



End-to-End Integration of Speech Recognition, Speech Enhancement, and Self-Supervised Learning Representation

Xuankai Chang¹, Takashi Maekaku², Yuya Fujita², Shinji Watanabe¹

¹Carnegie Mellon University, PA, USA

²Yahoo Japan Corporation, Tokyo, JAPAN

{xuankaic, swatanab}@andrew.cmu.edu, {tmaekaku, yuyfujit}@yahoo-corp.jp

Abstract

This work presents our end-to-end (E2E) automatic speech recognition (ASR) model targeting at robust speech recognition, called Integrated speech Recognition with enhanced speech Input for Self-supervised learning representation (IRIS). Compared with conventional E2E ASR models, the proposed E2E model integrates two important modules including a speech enhancement (SE) module and a self-supervised learning representation (SSLR) module. The SE module enhances the noisy speech. Then the SSLR module extracts features from enhanced speech to be used for speech recognition (ASR). To train the proposed model, we establish an efficient learning scheme. Evaluation results on the monaural CHiME-4 task show that the IRIS model achieves the best performance reported in the literature for the single-channel CHiME-4 benchmark (2.0% for the real development and 3.6% for the real test) thanks to the powerful pre-trained SSLR module and the fine-tuned SE module.

Index Terms: robust automatic speech recognition, self-supervised learning, speech enhancement, deep learning

1. Introduction

In the past decade, deep learning has significantly pushed the development of automatic speech recognition (ASR) moving forward. Many interesting models and technologies have been proposed. Deep neural network-hidden Markov model (DNN-HMM) based hybrid system [1] is one of the them. DNN-HMM hybrid ASR systems usually train a DNN to predict frame-aligned states, e.g. context-dependent phonemes. Recently, end-to-end speech recognition systems have become more and more popular. Several end-to-end ASR technologies were proposed, including connectionist temporal classification (CTC) [2], Transducer [3] and attention-based encoder-decoder [4, 5, 6]. A lot of existing speech recognition techniques exhibit strong performance in clean conditions. However, applying speech recognition in noisy environments is still challenging, especially in the monaural case. DNN-HMM hybrid ASR systems still outperform E2E ASR system on a well-known noisy speech corpus[7], CHiME-4 corpus [8].

Usually, speech signals recorded in the real scenarios contain unpredicted noise. The noise is from the environment or the device imperfections, which degrades the ASR performance. The existing solutions to address the noisy speech recognition can be summarized as two categories. One is to train the ASR model robust to noise [9, 10, 11]. The other is to use an dedicated model to improve the intelligibility of the noisy speech before sending it to the ASR model. Such preprocessing is one of the important topics in speech research, called speech enhancement (SE) or denoising [12]. The SE model and the ASR

model can be trained separately or jointly [13, 14, 15]. However, it is well known that the monaural SE techniques produce distortions which deteriorates the ASR performance [16, 17].

Recently, self-supervised learning representations (SSLR) have demonstrated great potential in improving the speech recognition [18, 19, 20, 21]. One primary drawback of current SSLR models is that the pre-training cost is too high for most of the research groups. As an alternative solution, some researchers fine-tune the pre-trained SSLR models to get their customized version [22]. In our previous study [21], we have shown that directly using the pre-trained Wav2Vec2.0 [18] and HuBERT [20] for feature extraction improves the ASR performance. However, the improvement on mismatched conditions is usually limited. The result of CHiME-4 corpus in [21] shows the word error rate (WER) reduction for the multi-channel data with beamforming are much better than those for the isolated single channel. Because the audio of the latter set is relatively noisier than the former one. We believe that it is due to the mismatch between the pre-training and the target task. Wav2Vec2.0 and HuBERT models were pre-trained on the LibriLight [23] data, a clean read English speech corpus. Later, WavLM [24] was proposed to learn a representation model on simulated noisy / overlapped speech. In another recent work, Wang *et al.* [25] proposed a noisy robust SSLR model based on Wav2Vec2.0, which also shows promising results on CHiME-4. But the model is not publicly available.

In this work, we propose a new model, called IRIS, for robust speech recognition, which integrates an SE module, an SSLR module and an ASR module into a single end-to-end model. We extensively investigate the benefits of the SE module and the SSLR module for robust speech recognition. Through experiments, we establish an efficient training scheme for the proposed E2E IRIS model. Finally, we show that our proposed model achieves state-of-the-art performance on the single-channel CHiME-4 ASR tasks. The scripts, configurations and the trained models and more detailed results of this work are publicly available in ESPnet¹.

2. IRIS Model

We describe the proposed IRIS model in this section. The model includes a speech enhancement module (SE) and a self-supervised learning representation (SSLR)-based ASR (SSLR-ASR) module, shown in Figure 1. Each module can be trained separately. Then whole model can be fine-tuned with the objectives of speech enhancement and recognition. For the convenience of the following discussions, we denote the noisy speech input as Y .

¹https://github.com/espnet/espnet/tree/master/egs2/chime4/enh_asr1

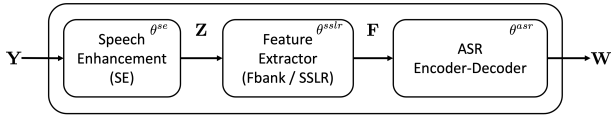


Figure 1: Overview of the proposed end-to-end model.

2.1. Speech Enhancement

Most of the data collected in real scenarios contains not only speech signal but also undesired noise and reverberation. The target of speech enhancement is to keep the speech signal from data and to suppress the undesired signals. We denote the SE process as the following:

$$\mathbf{Z} = \text{SE}(\mathbf{Y}; \theta^{\text{se}}), \quad (1)$$

where \mathbf{Z} is the enhanced speech and θ^{se} represents the parameters of the SE model.

A lot of powerful speech enhancement techniques have been proposed. In this work, we choose the Conv-TasNet proposed in [26] as the SE module. Conv-TasNet is a very successful model for end-to-end time-domain speech enhancement. Besides this, many other strong end-to-end time-domain speech enhancement models were proposed before [27, 28, 29]. The advantage of the time-domain speech enhancement model is that we do not need to care about the phase when we generate enhanced signals. This might be helpful to reduce the distortion generated by speech enhancement models. Without loss of generality, any speech enhancement models can be used in our model.

2.2. SSLR-ASR

2.2.1. E2E-ASR

To recognize the speech, we use an E2E-ASR model. If we denote the input speech signal as \mathbf{Z} , the feature of the speech as \mathbf{F} and the text as \mathbf{W} , we can write the ASR process as:

$$\mathbf{F} = \text{FeatureExtraction}(\mathbf{Z}), \quad (2)$$

$$\mathbf{W} = \text{ASR}(\mathbf{F}; \theta^{\text{asr}}), \quad (3)$$

where θ^{asr} represents the parameters of the ASR model. In this work, we use the joint CTC / attention-based encoder-decoder framework proposed in [5] to build our E2E-ASR model. More details can be referred to [5, 6]. It is worth to note that the choice of ASR is not limited to a specific architecture.

2.2.2. SSLR

Conventional ASR models use energy-based features such as log Mel-Filterbanks (Fbank) and mel-frequency cepstral coefficients (MFCC). In our previous work [21], we have shown that replacing the energy-based features with SSLRs can improve the performance of E2E-ASR. In this way, the Eq. 2 would be rewritten as:

$$\mathbf{F} = \text{SSLR}(\mathbf{Z}; \theta^{\text{sslr}}), \quad (4)$$

where θ^{sslr} represents the parameters of the SSLR model.

SSLR models are learning-based speech representations. SSLR models are trained using large amount of unlabelled data. In this study, we propose to use WavLM proposed in [24] to improve robust speech recognition. Similar to HuBERT [20], WavLM is trained to predict pseudo-labels of masked segments.

In this way, WavLM / HuBERT learns the linguistic information from speech. The HuBERT model we used is trained on 60k hours of Libri-Light [23] speech data. Whereas the WavLM learns to handle the noise from the speaker identification, separation, and diarization tasks by training on 60k hours of Libri-Light [23], 10k hours of GigaSpeech [30], and 24k hours of VoxPopuli [31]. This motivates us to use WavLM to extract features for noisy speech.

2.3. End-to-End IRIS Model

Although WavLM shows good performance on the noisy speech input in downstream tasks, such as speaker identification, separation and diarization tasks [24], it is still a question whether it can handle various noises. We propose to use a speech enhancement model to help the WavLM. Our end-to-end model can be written as:

$$\mathbf{W} = \text{ASR}(\text{SSLR}(\text{SE}(\mathbf{Y}; \theta^{\text{se}}); \theta^{\text{sslr}}); \theta^{\text{asr}}) \quad (5)$$

The proposed model adopts a modularized design, where the SE module enhances the input noisy speech, SSLR module extracts the feature and the ASR module generates the transcription. We directly use the pre-trained SSLR models from existing works, which are publicly available. Usually, SSLR models are very large, which makes it difficult to train the whole model. To address this issue, all three modules are initialized by pre-trained models with parameters $\hat{\theta}^{\text{se}}$, $\hat{\theta}^{\text{sslr}}$ and $\hat{\theta}^{\text{asr}}$, respectively. Then the parameters of SE and ASR are fine-tuned to get better performance.

3. Experimental Setup

3.1. CHiME-4 Corpus

We carried out all the experiments on the CHiME-4 corpus [8]. The dataset contains real and simulated six-channel noisy recordings of speech from Wall Street Journal (WSJ0) corpus. The recordings cover four noisy scenarios including bus, cafe, pedestrian and street. There are 1,600 real and 7,138 simulated utterances for training, 1,640 real and 1,640 simulated utterances for development, and 1,320 real and 1,320 simulated utterances for test.

All the channels of CHiME-4 simulated recordings are used to train the SE model. To train the ASR model, we exclude the second channel of the CHiME-4 training set. This brings slight improvement because the second channel faces backward. Besides the noisy utterances in CHiME-4, the clean Wall Street Journal (WSJ0 + WSJ1) utterances are also used to train the E2E ASR model, based on the original ESPnet CHiME-4 recipe². In fine-tuning, we use the same data as in ASR training. During evaluation, the single-channel development and test sets are used.

3.2. Configurations

We use a relatively small Conv-TasNet enhancement model to save the computation. The encoder consists of an 1-D convolution layer, with 256 output channel (N). The kernel and stride sizes are 40 and 20 respectively. The decoder has a reverse 1-D convolution layer with corresponding hyper-parameters of encoder. In the separation part, the temporal convolutional network (TCN), 4 convolutional blocks (X) are repeated twice (R).

²<https://github.com/espnet/espnet/tree/master/egs2/chime4/asr1>

The number of channels (H) and the kernel size (P) in convolutional blocks are 512 and 3, respectively. The bottleneck has 256 channels (B). More detailed meaning of the hyper-parameters can be referred to [26]. SI-SNR[27] is used to compute the enhancement loss between the reference signal and the enhanced signal. The enhancement model is optimized by adam algorithm with learning rate at 1×10^{-3} .

For the ASR model, we use Transformer block to build the encoder and the decoder. The ASR model contains 12 encoder and 6 decoder Transformer layers. For each Transformer layer, the number of attention heads is 4. The dimension of the linear projection is 2,048. The encoder uses two convolutional layers to downsample the input feature sequence and the total frame shift is 40ms. The dropout is set to be 0.1. In addition to the log Mel-Filterbank (Fbank), we use two SSLR models as feature extractor including the HuBERT-large and WavLM-large. When using the SSLR as feature extractor, the feature dimension is reduced from 1,024 to 128 with a linear layer before input to the encoder. The ASR model is optimized by adam algorithm with peak learning rate at 1×10^{-3} and 20k steps to warm up. SpecAug [32] is used for both Fbank and SSLR feature during training. In decoding, we use a Transformer language model based on character level, with weight 1.0 during beam search.

In the proposed IRIS model, each of the modules is initialized by pre-training. SE and ASR parameters are from the pre-trained models described above. The IRIS model is fine-tuned with 10 epochs using both the enhancement and ASR losses. The same optimizer algorithm for ASR model training is used, with learning rate at 5×10^{-4} . During the training of both ASR and IRIS models, the parameters of the SSLR models are not updated.

Unless otherwise mentioned, model averaging is performed over the 10 checkpoints with best accuracy during decoding.

4. Results

We present our experimental results and analysis in this section.

4.1. E2E-ASR Model with SSLRs

In this part, we show the evaluation results of ASR models on the monaural CHiME-4 corpus. The word error rates (WERs) of both simulated and real speech recordings are computed on the development and the test sets. The results are shown in Table 1. The results of systems 1-4 are from existing research works. Among them, system 4 is based on E2E-ASR. The rest systems are built by hybrid ASR systems. We can observe that the best performance is achieved by the hybrid ASR method. We have trained systems 5-7. In system 5, we use the conventional Fbank feature to train the E2E-ASR model, the performance of which is worse than system 3 by a large gap. In system 6 and 7, we use the HuBERT and WavLM models, which are pre-trained on large amount of unlabelled data, to extract feature. When using HuBERT to generate speech features, there is no consistent or obvious improvement across all the evaluation data. We conjecture that it is because the HuBERT is only pre-trained on the clean speech. This can be inferred from the performance of system 4 and 7. In system 4, the Wav2Vec2.0-based model was trained with noisy speech data, leading to similar performance as system 3. Likewise, system 7 using WavLM for feature extraction also achieves comparable performance with system 3. The WERs of simulated speech are 5.9% and 8.2% on dev and test sets, respectively, and those of real speech are 4.0% and 4.5%. Specially, in the test set, the WERs of real recordings

is 28% better than the previous best results. From this results, we find that it is important to use noisy data to train the robust speech SSLR.

Table 1: Single-channel CHiME-4 ASR performance (%WER) of the E2E-ASR model and previous studies on monaural dev and test sets. In system 6 and 7, HuBERT and WavLM are pre-trained models learned on different sets of external data.

ID	System	Model	Dev. Set		Test Set	
			Simu.	Real	Simu.	Real
1	Kaldi Baseline [33]	Hybrid	6.81	5.58	12.15	11.42
2	Du <i>et al.</i> [34]	Hybrid	6.61	4.55	11.81	9.15
3	Yang <i>et al.</i> [7]	Hybrid	4.99	3.35	8.61	6.25
4	Wav2Vec-Switch [25]	E2E	-	3.5	-	6.6
5	E2E Transformer - Fbank	E2E	11.32	9.43	19.67	17.99
6	E2E Transformer - HuBERT	E2E	11.56	9.13	18.02	20.41
7	E2E Transformer - WavLM	E2E	5.93	4.03	8.25	4.47

4.2. IRIS Model

Table 2: Monaural CHiME-4 ASR performance (%WER) of the IRIS model. Different combinations of fine-tuning SE (FT. SE) and fine-tuning ASR (FT. ASR) are evaluated.

Enhancement	Feature	FT. SE	FT. ASR	Dev. Set		Test Set	
				Simu.	Real	Simu.	Real
Conv-TasNet	Fbank	✗	✗	17.22	16.76	30.28	32.50
	Fbank	✗	✓	11.42	9.92	21.16	21.82
	Fbank	✓	✗	9.20	8.33	17.01	16.56
	Fbank	✓	✓	9.52	7.94	17.42	15.24
	WavLM	✗	✗	5.96	4.37	13.52	12.11
	WavLM	✗	✓	5.45	4.04	12.68	11.57
	WavLM	✓	✗	3.54	2.27	6.73	4.90
	WavLM	✓	✓	3.43	1.98	6.21	3.64

Table 3: ASR performance (%WER) comparison between the proposed IRIS model and the best existing single- and multi-channel systems.

System	Track	Dev. Set		Test Set	
		Simu.	Real	Simu.	Real
IRIS (proposed)	1ch	3.43	1.98	6.21	3.64
Yang <i>et al.</i> [7]	1ch	4.99	3.35	8.61	6.25
Du <i>et al.</i> [34]	2ch	3.46	2.33	5.74	3.91
Wang <i>et al.</i> [35]	2ch	2.17	1.99	2.53	3.19
Kaldi Baseline [33]	6ch	1.90	2.10	2.74	2.66
Wang <i>et al.</i> [35]	6ch	1.15	1.50	1.45	1.99

Next, we evaluate our proposed IRIS models. From the results in Table 1, we already know that WavLM is robust in the noisy condition. In this part, we further investigate if adding a speech enhancement module is beneficial to the model. As a reference, we did the similar evaluation on the E2E-ASR based on Fbank. Considering the computation cost when concatenated with the ASR model, we choose Conv-TasNet as the enhancement model and reduce the number of parameters by using a shallow architecture described in Sec. 3.2. The SI-SNRs of the pre-trained speech enhancement model are 9.55 dB and 9.71 dB on the development and test sets, respectively.

First, we directly concatenate the speech enhancement model and E2E-ASR models to perform the speech recognition. The results are shown in the Table 2. If the simple concatenation is used, both the performance of the Fbank-based system and that of the WavLM-based system are degraded, compared with the results of system 5 and 7 in the previous table. This indicates

that speech enhancement models do not necessarily improve the ASR performance on noisy speech, because the training objectives of speech enhancement and recognition are not very well aligned. It is a well-known phenomenon in previous research [17].

Second, if we keep the enhancement model fixed and fine-tune the ASR model with ASR loss, the performance of a WavLM-based system is slightly improved but not reaching the same level as system 7 in the previous table. We believe the artifacts from the enhancement model is difficult to handle by the WavLM. For the Fbank-based system, the performance degradation is mitigated. However, in the other way around, if we keep the ASR model fixed and fine-tune the enhancement model with both enhancement loss and the ASR loss, we find that the performance are significantly improved in WavLM-based model and Fbank-based model, especially on the simulation sets. We assume that the major reason is because only the simulation data is used to fine-tune the enhancement module.

As the last case, we fine-tune both enhancement and ASR models with the enhancement and ASR losses. We observe further improvements on both Fbank-based and WavLM-based models. For the WavLM system, the best performance is achieved. Compared to the system 7 in the previous table, WavLM without speech enhancement, WERs on all the evaluation sets are further improved with a nonneglectable improvement. In Table 3, we list the best result of existing systems from Table 1, the result of the 1st ranking system in CHiME-4 two- and six-channel track [34] and the result of our end-to-end IRIS system. Our system achieves a new state-of-the-art performance on the monaural CHiME-4 ASR task³, outperforming the best monaural system. More interestingly, the results are comparable to the CHiME-4 challenge best 2-channel results from [34].

The results indicate that the noise robust SSLR can still suffer from the degradation of noise. We can greatly alleviate the problem by introducing a speech enhancement as pre-processing. However, it is critical to fine-tune both models jointly to eliminate the mismatch. This rule can be applied to Fbank-based E2E-ASR model as well.

4.3. Analysis

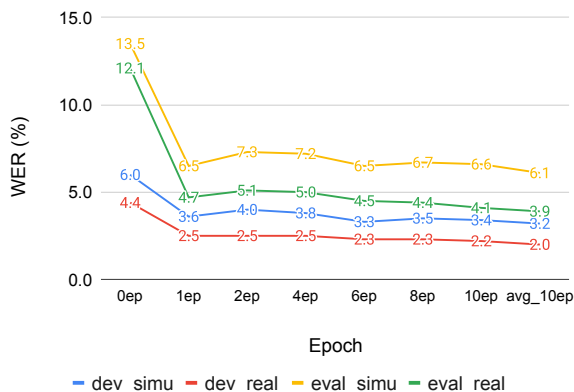


Figure 2: CHiME-4 ASR performance (WERs) of the IRIS model at different epochs during fine-tuning. Both SE and ASR are fine-tuned.

It is interesting to know how the fine-tuning improves the IRIS model. We show the ASR performance with the check-

³The pre-trained SSLR has more parameters and uses more data.

points in the middle of fine-tuning the IRIS model in Figure 2. It can be observed that the fine-tuning converges very fast. After only one epoch, the WERs can reach a very good level. With the model average over the first 10 epochs, the best performance can be observed.

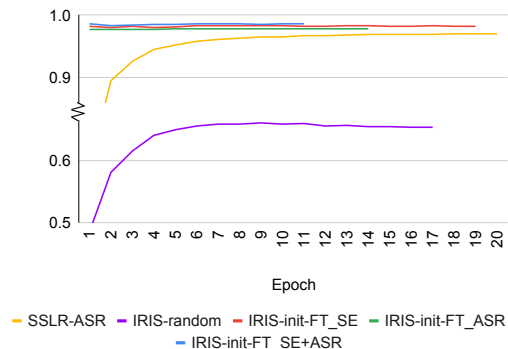


Figure 3: Accuracies on the development set for training and fine-tuning different models.

One difficult point is that the current IRIS model needs pre-training, taking extra efforts to prepare the individual enhancement and ASR models. In Figure 3, we show the training curves of the following models:

Model	Init. Param.	Update Param.
SSLR-ASR	$\hat{\theta}^{sslr}$	θ^{asr}
IRIS-random	$\hat{\theta}^{sslr}$	$\theta^{se}, \theta^{asr}$
IRIS-init-FT_SE	$\hat{\theta}^{se}, \hat{\theta}^{sslr}, \hat{\theta}^{asr}$	θ^{se}
IRIS-init-FT_AS+ASR	$\hat{\theta}^{se}, \hat{\theta}^{sslr}, \hat{\theta}^{asr}$	θ^{asr}
IRIS-init-FT_SE+ASR	$\hat{\theta}^{se}, \hat{\theta}^{sslr}, \hat{\theta}^{asr}$	$\theta^{se}, \theta^{asr}$

We can see that training the IRIS model from random initialization could not converge to a good point. We assume that the deep architecture of the SSLR models might disturb the gradient back-propagation from ASR to the enhancement. More training tricks are required. However, if we initialize the parameters of each module, the training reaches a good level after the 1st epoch.

5. Conclusions

We have proposed a new end-to-end model, IRIS, for robust speech recognition. The model contains three modules including an SE module, an SSLR module and an ASR module. For the implementation, we use Conv-TasNet as SE module, WavLM as SSLR module and a joint CTC/attention-based encoder-decoder as ASR module. In the evaluation on monaural CHiME-4 task, the IRIS model outperforms the current state-of-the-art system, which is based on the hybrid ASR model. It should be noted that the pre-training of SSLR model uses more data and more parameters. Moving forward, we plan to explore the generalization ability of our proposed system and to investigate the way to enable the end-to-end joint training from random initialization.

6. Acknowledgements

This work used the Extreme Science and Engineering Discovery Environment (XSEDE) [36], which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system [37], which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).

7. References

- [1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *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] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proc. ICML*, 2006, pp. 369–376.
- [3] A. Graves, A. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proc. ICASSP*, 2013, pp. 6645–6649.
- [4] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” in *Proc. ICASSP*, 2016, pp. 4960–4964.
- [5] S. Kim, T. Hori, and S. Watanabe, “Joint CTC-attention based end-to-end speech recognition using multi-task learning,” in *Proc. ICASSP*, 2017, pp. 4835–4839.
- [6] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, “ESPnet: End-to-end speech processing toolkit,” in *Proc. Interspeech*, 2018, pp. 2207–2211.
- [7] Y. Yang, P. Wang, and D. Wang, “A conformer based acoustic model for robust automatic speech recognition,” *arXiv preprint arXiv:2203.00725*, 2022.
- [8] 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.
- [9] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, “Deep speech: Scaling up end-to-end speech recognition,” *arXiv preprint arXiv:1412.5567*, 2014.
- [10] Y. Shinohara, “Adversarial multi-task learning of deep neural networks for robust speech recognition,” in *Proc. Interspeech*, 2016, pp. 2369–2372.
- [11] C. Kim, E. Variiani, A. Narayanan, and M. Bacchiani, “Efficient implementation of the room simulator for training deep neural network acoustic models,” *arXiv preprint arXiv:1712.03439*, 2017.
- [12] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.
- [13] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, “Multichannel end-to-end speech recognition,” in *Proc. ICML*, 2017, pp. 2632–2641.
- [14] A. Narayanan and D. Wang, “Joint noise adaptive training for robust automatic speech recognition,” in *Proc. ICASSP*, 2014, pp. 2504–2508.
- [15] A. S. Subramanian, X. Wang, M. K. Baskar, S. Watanabe, T. Taniguchi, D. Tran, and Y. Fujita, “Speech enhancement using end-to-end speech recognition objectives,” in *Proc. WASPAA*, 2019, pp. 234–238.
- [16] 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,” *arXiv preprint arXiv:2201.06685*, 2022.
- [17] W. Zhang, J. Shi, C. Li, S. Watanabe, and Y. Qian, “Closing the gap between time-domain multi-channel speech enhancement on real and simulation conditions,” in *Proc. WASPAA*, 2021, pp. 146–150.
- [18] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Proc. NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.
- [19] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, “Pushing the limits of semi-supervised learning for automatic speech recognition,” *arXiv preprint arXiv:2010.10504*, 2020.
- [20] W.-N. Hsu, Y.-H. H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, “HuBERT: How much can a bad teacher benefit ASR pre-training?” in *Proc. ICASSP*, 2021, pp. 6533–6537.
- [21] X. Chang, T. Maekaku, P. Guo, J. Shi, Y.-J. Lu, A. S. Subramanian, T. Wang, S.-w. Yang, Y. Tsao, H.-y. Lee *et al.*, “An exploration of self-supervised pretrained representations for end-to-end speech recognition,” in *Proc. ASRU*, 2021.
- [22] A. Pasad, J.-C. Chou, and K. Livescu, “Layer-wise analysis of a self-supervised speech representation model,” in *Proc. ASRU*, 2021.
- [23] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, “Libri-light: A benchmark for asr with limited or no supervision,” in *Proc. ICASSP*, 2020, pp. 7669–7673.
- [24] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *arXiv preprint arXiv:2110.13900*, 2021.
- [25] Y. Wang, J. Li, H. Wang, Y. Qian, C. Wang, and Y. Wu, “Wav2vec-switch: Contrastive learning from original-noisy speech pairs for robust speech recognition,” *arXiv preprint arXiv:2110.04934*, 2021.
- [26] Y. Luo and N. Mesgarani, “Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation,” *IEEE/ACM TASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [27] Y. Luo and Mesgarani, “Tasnet: time-domain audio separation network for real-time, single-channel speech separation,” in *Proc. ICASSP*, 2018, pp. 696–700.
- [28] A. Pandey and D. Wang, “Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain,” in *Proc. ICASSP*, 2019, pp. 6875–6879.
- [29] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 46–50.
- [30] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, “GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio,” in *Proc. Interspeech*, 2021.
- [31] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, “Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation,” in *Proc. ACL*, 2021, pp. 993–1003.
- [32] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” *Proc. Interspeech*, pp. 2613–2617, 2019.
- [33] 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,” *Proc. Interspeech*, pp. 1571–1575, 2018.
- [34] J. Du, Y.-H. Tu, L. Sun, F. Ma, H.-K. Wang, J. Pan, C. Liu, J.-D. Chen, and C.-H. Lee, “The ustc-ifytek system for chime-4 challenge,” *Proc. CHiME*, vol. 4, pp. 36–38, 2016.
- [35] Z.-Q. Wang, P. Wang, and D. Wang, “Complex spectral mapping for single-and multi-channel speech enhancement and robust asr,” *IEEE/ACM TASLP*, vol. 28, pp. 1778–1787, 2020.
- [36] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gauthier, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson *et al.*, “Xsede: Accelerating scientific discovery computing in science & engineering, 16 (5): 62–74, sep 2014,” *URL https://doi.org/10.1109/mcse*, vol. 128, 2014.
- [37] N. A. Nystrom, M. J. Levine, R. Z. Roskies, and J. R. Scott, “Bridges: A uniquely flexible HPC resource for new communities and data analytics,” in *Proc. XSEDE*, 2015, pp. 1–8.