zhang et al., 2022; Kanda et al., 2023) scenarios where the existing systems typically used classical filterbanks as feature representation for ASR as summarized in Table 1.

Apart from the progress in SSE and its E2E integration with ASR, self-supervised learning (SSL) models have shown promising performance improvements in single-speaker ASR (Yang et al., 2021; Tsai et al., 2022). SSL aims to learn useful feature representation (SSLR) by solving a pretext task defined without manual labels. Many SSL models, including Wav2Vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021), have been pre-trained only on clean single-speaker speech, and thus their strength under noisy reverberant conditions is limited (Chang et al., 2021). More recently, SSL models trained on noisy speech (Chen et al., 2022; Wang et al., 2022a,b; Huang et al., 2022) and on multi-speaker speech (Fazel-Zarandi and Hsu, 2023; Huang et al., 2023a) have been investigated. Among these models, WavLM (Chen et al., 2022), a robust variant of HuBERT, is particularly effective. It has achieved state-of-the-art (SOTA) performance on SUPERB benchmark tasks (Yang et al., 2021) and demonstrated promising recognition performance even in noisy reverberant environments (von Neumann et al., 2024; Cornell et al., 2024).

By fusing WavLM into the E2E integration of speech enhancement and ASR, the performance of single-speaker robust ASR has been significantly improved (Chang et al., 2022; Masuyama et al., 2023). IRIS (Chang et al., 2022) integrates ConvTasNet, WavLM, and the joint CTC/AED model, achieving SOTA performance on the CHiME-4 single-channel track. Multi-IRIS (Masuyama et al., 2023) extends IRIS to perform multi-channel speech enhancement by neural beamforming. These works focused only on single-speaker scenarios, and here we consider a multi-speaker extension for "cocktail party" scenarios.

In this paper, we advance the performance of robust multi-speaker ASR by proposing SIMO-/MIMO-IRIS: an E2E integration of monaural/multi-channel SSE, SSLR extraction, and ASR. Our systems are based on the literature of the E2E training of SSE and ASR models, but we leverage recently proposed powerful SSE and SSL models as summarized in Table 1. Specifically, we perform complex spectral mapping by using TF-GridNet (Wang et al., 2023a,b) and extract feature representation from WavLM (Chen et al.,

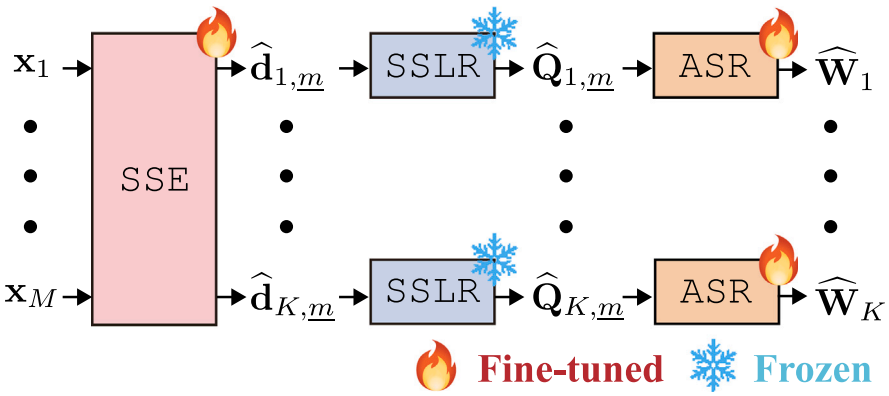

Fig. 1. Overview of SIMO/MIMO-IRIS via an E2E integration of SSE, SSLR extraction, and ASR. The input of the SSE model can be multi-channel as in MIMO-IRIS or monaural (only $\mathbf{x}_1$) as in SIMO-IRIS.

2022). The SSE model is pre-trained individually and then combined with SSL and ASR models. Afterwards, both the front-end (the SSE model) and back-end (the ASR model) are fine-tuned together based on an ASR criterion as illustrated in Fig. 1. This allows to simultaneously optimize the front-end for the subsequent ASR and adapt the back-end to the imperfect SSE outputs. We perform a comprehensive evaluation under anechoic/reverberant and clean/noisy conditions by using the spatialized WSJ0-2mix (Wang et al., 2018) and WHAMR! datasets (Maciejewski et al., 2020). We explore various SSE methods, including time–frequency masking and neural beamforming, and feature representations. The results confirm the advantage of the E2E integration of a strong TF-GridNet-based SSE front-end and the WavLM-based ASR back-end. The fine-tuning only with the ASR criterion degrades the signal-level separation quality while improving the word error rate (WER) as mentioned in von Neumann et al. (2020b). We explore this trade-off by following a multi-task learning framework with SSE and ASR criteria and perform a detailed analysis of the trade-off. Our training recipes and model checkpoints will be released through the end-to-end speech processing (ESPnet) toolkit (Watanabe et al., 2018; Li et al., 2021; Lu et al., 2022).

In our previous work (Masuyama et al., 2023), we introduced MIMO-IRIS and verified its efficacy with preliminary experiments. This paper extends the previous work by making substantial contributions in the following areas:

- developing SIMO-IRIS that integrates monaural SSE and SSLR-based ASR and performs the joint training;
- conducting an extensive evaluation of both SIMO-IRIS and MIMO-IRIS under various conditions;
- exploring various SSLRs and classical filterbank features in the multi-speaker ASR system;
- investigating the trade-off between separation and recognition performance in the E2E fine-tuning through a multi-task learning framework.

This paper is organized as follows. Section 2 reviews various monaural and multi-channel SSE methods under noisy reverberant conditions. Section 3 presents the proposed SIMO-/MIMO-IRIS and clarifies its relationship with other multi-speaker ASR frameworks. Sections 4 and 5 describe the experimental investigations of SIMO-IRIS and MIMO-IRIS, respectively. In Section 6, we draw conclusions and highlight future research directions.

2. Review of monaural and multi-channel SSE

2.1. Problem settings

Let an audio mixture with K speakers and noise be observed by M microphones under reverberant conditions. The mixing process of the observed time-domain signal $\mathbf{x}_m$ of length L can be formulated as

$$\begin{aligned}\mathbf{x}_m &= \sum_{k=1}^{K} \mathbf{y}_{k,m} + \mathbf{n}_m \\ &= \sum_{k=1}^{K} (\mathbf{d}_{k,m} + \mathbf{r}_{k,m}) + \mathbf{n}_m \\ &= \sum_{k=1}^{K} (\mathbf{h}_{k,m} \circledast \mathbf{s}_k + \mathbf{r}_{k,m}) + \mathbf{n}_m,\end{aligned} \tag{1}$$

where $\mathbf{y}_{k,m}$ is the reverberant source image of source k, $\mathbf{n}_m$ is the background noise signal, and $m = 1, \dots, M$ and $k = 1, \dots, K$ are the microphone and source indices, respectively. Each source image $\mathbf{y}_{k,m}$ is decomposed into the desired source image $\mathbf{d}_{k,m}$ and the

undesired late reverberation $\mathbf{r}_{k,m}$. In (1), the desired source image is further modeled by the convolution of the impulse response $\mathbf{h}_{k,m}$ and the dry source signal $\mathbf{s}_k$, where $\circledast$ denotes the convolution.

Let us denote the STFT coefficients of $\mathbf{x}_m$ by $\mathbf{X}_m = \mathtt{STFT}(\mathbf{x}_m) \in \mathbb{C}^{T\times F}$, where T and F are the number of time frames and frequency bins, respectively. The mixing process in (1) can be reformulated as follows:

$$\mathbf{X}_m = \sum_{k=1}^{K} (\mathbf{D}_{k,m} + \mathbf{R}_{k,m}) + \mathbf{N}_m, \tag{2}$$

where $\mathbf{D}_{k,m}$, $\mathbf{R}_{k,m}$, and $\mathbf{N}_m$ are the STFT coefficients of the desired source image, late reverberation, and noise, respectively. With the reference microphone $\underline{m} \in \{1, \dots, M\}$, the (t, f)th entry of $\mathbf{D}_{k,m}$ is approximated by

$$D_{k,m}(t, f) = a_{k,m}(f) D_{k,\underline{m}}(t, f), \tag{3}$$

where $\mathbf{a}_k(f) = [a_{k,1}(f), \dots, a_{k,M}(f)]^{\mathsf{T}}$ is the relative transfer function (Gannot et al., 2001), and $(\cdot)^{\mathsf{T}}$ denotes the transpose.

On the basis of the mixing process in (2), our SSE model aims to estimate the desired source image in the time–frequency domain:

$$\{\widehat{\mathbf{D}}_{1,\underline{m}}, \dots, \widehat{\mathbf{D}}_{K,\underline{m}}\} = \mathtt{SSE}(\mathbf{X}_1, \dots, \mathbf{X}_M), \tag{4}$$

where we need to not only separate each speaker but also suppress the noise and late reverberation. That is, the SSE model performs denoising and dereverberation along with separation. Building up on SSE, our goal is to predict the transcription sequence for each speaker $\mathbf{W}_k$ from the mixture:

$$\{\widehat{\mathbf{W}}_1, \dots, \widehat{\mathbf{W}}_K\} = \mathtt{MultiSpeakerASR}(\mathbf{X}_1, \dots, \mathbf{X}_M), \tag{5}$$

where we do not care for the speaker order of the predicted transcripts.

Throughout this paper, we assume that the array geometry, including the number of microphones M, and the number of speakers K are known and fixed during both training and inference. While the integration of SSE and ASR with variable numbers of microphones (Kanda et al., 2023) and speakers (von Neumann et al., 2020a) is interesting, we defer such directions to future work. We instead focus on demonstrating comprehensive analysis of multi-speaker ASR under static conditions.

2.2. Monaural SSE

When $M = 1$ and the input of SSE is only $\mathbf{x}_1$, it is called monaural SSE. For monaural SSE in the time–frequency domain, time–frequency masking and complex spectral mapping have been widely used (Wang and Chen, 2018). In the former masking approach, a neural network is used to estimate a mask $\widehat{\mathbf{G}}_k \in \mathbb{C}^{T\times F}$ for each speaker k, and then the mask is applied to the STFT of the mixture:

$$\{\widehat{\mathbf{G}}_{1,\underline{m}}, \dots, \widehat{\mathbf{G}}_{K,\underline{m}}\} = \mathtt{MaskNet}(\mathbf{X}_{\underline{m}}), \tag{6}$$

$$\widehat{\mathbf{D}}_{k,\underline{m}} = \widehat{\mathbf{G}}_{k,\underline{m}} \odot \mathbf{X}_{\underline{m}}, \tag{7}$$

where $\underline{m} = 1$ by definition, and $\odot$ denotes the Hadamard product. While the time–frequency mask is usually restricted to non-negative real value, it can be complex (Williamson et al., 2016). The separated STFT coefficients are then converted to the time domain by the inverse STFT (iSTFT):

$$\widehat{\mathbf{y}}_{\underline{m}} = \mathtt{iSTFT}(\widehat{\mathbf{Y}}_{\underline{m}}). \tag{8}$$

In the more modern encoder-masker-decoder approach, STFT and iSTFT are replaced by trainable 1D convolution and deconvolution layers, respectively (Luo and Mesgarani, 2019; Luo et al., 2020; Subakan et al., 2021).

Complex spectral mapping directly predicts the complex STFT coefficients of each speaker $\widehat{\mathbf{D}}_{k,\underline{m}}$ instead of the mask $\widehat{\mathbf{G}}_{k,\underline{m}}$ (Tan and Wang, 2020; Wang et al., 2020; Yang et al., 2022):

$$\{\widehat{\mathbf{D}}_{1,\underline{m}}, \dots, \widehat{\mathbf{D}}_{K,\underline{m}}\} = \mathtt{MappingNet}(\mathbf{X}_{\underline{m}}). \tag{9}$$

This approach has less restriction on the output and demonstrates comparable or better performance than the encoder-masker-decoder approach (Cornell et al., 2023a; Pan et al., 2023).

2.3. Multi-channel SSE

Multi-channel SSE exploits spatial information obtained from multiple microphones. It is typically realized in the time–frequency domain because we can efficiently implement spatial filtering with the narrow-band approximation (Gannot et al., 2017). As a front-end for ASR, mask-based beamforming has demonstrated promising results (Heymann et al., 2016; Erdogan et al., 2016), where the non-negative time–frequency masks $\{\widehat{\mathbf{G}}_{1,\underline{m}}, \dots, \widehat{\mathbf{G}}_{K,\underline{m}}\}$ in (6) are used to build the beamformers.

In mask-based beamforming, a concatenation of STFT of the M-channel mixture $\boldsymbol{\chi}(t,f) = [X_1(t,f),\dots,X_M(t,f)]^{\mathsf{T}} \in \mathbb{C}^M$ is considered, where $X_m(t,f)$ is the (t,f)th entry of $\mathbf{X}_m$. We first compute the spatial covariance matrix of the each speaker $\widehat{\mathbf{V}}_k(f)$ and their respective interference $\widehat{\mathbf{V}}_{\backslash k}(f)$ as follows:

$$\widehat{\mathbf{V}}_k(f) = \frac{1}{\sum_{t=1}^{T} \widetilde{G}_k(t,f)} \sum_{t=1}^{T} \widetilde{G}_k(t,f)\boldsymbol{\chi}(t,f)\boldsymbol{\chi}(t,f)^{\mathsf{H}}, \tag{10}$$

$$\widehat{\mathbf{V}}_{\backslash k}(f) = \left(\sum_{k'=1}^{K+1} \widehat{\mathbf{V}}_{k'}(f)\right) - \widehat{\mathbf{V}}_k(f), \tag{11}$$

$$\widetilde{G}_k(t,f) = \frac{1}{M}\sum_{m=1}^{M} \widehat{G}_{k,m}(t,f), \tag{12}$$

where $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose, and $\widehat{\mathbf{V}}_{K+1}$ represents the spatial covariance matrix of the noise. Then, the beamformers are constructed based on well-studied criteria such as the minimum variance distortionless response (MVDR) beamformer (Souden et al., 2010) and the generalized eigenvalue beamformer (Warsitz and Haeb-Umbach, 2007).

The MVDR beamformer can be implemented by leveraging the spatial covariance matrix of each speaker instead of the relative transfer function (Souden et al., 2010). Specifically, the spatial filter $\widehat{\mathbf{w}}_k(f)$ can be expressed as:

$$\widehat{\mathbf{w}}_k(f) = \frac{\widehat{\mathbf{V}}_{\backslash k}^{-1}(f)\widehat{\mathbf{V}}_k(f)}{\mathrm{trace}\left(\widehat{\mathbf{V}}_{\backslash k}^{-1}(f)\widehat{\mathbf{V}}_k(f)\right)}\mathbf{u}, \tag{13}$$

where $\mathrm{trace}(\cdot)$ denotes the matrix trace, and $\mathbf{u} \in \mathbb{R}^M$ is a one-hot vector indicating the reference microphone $\underline{m}$. Then, the beamformer is applied to the STFT of the mixture, and the results are converted to the time domain via iSTFT:

$$\widehat{D}_{k,\underline{m}}(t,f) = \widehat{\mathbf{w}}_k^{\mathsf{H}}(f)\boldsymbol{\chi}(t,f), \tag{14}$$

$$\widehat{\mathbf{d}}_{k,\underline{m}} = \mathtt{iSTFT}(\widehat{\mathbf{D}}_{k,\underline{m}}). \tag{15}$$

The process of the mask-based beamforming is differentiable, and thus we can optimize `MaskNet` in (6) with a loss function defined on the separated signals $\{\widehat{\mathbf{d}}_{1,\underline{m}},\dots,\widehat{\mathbf{d}}_{K,\underline{m}}\}$. The main advantage of mask-based beamforming is the fact that the obtained spatial filter is linear and based on well-studied array signal processing principles. These properties mitigate artifacts (Iwamoto et al., 2022) and improve the robustness and generalization capability (Chang et al., 2019; Zhang et al., 2021a; Masuyama et al., 2023).

Since the MVDR beamformer is mainly designed for separation and denoising, we need additional dereverberation under reverberant conditions. The weighted prediction error (WPE) (Nakatani et al., 2010) has been widely used to suppress the late reverberation in (2) (Kinoshita et al., 2016; Barker et al., 2018). WPE estimates the late reverberation by using an inverse filter and subtracts it. It is typically performed before beamforming, and its integration with beamforming has been explored (Nakatani and Kinoshita, 2019; Zhang et al., 2022).

Complex spectral mapping in (9) can be easily extended to multi-channel SSE as follows (Wang et al., 2020, 2021b; Tan et al., 2022):

$$\{\widehat{\mathbf{D}}_{1,\underline{m}},\dots,\widehat{\mathbf{D}}_{K,\underline{m}}\} = \mathtt{MappingNet}(\mathbf{X}_1,\dots,\mathbf{X}_M). \tag{16}$$

In (16), `MappingNet` implicitly performs time-varying nonlinear spatial filtering. Compared with the time-invariant beamformers, the output may reduce the interference more but can introduce more artifacts on the separated speech. Although prior studies showed that the time-invariant beamformers would be preferable for ASR (Chang et al., 2019; Zhang et al., 2021a), we expect that the combination of modern front-ends and back-ends can address the issue of artifacts.

2.4. Loss functions for SSE

In this subsection, we briefly review popular signal-level loss functions for SSE. During the training of an SSE model, the order of the separated signals $\widehat{\mathbf{d}}_{1,m},\dots,\widehat{\mathbf{d}}_{K,\underline{m}}$ is unconstrained and can be different from the original order. To compute the signal-level loss functions, however, we need to assign each estimate $\widehat{\mathbf{d}}_{k',m}$ to one of the desired source images $\mathbf{d}_{k,m}$. To address this problem, PIT computes all the possible permutations and uses the best assignment as follows (Kolbæk et al., 2017):

$$\mathcal{L}_{\mathrm{SSEPIT}} = \min_{\boldsymbol{\pi}\in\mathcal{P}_K} \sum_{k=1}^{K} \mathcal{L}(\mathbf{d}_{\pi_k}, \widehat{\mathbf{d}}_k), \tag{17}$$

where $\mathcal{P}_K$ is the set of $K!$ possible permutations on $\{1,\dots,K\}$, and π_k is the kth entry of the permutation. In (17), $\mathcal{L}$ denotes a loss function computed for each pair of signals.

Regarding the pair-wise loss function in (17), various loss functions have been developed. The scale-invariant signal-to-distortion ratio (SI-SDR) has been widely used not only for evaluation but also as a loss function:

$$\mathcal{L}_{\text{SI-SDR}} = 20\log_{10}\left(\frac{\|\alpha\mathbf{d}\|_2}{\|\alpha\mathbf{d}-\widehat{\mathbf{d}}\|_2}\right), \tag{18}$$

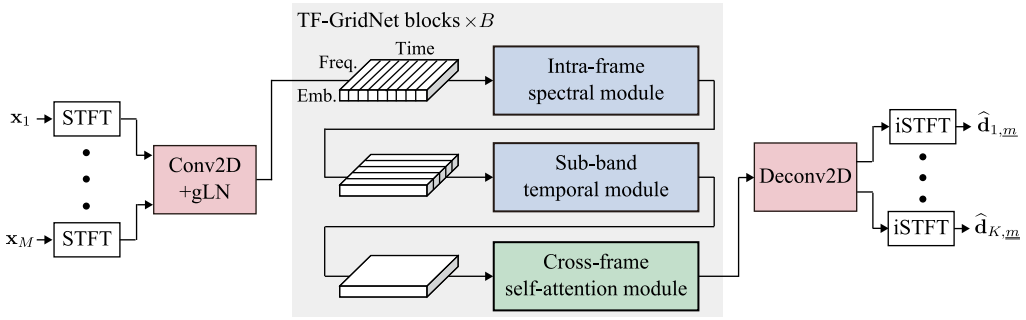

Fig. 2. Overview of TF-GridNet. Here, gLN stands for global layer normalization.

where $\alpha = \widehat{\mathbf{d}}^{\mathsf{T}}\mathbf{d}/\|\mathbf{d}\|_2^2$, and $\|\cdot\|_2$ denotes the ℓ_2 norm. In (18), α compensates for the scale of the target signal to fit to the estimate. While the SI-SDR loss is defined in the time domain, recent studies (Wang et al., 2023a) have shown the benefit of combining the time-domain and STFT-domain losses:

$$\mathcal{L}_{\text{MIX}} = \beta\left\|\mathbf{d} - \hat{\alpha}\widehat{\mathbf{d}}\right\|_1 + (1-\beta)\left\||\mathtt{STFT}(\mathbf{d})| - |\mathtt{STFT}(\hat{\alpha}\widehat{\mathbf{d}})|\right\|_1, \tag{19}$$

where $\hat{\alpha} = \widehat{\mathbf{d}}^{\mathsf{T}}\mathbf{d}/\|\widehat{\mathbf{d}}\|_2^2$, $|\cdot|$ computes magnitude in entry wise, and $\|\cdot\|_1$ denotes the ℓ_1 norm. Here, the estimate is rescaled instead of the target following (Ma et al., 2020; Wang et al., 2023a). In (19), the ℓ_1 norm of the STFT magnitude is computed for $|\mathtt{STFT}(\hat{\alpha}\widehat{\mathbf{d}})|$ instead of $|\hat{\alpha}\widehat{\mathbf{D}}|$. As a result, the second term takes into account the inconsistency issue in $\widehat{\mathbf{D}}$ (Masuyama et al., 2020; Wang et al., 2021c).

3. E2E multi-speaker ASR with SSLR

3.1. Overview of proposed SIMO- and MIMO-IRIS

Our multi-speaker ASR system consists of three modules: monaural/multi-channel SSE (SSE), SSLR extraction (SSLR), and E2E ASR (ASR), as illustrated in Fig. 1. The observed mixture $(\mathbf{x}_1, \ldots, \mathbf{x}_M)$ is first separated into K sources, where $M = 1$ in the monaural case. Then, we extract SSLR from each separated signal and feed it into E2E ASR as follows:

$$\{\widehat{\mathbf{D}}_{1,\underline{m}}, \ldots, \widehat{\mathbf{D}}_{K,\underline{m}}\} = \mathtt{SSE}(\mathbf{X}_1, \ldots, \mathbf{X}_M), \tag{20}$$

$$\widehat{\mathbf{Q}}_k = \mathtt{SSLR}(\mathtt{iSTFT}(\widehat{\mathbf{D}}_{k,\underline{m}})), \tag{21}$$

$$\widehat{\mathbf{W}}_k = \mathtt{ASR}(\widehat{\mathbf{Q}}_k), \tag{22}$$

where $\widehat{\mathbf{Q}}_k$ is the SSLR for the kth speaker. The modularity of our system allows to leverage powerful pre-trained models for each module. To recognize multi-speaker conversation, we can cascade pre-trained SSE and ASR models. This cascaded strategy without fine-tuning is not optimal because the ASR model is sensitive to the imperfect SSE outputs. To mitigate this issue, we optimize the entire multi-speaker ASR system in an E2E fashion using backpropagation.

3.2. Complex spectral mapping by TF-GridNet

While our system can use any SSE model, we leverage complex-spectral mapping to simultaneously perform separation, denoising, and dereverberation as it is seamlessly applicable to both monaural and multi-channel cases. Specifically, we use TF-GridNet that has shown SOTA performance in various SSE benchmarks (Hershey et al., 2016; Wang et al., 2018; Maciejewski et al., 2020; Cornell et al., 2023a; Pan et al., 2023). Its overview is illustrated in Fig. 2. The STFT of the observed mixture is expanded to C-dimensional embeddings for each time–frequency bin by using a 2D convolution layer followed by global layer normalization. The number of input channels of the 2D convolution layer is $2M$ by concatenating the real and imaginary parts for all the microphone channels. These embeddings are passed to B TF-GridNet blocks, each consisting of an intra-frame spectral module, a sub-band temporal module, and a cross-frame self-attention module. The real and imaginary parts of each speaker are obtained by applying a 2D deconvolution layer to the output of the final TF-GridNet block.

In the intra-frame spectral module, the embeddings of size $T \times F \times C$ are treated as T sequences of length F. For each sequence, we apply layer normalization and stack adjacent I embeddings along frequency with shift J. The resulting sequence of $C \times I$ dimensional features are passed to a bi-directional long short-term memory (BLSTM) layer and a 1D deconvolution layer. This module focuses on modeling spectral information at each time frame. Meanwhile, the sub-band temporal module views the embeddings as F sequences of length T and performs a similar procedure. This module handles the temporal information at each sub-band.

While the aforementioned modules handle the spectral and temporal relationships separately, the cross-frame self-attention module is designed to aggregate full-band information between distant time frames. In detail, we apply a point-wise 2D convolution

to the 2D embeddings and concatenate the embeddings for the frequency direction with layer normalization. The attention matrix is computed over the time frames, i.e., its size is $T \times T$. The attention outputs of multiple heads are concatenated for the channel direction, and then it is passed to another point-wise 2D convolution layer followed by layer normalization.

After B TF-GridNet blocks, the embeddings are passed to a 2D deconvolution layer with $2K$ output channels to obtain the real and imaginary parts of each speaker. Each estimate is converted back to the time domain as $\widehat{\mathbf{d}}_{k,\underline{m}}$ by iSTFT.

3.3. SSLR extraction by WavLM

Separated signals may contain residual noise, reverberation, and artifacts introduced by the SSE process. To robustly extract SSLRs, we use WavLM (Chen et al., 2022) that has shown remarkable performance in the SUPERB benchmark (Yang et al., 2021) and multi-speaker ASR (Cornell et al., 2023b; von Neumann et al., 2024). WavLM consists of a convolutional encoder block and L_{SSL} Transformer layers. During pre-training, the input speech is augmented with noise and interference speech, and the augmented speech is fed into the convolutional encoder block. Its output $\mathbf{Z}_{0,k}$ is passed to the Transformer layers after masking out some frames. We optimize the Transformer layers to predict the k-means cluster assignment of clean speech on the masked frames. Although this training strategy is similar to HuBERT, the training of WavLM uses multiple datasets (Kahn et al., 2020; Chen et al., 2021; Wang et al., 2021a) and significantly leverages data augmentation.

Following the SUPERB strategy (Yang et al., 2021), the output from each Transformer layer $\mathbf{Z}_{l,k}$ are averaged with trainable weights in our multi-speaker ASR system:

$$\widehat{\mathbf{Q}}_k = \sum_{l=0}^{L_{\mathrm{SSL}}} \gamma_l \mathbf{Z}_{l,k}, \tag{23}$$

where $l = 0$ indicates the convolution encoder output, $\gamma_l \in [0, 1]$ is the non-negative weight satisfying $\sum_{l=0}^{L_{\mathrm{SSL}}} \gamma_l = 1$. This trainable weight is jointly optimized with the following E2E ASR model while WavLM itself is frozen.

3.4. E2E ASR by joint CTC/AED

To predict the transcript of each speaker, we use the joint CTC/AED model (Watanabe et al., 2017) as a powerful E2E ASR framework. It leverages the CTC module with an encoder shared with AED, which enforces the alignment between the feature representation and transcription. In the joint CTC/AED model, SSLR from WavLM $\widehat{\mathbf{Q}}_k$ is passed to an encoder:

$$\widehat{\mathbf{E}}_k = \mathtt{ASREnc}(\widehat{\mathbf{Q}}_k), \tag{24}$$

where $\widehat{\mathbf{E}}_k = [\widehat{\mathbf{e}}_{1,k}, \dots, \widehat{\mathbf{e}}_{T',k}]$ is the encoder output, and $t' = 1, \dots, T'$ is the sub-sampled frame index. The CTC module predicts the posterior distribution of the alignment between the feature representation $\mathbf{Q}_k$ and the transcription $\mathbf{W}_k$:

$$p(\mathfrak{A}_k \mid \widehat{\mathbf{Q}}_k) \approx \prod_{t'=1}^{T'} p(\mathfrak{a}_{t',k} \mid \widehat{\mathbf{Q}}_k), \tag{25}$$

$$p(\mathfrak{a}_{t',k} \mid \widehat{\mathbf{Q}}_k) = \mathtt{CTC}(\widehat{\mathbf{e}}_{t',k}), \tag{26}$$

where $\mathfrak{A}_k = [\mathfrak{a}_{1,k}, \dots, \mathfrak{a}_{T',k}]$ is the alignment sequence, $\mathfrak{a}_{t',k}$ is an entry of the vocabulary or a blank token, and $p(\mathfrak{a}_{t',k} \mid \widehat{\mathbf{Q}}_k)$ denotes the frame-level posterior distribution. In (25), we introduce the conditional independence assumption between the frame-level outputs for the approximation. Then, the posterior distribution of the transcription is computed as follows (Graves et al., 2006):

$$p_{\mathrm{ctc}}(\mathbf{W}_k \mid \widehat{\mathbf{Q}}_k) = \sum_{\mathfrak{A}_k \in \mathcal{B}^{-1}(\mathbf{W}_k)} p_{\mathrm{ctc}}(\mathfrak{A}_k \mid \widehat{\mathbf{Q}}_k), \tag{27}$$

where $\mathcal{B}^{-1}(\mathbf{W}_k)$ indicates all the possible alignments compatible with the transcription $\mathbf{W}_k$.

Meanwhile, AED computes the posterior distribution without the conditional independence assumption:

$$p_{\mathrm{att}}(\mathbf{W}_k \mid \widehat{\mathbf{Q}}_k) = \prod_{\tau=1}^{\tau_k} p_{\mathrm{att}}(W_{\tau,k} \mid \widehat{\mathbf{Q}}_k, W_{1,k}, \dots, W_{\tau-1,k}), \tag{28}$$

$$p_{\mathrm{att}}(W_{\tau,k} \mid \widehat{\mathbf{Q}}_k, W_{1,k}, \dots, W_{\tau-1,k}) = \mathtt{ASRDec}(\widehat{\mathbf{Q}}_k, W_{1,k}, \dots, W_{\tau-1,k}), \tag{29}$$

where $\tau = 1, \dots, \tau_k$ is the index of tokens in the transcription for the kth speaker. During training, we use teacher forcing to condition the decoder. At the inference, we combine the posteriors from the CTC module and the decoder and apply beam search.

3.5. Training strategy of SIMO- and MIMO-IRIS

Training of a multi-speaker ASR system from scratch is computationally intensive and potentially leads to sub-optimal performance, as reported in previous studies (Chang et al., 2022; Masuyama et al., 2023). To address this issue, we first individually pre-train SSE, SSLR, and ASR, and then fine-tune SSE and ASR in an E2E manner. In detail, the SSE model is pre-trained with

the signal-level loss function in (19). Meanwhile, the ASR model is pre-trained on monaural clean speech datasets, e.g., the WSJ corpus, based on the following sum of two loss functions (Watanabe et al., 2017):

$$\mathcal{L}_{\mathrm{ASR}} = -\lambda \log p_{\mathrm{ctc}}(\mathbf{W}|\mathbf{Q}) - (1-\lambda) \log p_{\mathrm{att}}(\mathbf{W} \mid \mathbf{Q}), \tag{30}$$

where $\lambda \in [0, 1]$ is a hyperparameter for balancing two terms. In (30), we omit the source index because the pre-training is performed on a single-speaker dataset. While freezing the pre-trained WavLM to avoid overfitting, we optimize the trainable weights introduced in (23) together with the ASR model (Chang et al., 2022; Masuyama et al., 2023).

After pre-training, the integrated system is fine-tuned with the ASR criterion defined in (30) with PIT:

$$\mathcal{L}_{\mathrm{ASRPIT}} = \sum_{k=1}^{K} \mathcal{L}_{\mathrm{ASR}}(\mathbf{W}_{\pi_k^\star}, \widehat{\mathbf{Q}}_k), \tag{31}$$

$$\boldsymbol{\pi}^\star = \min_{\boldsymbol{\pi} \in \mathcal{P}_K} \sum_{k=1}^{K} -\log p_{\mathrm{ctc}}(\mathbf{W}_{\pi_k}, \widehat{\mathbf{Q}}_k), \tag{32}$$

where the CTC loss is used to determine the permutation $\boldsymbol{\pi}^\star$ that will be used in the subsequent computation of both terms in (30) for all the sources. This is because the CTC loss requires only K forward passes to determine the permutation, while the AED loss takes K^2 forward passes for considering every combination of the feature representation and the transcription for teacher forcing (Seki et al., 2018).

We can also include the SSE loss in (19) with PIT even during the fine-tuning according in a multi-task learning fashion:

$$\mathcal{L}_{\mathrm{MULTI}} = \mathcal{L}_{\mathrm{ASRPIT}}(\mathbf{W}_{1:K}, \widehat{\mathbf{Q}}_{1:K}) + \kappa \mathcal{L}_{\mathrm{SSEPIT}}(\mathbf{d}_{1:K}, \widehat{\mathbf{d}}_{1:K}), \tag{33}$$

where $\kappa \geq 0$ is a weight for the SSE loss, and the subscript $1:K$ indicates variables for all the sources. We choose the speaker permutation on the basis of the SSE loss as in (17) and use that permutation for the ASR loss (Settle et al., 2018; von Neumann et al., 2020b).

3.6. Relation to other multi-speaker ASR

Towards multi-speaker ASR, the classical strategy is to separate a mixture into single-speaker streams and then recognize each separated speech (Yoshioka et al., 2018a,b; Raj et al., 2021). This strategy employs the pre-trained SSE and ASR models in a cascade, which is not optimal as discussed in Section 3.5. This strategy has been extended to jointly train both SSE and ASR models (Qian et al., 2018; Settle et al., 2018; Chang et al., 2019; von Neumann et al., 2020b; Zhang et al., 2022). In particular, MIMO-Speech (Chang et al., 2019) integrates the mask-based beamformer and the joint CTC/AED model and trains the entire system with an ASR criterion in an E2E manner. Our multi-speaker ASR system advances this E2E integration by incorporating the robust SSL model, WavLM. We further provide a comprehensive analysis of various SSE strategies, including time–frequency masking, mask-based beamforming, and complex spectral mapping. Another relevant study (von Neumann et al., 2024) uses TF-GridNet and WavLM similarly to our work. The latter system, however, still cascades the pre-trained models and does not perform E2E training.

Another E2E strategy directly estimates multiple transcriptions without explicitly performing SSE (Yu et al., 2017a; Seki et al., 2018; Meng et al., 2023). While this strategy originally used PIT, the serialized output training (Kanda et al., 2020, 2022a; Sklyar et al., 2021; Li et al., 2023) and the heuristic error assignment training (L. Lu and Gong, 2021; Raj et al., 2022) have been developed to reduce computational complexity. This strategy has recently been extended to multi-channel cases (Kanda et al., 2023; Yifan et al., 2023). In contrast to these systems, our system preserves modularity and exploits pre-trained SSE and ASR models, which stabilizes the training of the combined multi-speaker ASR system. We can even pre-train each model on separate datasets.

Meanwhile, speaker-attributed ASR (Fiscus et al., 2007; Barker et al., 2018; Watanabe et al., 2020), which aims to jointly recognize "who said what", has gained much attention in conversation analysis (El Shafey et al., 2019; Kanda et al., 2022b; Cui et al., 2023; Cornell et al., 2024). Speaker-attributed ASR provides rich speaker attribution in addition to the transcription. On the other hand, our multi-speaker ASR system does not provide speaker attribution. It can be predicted by a subsequent diarization system or via continuous speech separation, which will be explored in future works.

Finally, SSL models handling multiple speakers have recently been developed (Fazel-Zarandi and Hsu, 2023; Huang et al., 2023a). These models are trained to predict the cluster assignments of each speaker from a mixture. In addition, WavLM has been adapted to extract SSLR of the target speaker by conditioning (Huang et al., 2023b). By feeding in the conditioning information of each speaker, the adapted SSL model can extract SSLRs for multi-speaker ASR. While our main experiments freeze the SSL model, we will also explore adapting the SSL model to multi-speaker ASR.

4. Experimental validation of SIMO-IRIS

We first evaluated SIMO-IRIS under noisy anechoic and reverberant conditions. Our experiments were conducted using the ESPnet-SE toolkit (Li et al., 2021; Lu et al., 2022) and S3PRL (Yang et al., 2021).

Table 2
WERs (%) on the clean WSJ corpus with different feature representation.

	dev93	eval92
Fbank	6.6	4.4
HuBERT Base	4.0	2.6
WavLM Base+	3.6	2.1
HuBERT Large	2.6	1.5
WavLM Large	**2.5**	**1.2**

4.1. Dataset

We used the WHAMR! dataset. It contains anechoic and reverberant two-speaker mixtures with noise recorded in urban environments (Maciejewski et al., 2020). Its dry source signals are from the WSJ0-2mix dataset (Hershey et al., 2016), but artificially simulated room impulse responses are convolved with each clean source signal to obtain reverberant signals. To exploit the pre-trained WavLM large model,[1] we used the source signals sampled at 16 kHz. The WHAMR! dataset provides two versions of overlapped mixtures: `min` and `max` versions. The longest of the two utterances is truncated in the `min` version, while the `max` version pads the shorter utterance by zero. We used the `min` version to pre-train the SSE model since it resulted in better separation performance compared with pre-training on the `max` version in our preliminary experiments. Meanwhile, the fine-tuning of the entire system was performed on the `max` version because the ASR loss in (31) requires the entire utterances of both speakers without truncation. We combined the anechoic and reverberant conditions during training and validation following (Zhang et al., 2022). For reference, the SI-SDRs of the input noisy mixtures are −4.5 dB and −6.1 dB under anechoic and reverberant conditions, respectively (Maciejewski et al., 2020).

For the pre-training of the ASR model, we used the WSJ corpus. Since it is also used as the source signal in the WHAMR! dataset, there is no domain mismatch regarding the speaking style or semantic content of the speech signal between the pre-training and fine-tuning data.

4.2. Experimental configuration

TF-GridNet was compared with a monaural SSE baseline, BLSTM-based time–frequency masking. STFT was implemented with the Hann window of 512 samples with a 256-sample shift. To estimate the time–frequency masks, we employed a 4-layer BLSTM along with time direction, where each layer has 600 units for each direction. Regarding the TF-GridNet, the kernel size of the initial 2D convolution and the final 2D deconvolution layers was 3, where the embedding dimension $C = 48$. The number of TF-GridNet blocks B was 6, where the BLSTM layer has 192 units in each block. We set I and J to 4 and 2, respectively. The cross-frame self-attention module leveraged 4 heads. The SSE models were pre-trained based on the combination of the time-domain and STFT domain losses in (19), where $\beta = 0.99$. The Adam optimizer was used with a learning rate of 0.001.

The ASR encoder consisted of 12 Conformer (Gulati et al., 2020) blocks, where each block had 4 attention heads with feed-forward layers of 2048 units. The kernel size of the convolution layers was set to 15. The ASR decoder consists of 6 Transformer (Karita et al., 2019) blocks with 4 attention heads. The dimensions of SSLR $\mathbf{Z}$ in (23) were reduced to 80 from 1024 by an additional feed-forward layer. The ASR model and the trainable weight in (23) were pre-trained on the WSJ corpus with the joint CTC/AED loss in (30). We used the Adam optimizer with a warmup scheduler, where the peak learning rate was 0.005. We performed model averaging over the 10 checkpoints with the highest accuracy.

Our entire system had 368 M parameters, where the TF-GridNet and the joint CTC/AED consisted of 8 M and 44 M parameters, respectively. The rest of the parameters came from the frozen WavLM large model. The SSE and ASR models were fine-tuned with the ASR criterion using stochastic gradient descent with a learning rate of 0.001 and a momentum of 0.9. This is because stochastic gradient descent typically performs better in fine-tuning (Zhou et al., 2020) and led to stable performance improvements in our previous study for robust single-speaker ASR (Masuyama et al., 2023). Training and inference scripts will be available through ESPnet.[2]

4.3. Experimental results with different feature representation

Before moving on to multi-speaker ASR, we investigate the recognition performance on the WSJ corpus with different feature representation: filterbanks, HuBERT, and WavLM. We explored two model sizes for HuBERT and WavLM. The recognition performance on the development and evaluation sets is summarized in Table 2. Consistent with the SUPERB benchmark results (Yang et al., 2021), SSLRs dramatically improved single-speaker recognition performance, and the WavLM large model performed the best.

Table 3 reports the WER as well as the substitution, deletion, and insertion errors for different feature representations in the monaural WHAMR! dataset. Here, the pre-trained TF-GridNet and the joint CTC/AED model are combined without fine-tuning, where

[1] https://huggingface.co/microsoft/wavlm-large
[2] https://github.com/espnet/espnet

Table 3
WERs (%) and their breakdown on the test set of monaural WHAMR! dataset with different feature representation. TF-GridNet is used for SSE.

	Noisy anechoic				*Noisy reverberant*			
	Sub.	Del.	Ins.	WER	Sub.	Del.	Ins.	WER
Fbank	17.3	2.1	10.1	29.5	17.4	1.7	10.2	29.3
HuBERT Base	12.4	1.6	7.5	21.5	11.9	1.3	6.4	19.5
WavLM Base+	9.3	1.5	5.8	16.7	8.7	1.2	4.3	14.2
HuBERT Large	8.0	1.3	**5.6**	14.9	7.2	1.0	**3.9**	12.1
WavLM Large	**6.1**	**1.1**	6.4	**13.6**	**5.8**	**0.9**	5.0	**11.6**

Table 4
Separation and recognition performance on the monaural WHAMR! test set with different training strategies. The shaded rows indicate the systems with the proposed E2E integration, i.e., SIMO-IRIS. During E2E fine-tuning, we used only the ASR loss except for the system with † that was fine-tuned with the multi-task learning framework in (33).

SSE model	Fine-tuned models	SDR	SIR	SAR	Sub.	Del.	Ins.	WER
Noisy anechoic								
Time–frequency masking	-	4.75	16.47	5.65	23.3	3.2	16.0	42.6
TF-GridNet	-	10.17	27.25	10.42	6.1	1.1	6.4	13.6
	ASR				4.0	0.9	1.9	6.7
	SSE, ASR	4.67	19.99	4.91	**2.3**	**0.4**	**0.8**	**3.6**
	SSE, ASR†	**11.92**	**29.81**	**12.04**	2.9	0.5	1.0	4.3
Noisy reverberant								
Ravenscroft et al. (2024)	SSE	-	-	-	-	-	-	26.7
Time–frequency masking	-	3.28	15.61	4.19	30.8	3.6	17.9	52.3
TF-GridNet	-	8.96	27.55	9.07	5.8	0.9	5.0	11.6
	ASR				3.6	0.7	1.3	5.7
	SSE, ASR	4.00	19.26	4.21	**2.2**	**0.3**	**0.6**	**3.1**
	SSE, ASR†	**10.77**	**29.21**	**10.86**	2.6	0.4	0.8	3.8

TF-GridNet results in SDRs of 10.17 dB and 8.96 dB under noisy anechoic and reverberant conditions, respectively. While the order of WERs is consistent with that on the clean WSJ corpus, the overall recognition performance was degraded even with TF-GridNet. This result shows that multi-speaker ASR under noisy conditions remains a challenge for a naïve cascaded method. Interestingly, all the systems achieved lower WERs under the reverberant condition, although TF-GridNet resulted in better SDR in the anechoic scenario. We suppose that this is because the residual interference and artifacts are masked out by the late reverberation. Such masking effect could improve the performance of subsequent ASR (Cord-Landwehr et al., 2022). In our informal listening tests, we heard the residual interference more clearly in the anechoic scenario.

4.4. Experimental results on monaural WHAMR!

We then compare our systems with different SSE models and training strategies. Table 4 summarizes both separation and recognition performance. Here, we also show WER of an existing method (Ravenscroft et al., 2024) that uses a time-domain SSE model fine-tuned with a loss defined on the ASR encoder output. Its WER was obtained by using Whisper (Radford et al., 2023). Even without fine-tuning, our combination of TF-GridNet, WavLM, and the Conformer-based joint CTC/AED model outperformed the existing system and the time–frequency masking baseline. This suggests that time–frequency masking might be too restrictive to restore clean speech, especially under noisy reverberant condition. The fine-tuning of the ASR model on the SSE outputs brought better recognition performance, notably reducing insertion errors. This result indicates that the fine-tuning stage mainly helps the ASR model to focus on the primary speaker in each separated signal, as observed in von Neumann et al. (2020b) with ConvTasNet for SSE. In the second bottom row, both the SSE and ASR models were fine-tuned with the ASR loss in (31), further improving recognition performance. This suggests that the fine-tuning of both models is essential for the best performing multi-speaker ASR system.

As shown in the second bottom row of Table 4, the fine-tuning of the SSE model degrades the separation performance, resulting in lower SDR and SAR even compared with the time–frequency masking baseline under the anechoic condition. TF-GridNet-based complex spectral mapping does not impose any constraints on its outputs and thus can easily generate artifacts. When SSE models are trained on signal-level loss functions, the generated artifacts degrade the recognition performance (Koizumi et al., 2022; Iwamoto et al., 2022). On the other hand, the artifacts improved WER in our case because our E2E fine-tuning enforces the SSE outputs to preserve critical information for subsequent ASR, as discussed in von Neumann et al. (2020b). The artifacts can be observed from examples of the separated signals shown in Fig. 3. Meanwhile, by incorporating the SSE loss as in (33) with $\kappa = 1.0$, the degradation

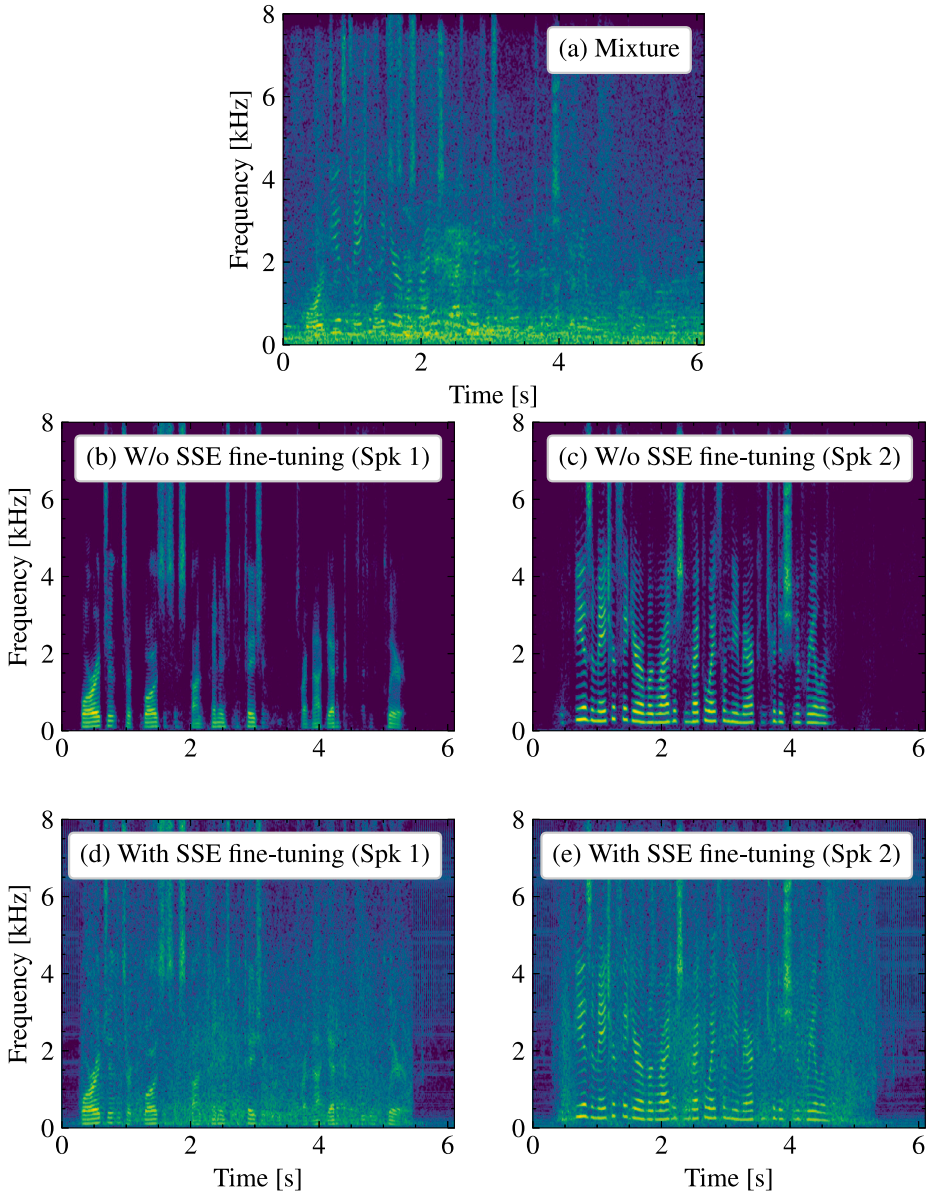

Fig. 3. Examples of log magnitude of the separated STFT coefficients with and without fine-tuning of the SSE model.

of separation metrics was mitigated[3] at the bottom row of Table 4. A similar trend was reported in a previous work using ConvTasNet for SSE and log-mel features instead of SSLR (von Neumann et al., 2020b). We therefore conclude that the degradation of the SSE performance due to fine-tuning only with an ASR criterion is mitigated by using also an SSE loss regardless of SSE models and feature representations.

Even with and without fine-tuning, the trainable weights in (23) for the SSLRs were mainly concentrated in the last layer. In detail, $\gamma_{L_{\mathrm{SSL}}}$ was 0.88 and 0.90 for with and without fine-tuning, respectively, which are similar to previous studies (Chang et al., 2021, 2022; Masuyama et al., 2023).

[3] The joint fine-tuning with the multi-task learning framework improved the separation performance even from the cascaded method in which the SSE model was pre-trained without the ASR loss. This is because the pre-training was performed on the `min` version and caused domain mismatch to the `max` version used in the fine-tuning and evaluation.

Table 5
Statistics of our two-speaker utterance groups.

		# of mixtures	Average duration (sec)	Total # of words
Train	SDM Original	19 605	5.0	328 800
	IHM-mix	19 528	5.1	328 640
Dev	SDM Original	2322	4.6	37 137
	IHM-mix	2315	4.6	37 123
Eval	SDM Original	2133	4.7	34 301

Table 6
WERs (%) and their breakdown on the two-speaker utterance groups in the development and evaluation sets of AMI SDM recordings. The scores for (Kanda et al., 2021) is from the original paper, where the score with † indicates that the training leveraged additional 900K multi-speaker mixtures simulated by using in-house recordings.

Fine-tuned models	*Development*				*Evaluation*			
	Sub.	Del.	Ins.	WER	Sub.	Del.	Ins.	WER
Kanda et al. (2021)	-	-	-	-	-	-	-	59.5
	-	-	-	-	-	-	-	19.6†
-	14.9	13.1	20.8	48.8	15.7	15.6	20.2	51.5
ASR	14.0	14.1	17.7	45.7	13.9	16.8	15.9	46.6
SSE, ASR	9.9	7.3	5.8	23.0	10.9	9.1	4.6	24.5

4.5. Experiment on two-speaker segments of AMI

To further investigate the performance in more realistic conditions, we tested SIMO-IRIS on the AMI dataset (Carletta et al., 2005). AMI consists of recordings of real long-form meetings, where each meeting has 3 to 5 participants. Here, we consider utterance-group based evaluation only (Kanda et al., 2023) and restrict ourselves to two speakers utterance groups. We followed the ESPnet AMI recipe (Watanabe et al., 2018) to partition the original AMI data into training, development, and evaluation sets. Afterwards, for each split, we extracted two-speaker utterance groups based on the given word-level segmentation annotation. The resulting dataset consists of sparsely overlapped single-channel noisy/reverberant mixtures of two speakers from a distant microphone (SDM). Following, Kanda et al. (2021), Cornell et al. (2024), we augmented the training data by generating mixtures from individual headset microphone (IHM) recordings. More in detail, we mixed headset recordings at two-speaker overlapped segments to obtain two-speaker mixtures. We used pyloudnorm (Steinmetz and Reiss, 2021) to align the loudness of the generated mixture and the corresponding SDM recording. We excluded utterance groups that were too short to compute the loudness. The statistics of these mixtures are summarized in Table 5.

The network architecture and training configurations are the same as in the previous experiment on the WHAMR! dataset. To construct SIMO-IRIS, we used the TF-GridNet model pre-trained on the WHAMR! dataset because there are no signal-level ground truth for the two-speaker utterance groups of SDM recordings of AMI. Although IHM provides relatively clean signals, these signals are typically still contaminated by their own noises and cross-talk speech (Wang et al., 2024). Training an SSE model on such imperfect target signals is still not an easy task (Maciejewski et al., 2023) and is out of the scope of this work. Meanwhile, the ASR model was pre-trained on the IHM recordings of AMI, which achieved WERs of 12.8% and 11.4% on the development and evaluation sets, respectively. Then, we fine-tuned only the ASR model or both SSE and ASR models with the ASR loss. The proposed E2E integration allows the adaptation of the entire system, including the SSE model, to realistic scenarios by using mixture recordings with transcripts. This is preferable because signal-level ground truth corresponding to a real mixture is difficult to obtain without a careful recording setup.

Table 6 shows the recognition performance on the two-speaker utterance groups from the development and evaluation sets. The system without the E2E fine-tuning resulted in poor performance. This is because TF-GridNet did not work well due to the domain mismatch between the WHAMR! and AMI datasets. Hence, the WERs were not improved significantly even with fine-tuning of the ASR model. By fine-tuning both SSE and ASR models, SIMO-IRIS achieved WER of 23.0% and 24.5% on the development and evaluation sets, respectively. While SIMO-IRIS still lags behind the best-performing system that was trained with huge amount of in-house recordings, these experiments confirm the advantage of the E2E fine-tuning with the ASR loss on real meeting scenarios.

5. Experimental validation of MIMO-IRIS

In this section, we demonstrate the efficacy of MIMO-IRIS under various conditions. Then, we explore multi-task learning during the joint fine-tuning of SSE and ASR models.

5.1. Dataset

We used the spatialized WSJ0-2mix dataset (Wang et al., 2018) as a dataset of multi-channel two-speaker mixtures without noise. It convolved the simulated room impulse responses with the dry source signals provided in WSJ0-2mix dataset, and we used the

Table 7
Separation and recognition performance on the two-channel reverberant test set of spatialized WSJ0-2mix. The shade indicates the system with the proposed E2E integration, i.e., MIMO-IRIS, where the systems were fine-tuned only with the ASR loss.

SSE model	Fine-tuned models	SDR	SIR	SAR	Sub.	Del.	Ins.	WER
Zhang et al. (2020b)	-	-	-	-	-	-	-	25.3
WPE-MVDR	-	4.77	8.75	8.40	20.2	0.6	27.3	48.1
	SSE, ASR	4.08	8.64	7.17	9.1	1.5	3.2	13.8
TF-GridNet	-	**15.06**	**31.29**	**15.21**	**1.3**	**0.1**	0.6	2.0
	SSE, ASR	6.75	25.98	6.82	**1.3**	**0.1**	**0.3**	**1.6**

first two channels as observations. In this dataset, SDR with respect to the input mixture is 0.07 dB. We combined both anechoic and reverberant mixtures during training as in Zhang et al. (2022).

The two-channel version of the WHAMR! dataset was also used to investigate the performance of MIMO-IRIS in challenging noisy environments. The real noise in the WAHMR! dataset was recorded in two-channels (Maciejewski et al., 2020), and thus we can add the noise to the simulated clean two-channel mixtures.

5.2. Experimental configuration

TF-GridNet was compared with a multi-channel SSE baseline employing mask-based beamforming. For the mask-based beamforming, we computed the spatial covariance matrix of each speaker by using time–frequency masks as in (10) and used the MVDR beamformer as formulated in (13). The time–frequency masks were predicted by a 3-layer BLSTM, where each layer has 512 units for each direction (forward/backward along the frame dimension). In addition, we performed dereverberation by using WPE before feeding the mixture to the beamformer. WPE leveraged a time–frequency mask calculated by another 3-layer BLSTM to compute the spatial filter following (Zhang et al., 2020a). TF-GridNet was adapted to the two-channel input by increasing the number of input channel of the initial convolution layer from 2 to 4. Both multi-channel SSE front-ends were pre-trained with the loss function in (19).

The ASR model was pre-trained on the WSJ corpus similar to the monaural case, and then we fine-tuned the system with (31). We used the same configuration as in the previous experiment.

5.3. Experimental results on spatialized WSJ0-2mix

The separation and recognition performance on the two-channel reverberant test set of the spatialized WSJ0-2mix dataset is summarized in Table 7. We include the score of an existing time-domain method (Zhang et al., 2020b). Although our mask-based beamforming lagged behind the existing method in terms of WER, its E2E fine-tuning resulted in better WER with a small degradation in separation performance. That is, the E2E fine-tuning is beneficial in multi-channel scenarios, similar to monaural cases. TF-GridNet without joint training consistently outperformed the mask-based beamformer. That is, the unconstrained complex spectral mapping is advantageous for ASR when the number of microphones is limited. In contrast to the mask-based beamforming, we observed a severe separation performance degradation in the E2E fine-tuning with TF-GridNet. This is because TF-GridNet can easily generate artifacts through time-varying nonlinear filtering, whereas the mask-based MVDR beamforming is a distortionless linear filtering. We emphasize that the E2E integration of TF-GridNet and the Conformer-based joint CTC/AED model achieved WER of 1.6%, comparable to the performance on the clean WSJ corpus in Table 2.

5.4. Experimental results on two-channel WHAMR!

Table 8 shows the separation and recognition performance on the two-channel WHAMR! dataset. Compared with existing methods based on mask-based beamforming (Zhang et al., 2022) and time-domain target speaker extraction (Zhang et al., 2021b), our combination of TF-GridNet, WavLM, and Conformer-based joint CTC/AED model dramatically improved WER. It also outperformed SIMO-IRIS in Table 4 by successfully leveraging spatial information. We then fine-tuned the ASR model while freezing the SSE model, obtaining a WER of 3.6% and 3.9% under anechoic and reverberant conditions, respectively. By fine-tuning WavLM in addition to the ASR model, we further improved WER. These fine-tuning strategies are advantageous for the front-end because we can preserve the original SSE model.

Our MIMO-IRIS with the E2E fine-tuning of both SSE and ASR models decreased the separation performance but yielded WERs of 2.3% under both anechoic and reverberant conditions. The fine-tuning of the SSE model is more beneficial than that of the SSL model for multi-speaker ASR. Finally, we further fine-tuned the entire model, including WavLM, after fine-tuning the SSE and ASR models. Its performance gain was marginal, which highlights the importance of fine-tuning the SSE model. We emphasize that the WER of 2.3% under the noisy reverberant condition is remarkable because the input SI-SDR is around −6.1 dB, demonstrating the potential of E2E integration of modern SSE and ASR models.

Table 8
Separation and recognition performance on the two-channel WHAMR! test set with different training strategies. The shaded rows are for MIMO-IRIS. The fine-tuning of the systems was performed only with the ASR loss.

SSE model	Fine-tuned models	SDR	SIR	SAR	Sub.	Del.	Ins.	WER
Noisy anechoic								
TF-GridNet	-	**13.2**	**32.9**	**13.3**	3.0	3.6	0.3	6.9
	ASR				2.3	0.3	1.0	3.6
	SSLR, ASR				2.1	0.3	0.8	3.1
	SSE, ASR	8.9	28.9	9.0	**1.5**	**0.2**	0.6	2.3
	SSE, SSLR, ASR	8.7	29.1	8.8	**1.5**	**0.2**	**0.5**	**2.2**
Noisy reverberant								
Zhang et al. (2022)	SSE, ASR	-2.27	-	-	-	-	-	28.9
Zhang et al. (2021b)	-	-	-	-	-	-	-	20.9
TF-GridNet	-	**11.1**	**30.2**	**11.3**	3.4	0.3	4.6	8.3
	ASR				2.5	0.3	1.1	3.9
	SSLR, ASR				2.2	0.2	0.8	3.2
	SSE, ASR	7.9	25.0	8.0	**1.6**	**0.2**	**0.5**	**2.3**
	SSE, SSLR, ASR	7.7	25.8	7.8	**1.6**	**0.2**	**0.5**	**2.3**

5.5. Multi-task learning framework with SSE and ASR losses

In this subsection, we investigate the impact of the multi-task learning with the SSE loss in addition to the ASR loss during the fine-tuning. Similar to Section 5.4, we used the two-channel anechoic and reverberant mixtures from the WHAMR! dataset. Since the dataset provides the clean source images for each mixture, we can seamlessly investigate the effect of the SSE loss in (19) by changing κ in (33). The previous multi-channel experiments does not use the SSE loss, corresponding to $\kappa = 0$. The joint fine-tuning of the SSE and ASR models was conducted on the `max` version to compute the ASR loss, while the SSE model was pre-trained on the `min` version. As a result, the separation performance on the `max` version could be improved through fine-tuning with the SSE loss.

Fig. 4 shows the relationship between SDR and WER for different weights on the SSE loss κ. Even with small weights, multi-task learning substantially mitigated the degradation of the separation performance. In detail, WERs were inversely proportional to SDRs under both anechoic and reverberant conditions across multi-task learning systems. The SSE loss in (33) enforced the model to achieve better signal-level separation and thus improved SDR. Meanwhile, E2E fine-tuning without the SSE loss is optimal for ASR applications, and WERs deteriorated as the weight κ increases. This could be because, for $\kappa > 0$, the SSE model separates sources more effectively but also introduces undesirable artifacts (Koizumi et al., 2022; Iwamoto et al., 2022), negatively impacting ASR performance. Meanwhile, the ASR loss may allow the SSE model to ignore the signal parts that are unimportant for the subsequent ASR, which degrades the signal-level metrics. These results reveal a trade-off between the SDR and WER, where improvements in one metric come at the expense of the other. This finding suggests the importance of task-specific fine-tuning because improving popular signal-level metrics might compete with down-stream objectives.

We emphasize that models with multi-task fine-tuning still outperformed existing systems (Zhang et al., 2022, 2021b) in Table 8 and our system without any fine-tuning.

6. Conclusions

We propose SIMO/MIMO-IRIS, a monaural/multi-channel multi-speaker ASR system, which integrates monaural/multi-channel SSE, SSLR extraction, and ASR in an E2E manner. By integrating TF-GridNET, WavLM, and a Conformer-based joint CTC/AED model, we obtained significant WER improvements and SOTA results on the spatialized WSJ0-2mix and WHAMR! datasets. Our experimental results demonstrated the importance of the E2E fine-tuning of the pre-trained SSE and ASR models, even with powerful pre-trained models including WavLM. E2E fine-tuning with only the ASR loss significantly decreased SDR by introducing artifacts, which can be mitigated via multi-task learning with an SSE loss.

In this paper, we focused on ideal settings where the numbers of microphones and speakers are time-invariant and known. In addition, our experiments were based on simulated reverberation, while using real noise recordings. In the future, we would like to extend SIMO/MIMO-IRIS to handle a time-varying number of speakers and long-form recordings e.g., via continuous speech separation (Chen et al., 2020; von Neumann et al., 2024). Such an extension would make the proposed system applicable to real multi-speaker conversations featured in the recent CHiME challenges (Watanabe et al., 2020; Cornell et al., 2023b).

CRediT authorship contribution statement

Yoshiki Masuyama: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Xuankai Chang:** Writing – review & editing, Software, Methodology, Investigation, Conceptualization. **Wangyou Zhang:** Writing – review & editing, Software, Methodology, Investigation, Conceptualization.

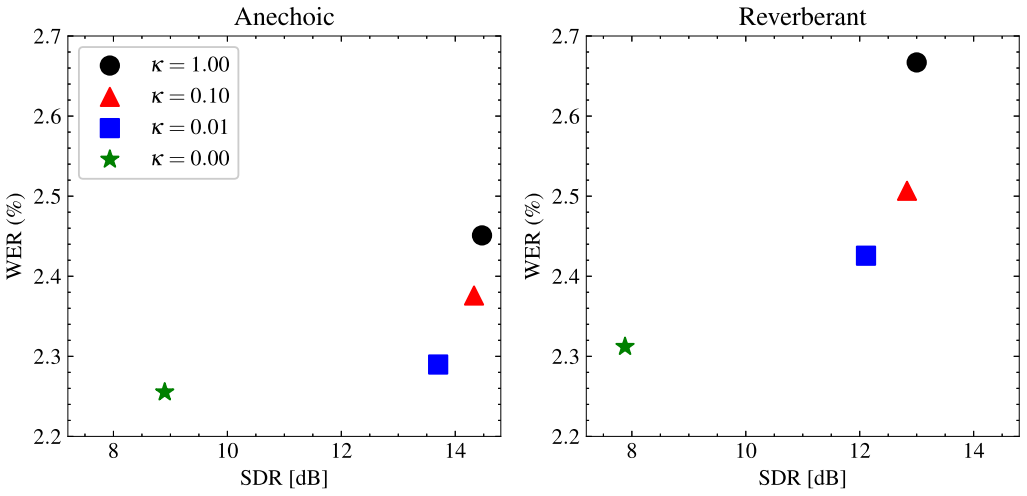

Fig. 4. SDR and WER with different weights κ on the SSE loss in multi-task learning framework. The star, $\kappa = 0$, indicates the fine-tuning without the SSE loss.

Samuele Cornell: Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Formal analysis. **Zhong-Qiu Wang:** Writing – review & editing, Supervision, Software, Methodology. **Nobutaka Ono:** Writing – review & editing, Supervision, Funding acquisition. **Yanmin Qian:** Writing – review & editing, Supervision, Funding acquisition. **Shinji Watanabe:** Writing – review & editing, Supervision, Project administration, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yoshiki Masuyama reports financial support was provided by Japan Society for the Promotion of Science. Xuankai Chang, Wangyou Zhang, Zhong-Qiu Wang reports financial support was provided by National Center for Supercomputing Applications. Samuele Cornell reports financial support was provided by Marche Region. Nobutaka Ono reports financial support was provided by Japan Science and Technology Agency. Shinji Watanabe (co-author) is a guest editor of the special issue we submit this manuscript. Samuele Cornell (co-author) is another guest editor of the special issue we submit this manuscript. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Y. Masuyama was partially supported by JSPS, Japan KAKENHI Grant Numbers JP21J21371. X. Chang, W. Zhang, and Z.-Q. Wang used the Bridges2 system at PSC and Delta system at NCSA through allocation CIS210014 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program. S. Cornell was partially supported by Marche Region, Italy within the funded project “Miracle” POR MARCHE FESR 2014–2020. N. Ono was partially supported by JST CREST, Japan Grant Number JPMJCR19A3.

Data availability

The datasets we used have been widely used in the literature. Clean speech is from a corpus that is available under the LDC, and noise signals are publicly available.

References

Baevski, A., Zhou, Y., Mohamed, A., Auli, M., 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Proc. NeurIPS, vol. 33, pp. 12449–12460.

Barker, J., Watanabe, S., Vincent, E., Trmal, J., 2018. The fifth ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines. In: Proc. Interspeech. pp. 1561–1565.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., et al., 2005. The AMI meeting corpus: A pre-announcement. In: Proc. ICMI-MLMI. pp. 28–39.

Chan, W., Jaitly, N., Le, Q., Vinyals, O., 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: Proc. ICASSP. pp. 4960–4964.

Chang, X., Maekaku, T., Fujita, Y., Watanabe, S., 2022. End-to-end integration of speech recognition, speech enhancement, and self-supervised learning representation. In: Proc. Interspeech. pp. 3819–3823.

Chang, X., Maekaku, T., Guo, P., Shi, J., Lu, Y.-J., Subramanian, A.S., Wang, T., Yang, S.-w., Tsao, Y., Lee, H.-y., Watanabe, S., 2021. An exploration of self-supervised pretrained representations for end-to-end speech recognition. In: Proc. ASRU. pp. 228–235.
Chang, X., Zhang, W., Qian, Y., Le Roux, J., Watanabe, S., 2019. MIMO-speech: End-to-end multi-channel multi-speaker speech recognition. In: Proc. ASRU. pp. 237–244.
Chang, X., Zhang, W., Qian, Y., Le Roux, J., Watanabe, S., 2020. End-to-end multi-speaker speech recognition with transformer. In: Proc. ICASSP. pp. 6134–6138.
Chen, G., Chai, S., Wang, G.-B., Du, J., Zhang, W.-Q., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J., Jin, M., Khudanpur, S., Watanabe, S., Zhao, S., Zou, W., Li, X., Yao, X., Wang, Y., You, Z., Yan, Z., 2021. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 h of transcribed audio. In: Proc. Interspeech. pp. 3670–3674.
Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., Wei, F., 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. IEEE J. Sel. Top. Signal Process. 16 (6), 1505–1518.
Chen, Z., Yoshioka, T., Lu, L., Zhou, T., Meng, Z., Luo, Y., Wu, J., Xiao, X., Li, J., 2020. Continuous speech separation: Dataset and analysis. In: Proc. ICASSP. pp. 7284–7288.
Chiu, C.-C., Sainath, T.N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R.J., Rao, K., Gonina, E., Jaitly, N., Li, B., Chorowski, J., Bacchiani, M., 2018. State-of-the-art speech recognition with sequence-to-sequence models. In: Proc. ICASSP. pp. 4774–4778.
Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y., 2015. Attention-based models for speech recognition. In: Proc. NeurIPS.
Cord-Landwehr, T., Boeddeker, C., Von Neumann, T., Zorilă, C., Doddipatla, R., Haeb-Umbach, R., 2022. Monaural source separation: From anechoic to reverberant environments. In: Proc. IWAENC.
Cornell, S., Jung, J.-w., Watanabe, S., Squartini, S., 2024. One model to rule them all? Towards end-to-end joint speaker diarization and speech recognition. In: Proc. ICASSP. pp. 11856–11860.
Cornell, S., Wang, Z.-Q., Masuyama, Y., Watanabe, S., Pariente, M., Ono, N., 2023a. Multi-channel target speaker extraction with refinement: The WAVLab submission to the second clarity enhancement challenge. In: Proc. Clarity.
Cornell, S., Wiesner, M., Watanabe, S., Raj, D., Chang, X., Garcia, P., Maciejewski, M., Masuyama, Y., Wang, Z.-Q., Squartini, S., Khudanpur, S., 2023b. The CHiME-7 DASR challenge: Distant meeting transcription with multiple devices in diverse scenarios. In: Proc. CHiME.
Cui, C., Sheikh, I., Sadeghi, M., Vincent, E., 2023. End-to-end multichannel speaker-attributed ASR: Speaker guided decoder and input feature analysis. In: Proc. ASRU.
El Shafey, L., Soltau, H., Shafran, I., 2019. Joint speech recognition and speaker diarization via sequence transduction. In: Proc. Interspeech. pp. 396–400.
Erdogan, H., Hershey, J.R., Watanabe, S., Mandel, M.I., Le Roux, J., 2016. Improved MVDR beamforming using single-channel mask prediction networks. In: Proc. Interspeech. pp. 1981–1985.
Fazel-Zarandi, M., Hsu, W.N., 2023. Cocktail HuBERT: Generalized self-supervised pre-training for mixture and single-source speech. In: Proc. ICASSP. pp. 1–5.
Fiscus, J.G., Ajot, J., Garofolo, J.S., 2007. The rich transcription 2007 meeting recognition evaluation. In: Proc. Multimodal Tech. Percept. Hum.. pp. 373–389.
Gannot, S., Burshtein, D., Weinstein, E., 2001. Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans. Signal Process. 49 (8), 1614–1626.
Gannot, S., Vincent, E., Markovich-Golan, S., Ozerov, A., 2017. A consolidated perspective on multi-microphone speech enhancement and source separation. IEEE/ACM Trans. Audio, Speech, Lang. Process. 25, 692–730.
Graves, A., 2012. Sequence transduction with recurrent neural networks. In: Proc. ICML Workshop Represent. Learn..
Graves, A., Fernández, S., Gomez, F., Schmidhuber, J., 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proc. ICML. pp. 369–376.
Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., Pang, R., 2020. Conformer: Convolution-augmented transformer for speech recognition. In: Proc. Interspeech. pp. 5036–5040.
Hershey, J.R., Chen, Z., Le Roux, J., Watanabe, S., 2016. Deep clustering: Discriminative embeddings for segmentation and separation. In: Proc. ICASSP. pp. 31–35.
Heymann, J., Drude, L., Boeddeker, C., Hanebrink, P., Haeb-Umbach, R., 2017. Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system. In: Proc. ICASSP. pp. 5325–5329.
Heymann, J., Drude, L., Haeb-Umbach, R., 2016. Neural network based spectral mask estimation for acoustic beamforming. In: Proc. ICASSP. pp. 196–200.
Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B., 2012. Deep neural networks for acoustgic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process. Mag. 26 (6), 82–97.
Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A., 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio, Speech, Lang. Process. 29, 3451–3460.
Huang, Z., Chen, Z., Kanda, N., Wu, J., Wang, Y., Li, J., Yoshioka, T., Wang, X., Wang, P., 2023a. Self-supervised learning with bi-label masked speech prediction for streaming multi-talker speech recognition. In: Proc. ICASSP. pp. 1–5.
Huang, K.P., Fu, Y.-K., Zhang, Y., Lee, H.-y., 2022. Improving distortion robustness of self-supervised speech processing tasks with domain adaptation. In: Proc. Interspeech. pp. 2193–2197.
Huang, Z., Raj, D., García, P., Khudanpur, S., 2023b. Adapting self-supervised models to multi-talker speech recognition using speaker embeddings. In: Proc. ICASSP. pp. 1–5.
Iwamoto, K., Ochiai, T., Delcroix, M., Ikeshita, R., Sato, H., Araki, S., Katagiri, S., 2022. How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR. In: Proc. Interspeech. pp. 5418–5422.
Kahn, J., Rivière, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., Likhomanenko, T., Synnaeve, G., Joulin, A., Mohamed, A., Dupoux, E., 2020. Libri-light: A benchmark for ASR with limited or no supervision. In: Proc. ICASSP. pp. 7669–7673.
Kanda, N., Gaur, Y., Wang, X., Meng, Z., Yoshioka, T., 2020. Serialized output training for end-to-end overlapped speech recognition. In: Proc. Interspeech. pp. 2797–2801.
Kanda, N., Wu, J., Wang, X., Chen, Z., Li, J., Yoshioka, T., 2023. VarArray meets T-SOT: Advancing the state of the art of streaming distant conversational speech recognition. In: Proc. ICASSP. pp. 1–5.
Kanda, N., Wu, J., Wu, Y., Xiao, X., Meng, Z., Wang, X., Gaur, Y., Chen, Z., Li, J., Yoshioka, T., 2022a. Streaming multi-talker ASR with token-level serialized output training. In: Proc. Interspeech. pp. 3774–3778.
Kanda, N., Xiao, X., Gaur, Y., Wang, X., Meng, Z., Chen, Z., Yoshioka, T., 2022b. Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed ASR. In: Proc. ICASSP. pp. 8082–8086.
Kanda, N., Ye, G., Wu, Y., Gaur, Y., Wang, X., Meng, Z., Chen, Z., Yoshioka, T., 2021. Large-scale pre-training of end-to-end multi-talker ASR for meeting transcription with single distant microphone. In: Interspeech. pp. 3430–3434.
Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N.E.Y., Yamamoto, R., Wang, X., Watanabe, S., Yoshimura, T., Zhang, W., 2019. A comparative study on transformer vs rnn in speech applications. In: Proc. ASRU. pp. 449–456.
Kinoshita, K., Delcroix, M., Gannot, S., P. Habets, E.A., Haeb-Umbach, R., Kellermann, W., Leutnant, V., Maas, R., Nakatani, T., Raj, B., Sehr, A., Yoshioka, T., 2016. A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. EURASIP J. Adv. Signal Process. 7.
Koizumi, Y., Karita, S., Narayanan, A., Panchapagesan, S., Bacchiani, M., 2022. SNRi target training for joint speech enhancement and recognition. In: Proc. Interspeech. pp. 1173–1177.

Kolbæk, M., Yu, D., Tan, Z.H., Jensen, J., 2017. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. 25 (10), 1901–1913.
L. Lu, J.L., Gong, Y., 2021. Streaming end-to-end multi-talker speech recognition. IEEE Signal Process. Lett. 28, 803–807.
Li, C., Qian, Y., Chen, Z., Kanda, N., Wang, D., Yoshioka, T., Qian, Y., Zeng, M., 2023. Adapting multi-lingual ASR models for handling multiple talkers. In: Proc. Interspeech. pp. 1314–1318.
Li, B., Sainath, T.N., Narayanan, A., Caroselli, J., Bacchiani, M., Misra, A., Shafran, I., Sak, H., Pundak, G., Chin, K., Sim, K.C., Weiss, R.J., Wilson, K.W., Variani, E., Kim, C., Siohan, O., Weintraub, M., McDermott, E., Rose, R., Shannon, M., 2017. Acoustic modeling for google home. In: Proc. Interspeech. pp. 399–403.
Li, B., Sainath, T.N., Weiss, R.J., Wilson, K.W., Bacchiani, M., 2016. Neural network adaptive beamforming for robust multichannel speech recognition. In: Proc. Interspeech. pp. 1976–1980.
Li, C., Shi, J., Zhang, W., Subramanian, A.S., Chang, X., Kamo, N., Hira, M., Hayashi, T., Boeddeker, C., Chen, Z., Watanabe, S., 2021. ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration. In: Proc. SLT. pp. 785–792.
Liang, Y., Shi, M., Yu, F., Li, Y., Zhang, S., Du, Z., Chen, Q., Xie, L., Qian, Y., Wu, J., Chen, Z., Lee, K.A., Yan, Z., Bu, H., 2023. The second multi-channel multi-party meeting transcription challenge (M2MeT 2.0): A benchmark for speaker-attributed ASR. In: Proc. ASRU.
Lu, Y.-J., Chang, X., Li, C., Zhang, W., Cornell, S., Ni, Z., Masuyama, Y., Yan, B., Scheibler, R., Wang, Z.-Q., et al., 2022. ESPnet-SE++: Speech enhancement for robust speech recognition, translation, and understanding. In: Proc. Interspeech. pp. 5458–5462.
Luo, Y., Chen, Z., Yoshioka, T., 2020. Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation. In: Proc. ICASSP. pp. 46–50.
Luo, Y., Mesgarani, N., 2019. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (8), 1256–1266.
Ma, C., Li, D., Jia, X., 2020. Optimal scale-invariant signal-to-noise ratio and curriculum learning for monaural multi-speaker speech separation in noisy environment. In: Proc. APSIPA ASC. pp. 711–715.
Maciejewski, M., Shi, J., Watanabe, S., Khudanpur, S., 2023. A dilemma of ground truth in noisy speech separation and an approach to lessen the impact of imperfect training data. Comput. Speech, Lang. 77, 101410.
Maciejewski, M., Wichern, G., McQuinn, E., Le Roux, J., 2020. WHAMR!: Noisy and reverberant single-channel speech separation. In: Proc. ICASSP. pp. 696–700.
Masuyama, Y., Chang, X., Cornell, S., Watanabe, S., Ono, N., 2023. End-to-end integration of speech recognition, dereverberation, beamforming, and self-supervised learning representation. In: Proc. SLT. pp. 260–265.
Masuyama, Y., Togami, M., Komatsu, T., 2020. Consistency-aware multi-channel speech enhancement using deep neural networks. In: Proc. ICASSP. pp. 821–825.
Meng, L., Kang, J., Cui, M., Wang, Y., Wu, X., Meng, H., 2023. A sidecar separator can convert a single-talker speech recognition system to a multi-talker one. In: Proc. ICASSP. pp. 1–5.
Miao, Y., Gowayyed, M., Metze, F., 2015. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. In: Proc. ASRU. pp. 167–174.
Minhua, W., Kumatani, K., Sundaram, S., Ström, N., Hoffmeister, B., 2019. Frequency domain multi-channel acoustic modeling for distant speech recognition. In: Proc. ICASSP. pp. 6640–6644.
Nakatani, T., Kinoshita, K., 2019. A unified convolutional beamformer for simultaneous denoising and dereverberation. IEEE Signal Process. Lett. 26 (6), 903–907.
Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., Juang, B., 2010. Speech dereverberation based on variance-normalized delayed linear prediction. IEEE Trans. Audio, Speech, Lang. Process. 18 (7), 1717–1731.
von Neumann, T., Boeddeker, C., Cord-Landwehr, T., Delcroix, M., Haeb-Umbach, R., 2024. Meeting recognition with continuous speech separation and transcription-supported diarization. In: Proc. ICASSP. pp. 775–779.
von Neumann, T., Boeddeker, C., Drude, L., Kinoshita, K., Delcroix, M., Nakatani, T., Haeb-Umbach, R., 2020a. Multi-talker ASR for an unknown number of sources: Joint training of source counting, separation and ASR. In: Proc. Interspeech. pp. 3097–3101.
von Neumann, T., Kinoshita, K., Drude, L., Boeddeker, C., Delcroix, M., Nakatani, T., Haeb-Umbach, R., 2020b. End-to-end training of time domain audio separation and recognition. In: Proc. ICASSP. pp. 7004–7008.
Ochiai, T., Watanabe, S., Hori, T., Hershey, J.R., 2017. Multichannel end-to-end speech recognition. In: Proc. ICML. pp. 2632–2641.
Pan, Z., Wichern, G., Masuyama, Y., Germain, F.G., Khurana, S., Hori, C., Le Roux, J., 2023. Scenario-aware audio-visual TF-GridNet for target speech extraction. In: Proc. ASRU.
Qian, Y., Chang, X., Yu, D., 2018. Single-channel multi-talker speech recognition with permutation invariant training. Speech Commun. 104, 1–11.
Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2023. Robust speech recognition via large-scale weak supervision. In: Proc. ICML. pp. 28492–28518.
Raj, D., Denisov, P., Chen, Z., Erdogan, H., Huang, Z., He, M., Watanabe, S., Du, J., Yoshioka, T., Luo, Y., Kanda, N., Li, J., Wisdom, S., Hershey, J.R., 2021. Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis. In: Proc. SLT. pp. 897–904.
Raj, D., Lu, L., Chen, Z., Gaur, Y., Li, J., 2022. Continuous streaming multi-talker ASR with dual-path transducers. In: Proc. ICASSP. pp. 7317–7321.
Ravenscroft, W., Close, G., Goetze, S., Hain, T., Soleymanpour, M., Chowdhury, A., Fuhs, M.C., 2024. Transcription-free fine-tuning of speech separation models for noisy and reverberant multi-speaker automatic speech recognition. In: Proc. Interspeech. pp. 4998–5002.
Seki, H., Hori, T., Watanabe, S., Le Roux, J., Hershey, J.R., 2018. A purely end-to-end system for multi-speaker speech recognition. In: Proc. ACL. pp. 2620–2630.
Seltzer, M.L., Raj, B., Stern, R.M., 2004. Likelihood-maximizing beamforming for robust hands-free speech recognition. IEEE Trans. Speech, Audio Process. 12 (5), 489–498.
Settle, S., Le Roux, J., Hori, T., Watanabe, S., Hershey, J.R., 2018. End-to-end multi-speaker speech recognition. In: Proc. ICASSP. pp. 4819–4823.
Shi, J., Chang, X., Guo, P., Watanabe, S., Fujita, Y., Xu, J., Xu, B., Xie, L., 2020. Sequence to multi-sequence learning via conditional chain mapping for mixture signals. In: Proc. NeurIPS. 33, pp. 3735–3747.
Sklyar, I., Piunova, A., Liu, Y., 2021. Streaming multi-speaker ASR with RNN-T. In: Proc. ICASSP. pp. 6903–6907.
Souden, M., Benesty, J., Affes, S., 2010. On optimal frequency-domain multichannel linear filtering for noise reduction. IEEE Trans. Audio, Speech, Lang. Process. 18 (2), 260–276.
Steinmetz, C.J., Reiss, J.D., 2021. pyloudnorm: A simple yet flexible loudness meter in python. In: Proc. AES.
Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., Zhong, J., 2021. Attention is all you need in speech separation. In: Proc. ICASSP. pp. 21–25.
Tan, K., Wang, D., 2020. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 28, 380–390.
Tan, K., Wang, Z.-Q., Wang, D., 2022. Neural spectrospatial filtering. IEEE/ACM Trans. Audio, Speech, Lang. Process. 30, 605–621.
Tsai, H.-S., Chang, H.-J., Huang, W.-C., Huang, Z., Lakhotia, K., Yang, S.-w., Dong, S., Liu, A., Lai, C.-I., Shi, J., Chang, X., Hall, P., Chen, H.-J., Li, S.-W., Watanabe, S., Mohamed, A., Lee, H.-y., 2022. SUPERB-SG: Enhanced speech processing universal performance benchmark for semantic and generative capabilities. In: Proc. ACL. pp. 8479–8492.
Wang, D., Chen, J., 2018. Supervised speech separation based on deep learning: An overview. IEEE/ACM Trans. Audio, Speech, Lang. Process. 26 (10), 1702–1726.
Wang, Z.Q., Cornell, S., Choi, S., Lee, Y., Kim, B.Y., Watanabe, S., 2023a. TF-GridNet: Integrating full- and sub-band modeling for speech separation. IEEE/ACM Trans. Audio, Speech, Lang. Process. 31, 3221–3236.

Wang, Z.Q., Cornell, S., Choi, S., Lee, Y., Kim, B.Y., Watanabe, S., 2023b. TF-GridNet: Making time-frequency domain models great again for monaural speaker separation. In: Proc. ICASSP.
Wang, Z.-Q., Kumar, A., Watanabe, S., 2024. Cross-talk reduction. In: Proc. IJCAI. pp. 5171–5180.
Wang, Z.-Q., Le Roux, J., Hershey, J.R., 2018. Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation. In: Proc. ICASSP. pp. 1–5.
Wang, Y., Li, J., Wang, H., Qian, Y., Wang, C., Wu, Y., 2022a. Wav2vec-switch: Contrastive learning from original-noisy speech pairs for robust speech recognition. In: Proc. ICASSP. pp. 7097–7101.
Wang, H., Qian, Y., Wang, X., Wang, Y., Wang, C., Liu, S., Yoshioka, T., Li, J., Wang, D., 2022b. Improving noise robustness of contrastive speech representation learning with speech reconstruction. In: Proc. ICASSP. pp. 6062–6066.
Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., Dupoux, E., 2021a. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In: Proc. ACL. pp. 993–1003.
Wang, Z.-Q., Wang, P., Wang, D., 2020. Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR. IEEE/ACM Trans. Audio, Speech, Lang. Process. 28, 1778–1787.
Wang, Z.-Q., Wang, P., Wang, D., 2021b. Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation. IEEE/ACM Trans. Audio, Speech, Lang. Process. 29, 2001–2014.
Wang, Z.Q., Wichern, G., Le Roux, J., 2021c. On the compensation between magnitude and phase in speech separation. IEEE Signal Process. Lett. 28, 2018–2022.
Warsitz, E., Haeb-Umbach, R., 2007. Blind acoustic beamforming based on generalized eigenvalue decomposition. IEEE Trans. Audio, Speech, Lang. Process. 15 (5), 1529–1539.
Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Enrique Yalta Soplin, N., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., Ochiai, T., 2018. Espnet: End-to-end speech processing toolkit. In: Proc. Interspeech. pp. 2207–2211.
Watanabe, S., Hori, T., Kim, S., Hershey, J.R., Hayashi, T., 2017. Hybrid CTC/Attention architecture for end-to-end speech recognition. IEEE J. Sel. Top. Signal Process. 11 (8), 1240–1253.
Watanabe, S., Mandel, M., Barker, J., Vincent, E., Arora, A., Chang, X., Khudanpur, S., Manohar, V., Povey, D., Raj, D., et al., 2020. CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. In: Proc. CHiME.
Williamson, D.S., Wang, Y., Wang, D., 2016. Complex ratio masking for monaural speech separation. IEEE/ACM Trans. Audio, Speech, Lang. Process. 24 (3), 483–492.
Yang, S.-w., Chi, P.-H., Chuang, Y.-S., Lai, C.-I.J., Lakhotia, K., Lin, Y.Y., Liu, A.T., Shi, J., Chang, X., Lin, G.-T., Huang, T.-H., Tseng, W.-C., Lee, K.-t., Liu, D.-R., Huang, Z., Dong, S., Li, S.-W., Watanabe, S., Mohamed, A., Lee, H.-y., 2021. SUPERB: Speech processing universal performance benchmark. In: Proc. Interspeech. pp. 1194–1198.
Yang, L., Liu, W., Wang, W., 2022. TFPSNet: Time-frequency domain path scanning network for speech separation. In: Proc. ICASSP. pp. 6842–6846.
Yifan, G., Yao, T., Hongbin, S., Yulong, W., 2023. Multi-channel multi-speaker transformer for speech recognition. In: Proc. Interspeech. pp. 4918–4922.
Yoshioka, T., Erdogan, H., Chen, Z., Alleva, F., 2018a. Multi-microphone neural speech separation for far-field multi-talker speech recognition. In: Proc. ICASSP. pp. 5739–5743.
Yoshioka, T., Erdogan, H., Chen, Z., Xiao, X., Alleva, F., 2018b. Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks. In: Proc. Interspeech. pp. 3038–3042.
Yu, D., Chang, X., Qian, Y., 2017a. Recognizing multi-talker speech with permutation invariant training. In: Proc. Interspeech. pp. 2456–2460.
Yu, D., Kolbæk, M., Tan, Z.-H., Jensen, J., 2017b. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In: Proc. ICASSP. pp. 241–245.
Zhang, W., Chang, X., Boeddeker, C., Nakatani, T., Watanabe, S., Qian, Y., 2022. End-to-end dereverberation, beamforming, and speech recognition in a cocktail party. IEEE/ACM Trans. Audio, Speech, Lang. Process. 30, 3173–3188.
Zhang, W., Shi, J., Li, C., Watanabe, S., Qian, Y., 2021a. Closing the gap between time-domain multi-channel speech enhancement on real and simulation conditions. In: Proc. WASPAA. pp. 146–150.
Zhang, W., Subramanian, A.S., Chang, X., Watanabe, S., Qian, Y., 2020a. End-to-end far-field speech recognition with unified dereverberation and beamforming. In: Proc. Interspeech. pp. 324–328.
Zhang, J., Zorila, C., Doddipatla, R., Barker, J., 2021b. Time-domain speech extraction with spatial information and multi speaker conditioning mechanism. In: Proc. ICASSP. pp. 6084–6088.
Zhang, J., Zorilă, C., Doddipatla, R., Barker, J., 2020b. On end-to-end multi-channel time domain speech separation in reverberant environments. In: Proc. ICASSP. pp. 6389–6393.
Zhou, P., Feng, J., Ma, C., Xiong, C., Hoi, S.C.H., E, W., 2020. Towards theoretically understanding why sgd generalizes better than adam in deep learning. In: Proc. NeurIPS, vol. 33, pp. 21285–21296.