INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Dual-Path Style Learning for End-to-End Noise-Robust Speech Recognition

Yuchen Hu, Nana Hou, Chen Chen, Eng Siong Chng*

School of Computer Science and Engineering, Nanyang Technological University, Singapore
yuchen005@e.ntu.edu.sg

Abstract

Automatic speech recognition (ASR) systems degrade significantly under noisy conditions. Recently, speech enhancement (SE) is introduced as front-end to reduce noise for ASR, but it also suppresses some important speech information, *i.e.*, over-suppression. To alleviate this, we propose a dual-path style learning approach for end-to-end noise-robust speech recognition (DPSL-ASR). Specifically, we first introduce clean speech feature along with the fused feature from IFF-Net as dual-path inputs to recover the suppressed information. Then, we propose style learning to map the fused feature close to clean feature, in order to learn latent speech information from the latter, *i.e.*, clean "speech style". Furthermore, we also minimize the distance of final ASR outputs in two paths to improve noise-robustness. Experiments show that the proposed approach achieves relative word error rate (WER) reductions of 10.6% and 8.6% over the best IFF-Net baseline, on RATS and CHiME-4 datasets respectively[1].

Index Terms: Dual-path style learning, consistency loss, noise-robust speech recognition, over-suppression problem

1. Introduction

Automatic speech recognition (ASR) has achieved a great success with recent advances of deep learning techniques [1–7], which has been widely used in practical applications. However, it is still a challenging task when put under extremely noisy conditions, particularly in radio communication speech [8], which are distorted by ambient noise as well as communication channel due to limited transfer bandwidth.

Prior works usually introduce speech enhancement (SE) [9–15] as a front-end pre-processing module to reduce the additive noise and improve speech quality for downstream ASR task. However, recent works [16–18] observe that the SE processing could not always improve ASR performance, as some important speech content information for ASR in original noisy speech [19] are also reduced by SE, together with the additive noise. Such over-suppression problem is usually undetected during the speech enhancement stage, but could significantly degrade the performance of downstream ASR task.

To alleviate this issue, prior work [19] proposes an interactive feature fusion network (IFF-Net) with joint SE-ASR framework. Specifically, it designs a CNN-Attention based architecture to interactively fuse the over-suppressed enhanced speech with original noisy speech, in order to recover the lost information in enhanced speech for downstream ASR. It achieves improvements but over-suppression problem can still be observed.

In this paper, we propose a novel dual-path style learning approach for end-to-end noise-robust automatic speech recognition (DPSL-ASR) to further alleviate over-suppression problem. Specifically, we first introduce clean speech feature along with the fused feature from IFF-Net [19] as dual-path inputs to the back-end ASR module, where the clean feature can complement the suppressed information in corresponding fused feature. We then propose a style learning method to map fused feature close to clean feature, in order to learn latent speech information from the latter, *i.e.*, clean "speech style". Furthermore, we employ consistency loss to minimize the distance of ASR outputs in two paths to improve noise-robustness. Experimental results show that the proposed approach significantly outperforms the best IFF-Net baseline, and further visualizations of intermediate embeddings indicate that DPSL-ASR can effectively recover the over-suppressed information by SE.

2. DPSL-ASR Architecture

2.1. Overview

In this work, we propose a novel dual-path style learning system for end-to-end noise-robust automatic speech recognition (DPSL-ASR), which is illustrated in Figure 1 (c).

We first examine a joint training system [12, 15] in Figure 1(a) by cascading the front-end SE module and back-end ASR module via multi-task learning strategy. However, the over-suppressed speech generated by speech enhancement module could significantly degrade the performance of downstream ASR task. To alleviate this, recent study [19] proposed an IFF-Net to combine the enhanced Fbank feature X_E and original noisy Fbank feature X_N as a fused feature X_F to recover the over-suppressed information for ASR, as shown in Figure 1(b). The IFF-Net has achieved significant improvement of ASR performance, but over-suppression phenomenon can still be observed in its generated fused feature X_F.

To further alleviate the over-suppression problem, we first introduce clean speech feature X_C along with the fused feature X_F from IFF-Net [19] as dual-path inputs to the back-end ASR module, as shown in Figure 1(c). Such parallel clean feature can provide more complementary information that has been suppressed in fused feature. In addition, it can also benefit the ASR training, especially at the early training stage when the poorly-trained speech enhancement module and IFF-Net cannot provide high-quality fused feature X_F for ASR. Then, we propose a novel style learning method to map fused feature close to the clean feature, in order to learn latent speech information from the latter, *i.e.*, clean "speech style". Furthermore, we employ consistency loss to minimize the distance of ASR outputs in two paths to improve noise-robustness. As a result, the back-end ASR module could learn more noise-robust speech repre-

*Nana Hou contributed to this work before leaving Nanyang Technological University, Singapore.

[1]https://github.com/YUCHEN005/DPSL-ASR

10.21437/Interspeech.2023-101

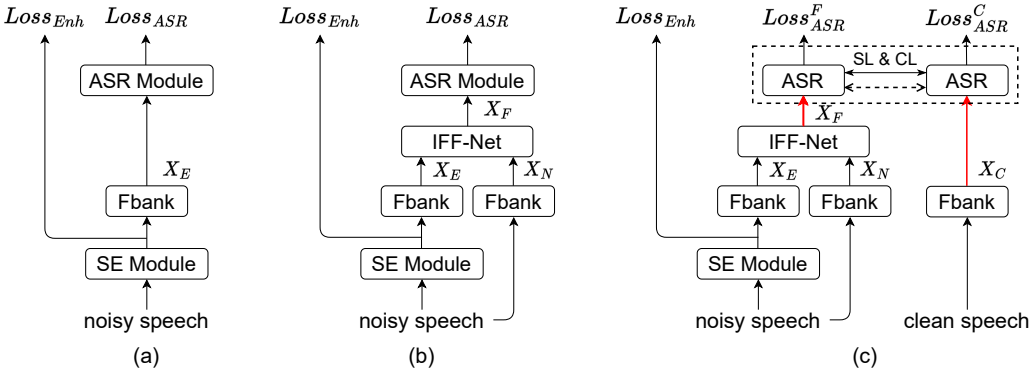

Figure 1: *Block diagrams of (a) joint SE-ASR approach, (b) IFF-Net baseline, (c) the proposed DPSL-ASR approach. The red arrows in (c) highlight the dual-path inputs. "SL & CL" in the black dashed box denotes the style learning and consistency loss. The black dashed arrow between ASR modules denotes sharing parameters, so that DPSL-ASR contains same number of parameters as IFF-Net.*

sentations to alleviate the over-suppression problem, and thus yield better performance. Since no clean data in test set is available, the clean path is discarded during inference stage.

2.2. Style Learning

Inspired by the image style transfer algorithm in prior work [20], we propose a style learning method to transfer "speech style" from clean encoded embeddings to fused embeddings, as shown in Figure 2.

Specifically, we first send clean feature and fused feature $X_C, X_F \in \mathbb{R}^{T\times F}$ into the ASR module in two paths, where T is number of time frames and F is number of frequency-bins. The ASR module in two paths share parameters with each other, which consists of L Conformer layers in the encoder and $L/2$ Transformer layers in the decoder, following the prior work [21]. Based on this, we denote the clean embedding from l-th encoder layer as E^l_C and the fused embedding from l-th encoder layer as E^l_F, where $E^l_C, E^l_F \in \mathbb{R}^{T\times D}, l \in \{1, ..., L\}$ and D is embedding size.

The "speech style" of clean embedding E^l_C and fused embedding E^l_F from l-th encoder layer are then formulated as:

$$\begin{aligned} S^l_C &= (E^l_C)^T \cdot E^l_C, \\ S^l_F &= (E^l_F)^T \cdot E^l_F, \end{aligned} \tag{1}$$

Equation 1 conducts dot product between the encoded embeddings and their own transpose, generating the style matrixes $S^l_C, S^l_F \in \mathbb{R}^{D\times D}$ that indicate correlations between every two embedding channels. Similar to the "image style" learned by neural style transfer algorithm [20], here the calculated style matrixes contain abundant latent information inside the speech feature, which is thus defined as "speech style". Then we formulate the style loss $Loss_{SL}$ as follows:

$$Loss_{SL} = \frac{1}{LD^2}\sum_{l=1}^{L} \|S^l_C - S^l_F\|^2_2, \tag{2}$$

where $\|\cdot\|_2$ denotes L_2 norm. Without specified, we employ the embeddings from all of L encoder layers to calculate the style loss. In this way, the ASR encoder would be optimized by style loss to learn abundant latent speech information from the clean embeddings, which could recover the suppressed information in fused embeddings and thus alleviate the existed over-suppression problem in IFF-Net.

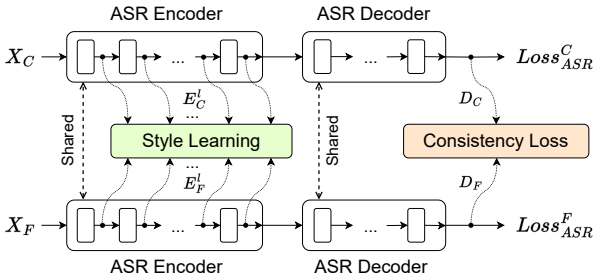

Figure 2: *Block diagram of the back-end ASR module with style learning and consistency loss in the proposed DPSL-ASR. The dashed arrows between ASR encoders and decoders both denote sharing model parameters.*

2.3. Consistency Loss

The consistency loss is designed to minimize the distance of outputs in two paths after ASR decoder via the symmetric Kullback-Leibler (KL) divergence [22,23], which is formulated as follows:

$$Loss_{CL} = Div_{KL}(D_C \parallel D_F) + Div_{KL}(D_F \parallel D_C), \tag{3}$$

where $D_C, D_F \in \mathbb{R}^{U\times V}$ are posterior outputs of ASR decoder (before softmax), in clean path and fused path respectively. U is the number of output tokens and V is the vocabulary size.

We aim to minimize the KL-divergence-based consistency loss to push the decoded embedding in fused path closer to that in clean path. As a result, the ASR output could be more robust against the noisy and over-suppression conditions.

2.4. Training Strategy

The proposed DPSL-ASR system adopts the following optimization strategy during training process:

- Seeking parameters of the SE module to minimize the mean square error loss of the speech enhancement task $Loss_{Enh}$;
- Seeking parameters of the IFF-Net and ASR module to minimize the ASR loss $Loss^C_{ASR}$ for clean path and $Loss^F_{ASR}$ for fused path;
- Seeking parameters of the ASR module to simultaneously minimize the style loss $Loss_{SL}$ for ASR encoder and the consistency loss $Loss_{CL}$ for ASR decoder.

The DPSL-ASR system is trained in an end-to-end manner, where the overall training objective $Loss_{final}$ is formulated as:

$$
\begin{aligned}
Loss_{final} &= (1-\lambda_{ASR}) \cdot Loss_{Enh} + \lambda_{ASR} \cdot Loss_{ASR} \\
&\quad + \lambda_{SL} \cdot Loss_{SL} + \lambda_{CL} \cdot Loss_{CL} \\
Loss_{ASR} &= (1-\lambda_{fused}) \cdot Loss_{ASR}^{C} + \lambda_{fused} \cdot Loss_{ASR}^{F}
\end{aligned} \tag{4}
$$

where $\lambda_{ASR}, \lambda_{SL}, \lambda_{CL}$ and λ_{fused} are weighting parameters to balance different training objectives.

3. Experiments and Results

3.1. Datasets

We conduct experiments on two corpora: one is Robust Automatic Transcription of Speech (RATS) [24], which is extremely noisy radio communication data; another is the far-field CHiME-4 dataset [25] that contains normally noisy speech.

RATS dataset consists of eight parallel channels and we only use the Channel-A subset in this work, which contains 44-hour training data, 5-hour valid data and 8-hour test data. Although RATS dataset is chargeable by LDC, a Fbank feature version is publicly available online[2].

The CHiME-4 dataset consists of three subsets: clean data, real noisy data and simulated noisy data. The clean data comes from the WSJ0 [26] training data. The real noisy data is recorded in four noisy environments, including bus, cafe, pedestrian area and street junction. The simulated noisy data is generated by mixing the clean data with background noise recorded in above four environments. We use real and simulated noisy data of 1-channel track in this work.

3.2. Experimental Setup

3.2.1. Network Configurations

The proposed DPSL-ASR system consists of three modules: the SE module, the IFF-Net and the ASR module. The SE module is same as that in [19], which utilizes 3 layer of 896-unit bidirectional long short-term memory (BLSTM) [27], followed by a 257-unit linear layer and a ReLU [28] activation function to predict masks for noisy magnitude features. The IFF-Net consists of 4 residual attention blocks with 64 filters, following the best configurations in prior work [19]. The ASR module contains L =12 Conformer [21] layers in encoder, and 6 Transformer [4] layers in decoder. The embedding dimension/feed-forward dimension/attention heads are set to 256/2048/4 for all the Conformer and Transformer layers. We also employ 1000 byte-pair-encoding (BPE) [29] tokens as ASR output.

The network is optimized by Adam algorithm [30], where the learning rate first warms up linearly to 0.002 in 25,000 steps and then decreases proportional to the inverse square root of the step number. The batch size is set to 64. The training epoch is set to 50 for experiments on RATS dataset and 100 for experiments on CHiME-4 dataset. The weighting parameters $\lambda_{ASR}, \lambda_{SL}, \lambda_{CL}$ and λ_{fused} are set to 0.7, 0.01, 0.4 and 0.3 respectively, where we first tune λ_{fused} to build $Loss_{ASR}$, and then tune λ_{ASR} to combine $Loss_{Enh}$, followed by λ_{SL} to add $Loss_{SL}$, and λ_{CL} to add $Loss_{CL}$. All the hyper-parameters are tuned on validation set.

3.2.2. Reference Baselines

We implement four competitive baselines to evaluate our proposed DPSL-ASR approach:

- **E2E-ASR** [21]: a Conformer-based end-to-end ASR system.
- **Cascaded SE-ASR** [11]: a cascaded SE and ASR system that only uses final ASR loss for optimization.

[2]https://github.com/YUCHEN005/RATS-Channel-A-Speech-Data

Table 1: *WER% results of the proposed DPSL-ASR and four baselines on RATS Channel-A dataset. "Use SL & CL" denotes whether use the style learning and consistency loss.*

Method	Use SL & CL	WER(%)
E2E-ASR [21]	✗	54.3
Cascaded SE-ASR [11]	✗	53.1
Joint SE-ASR [12]	✗	51.8
IFF-Net [19]	✗	46.2
DPSL-ASR (ours)	✗	44.1
	✓	**41.3**

Table 2: *WER% results of the proposed DPSL-ASR and four baselines on CHiME-4 1-Channel Track dataset. "Use SL & CL" denotes whether use the style learning and consistency loss. "Dev" and "Test" denote WER% results on development set and test set, respectively. "real" and "simu" denote the real noisy subset and simulated noisy subset, respectively.*

Method	Use SL & CL	Dev		Test	
		real	simu	real	simu
E2E-ASR [21]	✗	8.1	9.6	14.9	16.1
Cascaded SE-ASR [11]	✗	7.7	9.2	14.4	15.6
Joint SE-ASR [12]	✗	7.2	8.7	13.8	14.9
IFF-Net [19]	✗	6.4	7.9	12.4	13.4
DPSL-ASR (ours)	✗	6.2	7.6	12.0	12.9
	✓	**5.9**	**7.2**	**11.3**	**12.2**

- **Joint SE-ASR** [12]: a cascaded SE and ASR system that employs multi-task learning strategy for joint SE-ASR training, as shown in Figure 1(a).
- **IFF-Net** [19]: a joint SE-ASR system with IFF-Net, as shown in Figure 1(b). The IFF-Net fuses enhanced speech and original noisy speech to learn a fused representation for ASR. Therefore, the ASR module could learn complementary information to alleviate over-suppression problem.

For fair comparison, the SE modules, IFF-Nets and ASR modules in all baselines are in same structures and configurations with the proposed DPSL-ASR. Therefore, our DPSL-ASR contains same number of parameters as the IFF-Net baseline.

3.3. Results

We report experiment results in terms of word error rate (WER), as our target is ASR performance while SE is just auxiliary task.

3.3.1. DPSL-ASR vs. Other Competitive Methods

Table 1 summarizes the comparison between the proposed DPSL-ASR and the four baselines on RATS Channel-A dataset. We observe that the E2E-ASR system achieves 54.3% WER result, indicating the high difficulty of recognizing speech under extremely noisy conditions. Compared with E2E-ASR system, the cascaded SE-ASR approach reduces 1.2% WER absolutely by introducing speech enhancement as the front-end module to reduce noise for downstream ASR task. Based on this, the joint SE-ASR approach obtains another 1.3% absolute WER reduction by optimizing the SE and ASR modules simultaneously with multi-task learning strategy. Furthermore, the IFF-Net baseline improves significantly with 5.6% absolute WER reduction (51.8%→46.2%), by interactively fusing the enhanced speech and original noisy speech to complement the lost information caused by over-suppression. Finally, we observe that our proposed DPSL-ASR further alleviates this problem and obtains the best result with 10.6% relative WER reduction over the strongest IFF-Net baseline (46.2%→41.3%).

Table 3: *WER% results of the style learning and consistency loss with proposed DPSL-ASR on RATS Channel-A dataset. "Use SL" denotes whether use style learning, "Use CL" denotes whether use consistency loss. The weighting parameters λ_{SL} and λ_{CL} are kept same as those described in Section 3.2.1.*

Method	Use SL	Use CL	WER(%)
DPSL-ASR	✗	✗	44.1
	✓	✗	42.4
	✗	✓	43.3
	✓	✓	**41.3**

Table 4: *WER% results of style learning on different encoder layers, with RATS Channel-A dataset. "0" denotes do not use style learning, "1-3" denotes use it on first 3 layers, "1-12" denotes use it on all 12 layers, etc. The style losses on selected encoder layers are averaged like Equation 2, and the weighting parameter λ_{SL} is kept same as before.*

Method	Encoder Layers	WER(%)
DPSL-ASR	0	44.1
	1-3	43.9
	1-6	43.4
	10-12	43.3
	7-12	42.7
	1-12	**42.4**

Table 2 further compares the proposed DPSL-ASR with reference baselines on CHiME-4 1-Channel Track dataset. We observe that our DPSL-ASR approach obtains average relative WER reduction of 8.6% over the best IFF-Net baseline, which verifies its effectiveness under normally noisy conditions.

As a result, our proposed DPSL-ASR achieves superior performance on both extremely noisy radio-channel RATS data and the normally noisy far-field CHiME-4 data.

3.3.2. *Effect of Style Learning and Consistency Loss*

We further report the effect of proposed style learning and consistency loss on the performance of DPSL-ASR system, as presented in Table 3. Firstly we build the DPSL-ASR system with dual-path inputs, which achieves a WER of 44.1% (vs. 46.2% in IFF-Net baseline). Then, we observe that using style learning method to learn clean "speech style" can improve the WER result by 1.7% absolutely (44.1%→42.4%). Besides, applying consistency loss on decoded embeddings can also improve the performance (44.1%→43.3%). Furthermore, we can achieve the best result with 2.8% absolute WER improvement (44.1%→41.3%) by using both of them. Therefore, we conclude that both of the proposed style learning and consistency loss contribute positively to the superior performance of DPSL-ASR, where style learning plays the dominant role.

3.3.3. *Effect of Style Learning on Different Encoder Layers*

To understand the principles of style learning, we report the WER results of applying it on different encoder layers in Table 4. We observe that the style learning on low encoder layers can improve some WER performance (44.1%→43.9%/43.4%), while applying it on high encoder layers achieves significantly more improvements (44.1%→43.3%/42.7%). The reason could be that the low-layer encoded embeddings focus more on speech features with phonetic information, while high-layer embeddings contain richer semantic representations with linguistic information that is closer related to ASR task, so that more valuable clean "speech style" could be learned on high encoder layers. The best result is achieved by performing style learning on all 12 encoder layers (44.1%→42.4%). Therefore, we conclude that both low-layer and high-layer style learning can improve the ASR performance, where the latter is more effective.

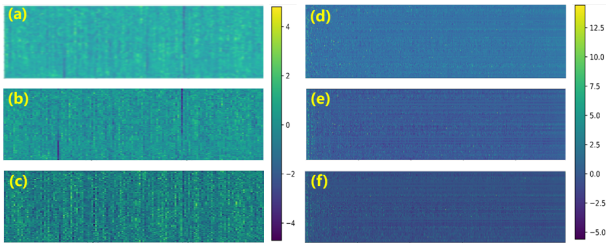

Figure 3: *Visualizations of ASR intermediate embeddings in DPSL-ASR with a RATS test sample: encoded embeddings in (a) fused path without style learning, (b) fused path with style learning, (c) clean path without style learning; and decoded embeddings in (d) fused path without consistency loss, (e) fused path with consistency loss, (f) clean path without consistency loss. The x-axis denotes embedding units and y-axis denotes time frames. The two methods are used separately with same weighting parameters ($\lambda_{SL}, \lambda_{CL}$) as before.*

3.3.4. *Visualizations of ASR Intermediate Embeddings*

To further analyze the contribution of proposed style learning and consistency loss, we visualize the ASR intermediate embeddings in DPSL-ASR with a RATS Channel-A test sample (fe_03_14342-04710-A-024073-025522-A.wav), as shown in Figure 3. We first present the encoded embeddings after final ASR encoder layer with and without style learning in (a-c). We observe that the encoded embedding (a) in fused path loses many latent speech information compared to (c) in clean path, where figure (a) looks blurred while (c) contains clear textures. After applying style learning to map encoded embeddings from fused path to clean path, the fused embedding (b) learns abundant clean "speech style" to alleviate over-suppression, *i.e.*, it shows clearer textures than (a). Furthermore, we visualize the decoded embeddings after ASR decoder with and without consistency loss in (d-f). We first observe clear difference between the decoded embedding (d) in fused path and (f) in clean path, where (d) looks bright but (f) looks dark. After applying the consistency loss, the fused embedding (e) is pushed much closer to the clean path, *i.e.*, it looks darker than (d) in figure.

4. Conclusion

In this paper, we propose a dual-path style learning approach for end-to-end noise-robust automatic speech recognition (DPSL-ASR) to alleviate the over-suppression problem existed in prior methods. In particular, we first introduce clean speech feature along with the fused feature from IFF-Net [19] as dual-path inputs to recover the over-suppressed information. Then, we propose a style learning method to map the fused feature close to clean feature, in order to learn abundant latent speech information from the latter, *i.e.*, clean "speech style". Furthermore, we also employ consistency loss to minimize the distance of final ASR outputs in two paths to improve noise-robustness. Experimental results on RATS Channel-A and CHiME-4 1-Channel Track datasets show that the proposed DPSL-ASR approach effectively alleviates the over-suppression problem and significantly outperforms the competitive baselines.

5. References

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[2] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[3] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[5] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[6] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," Technical report, OpenAI, 2022. URL https://cdn. openai. com/papers/whisper. pdf, Tech. Rep., 2022.

[8] N. Hou, C. Xu, J. T. Zhou, E. S. Chng, and H. Li, "Multi-task learning for end-to-end noise-robust bandwidth extension." in *INTERSPEECH*, 2020, pp. 4069–4073.

[9] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[10] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[11] 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 *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 234–238.

[12] B. Liu, S. Nie, S. Liang, W. Liu, M. Yu, L. Chen, S. Peng, and C. Li, "Jointly adversarial enhancement training for robust end-to-end speech recognition," in *Interspeech*, 2019, pp. 491–495.

[13] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust asr," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1778–1787, 2020.

[14] C. Zheng, X. Peng, Y. Zhang, S. Srinivasan, and Y. Lu, "Interactive speech and noise modeling for speech enhancement," in *AAAI*, 2021.

[15] A. Pandey, C. Liu, Y. Wang, and Y. Saraf, "Dual application of speech enhancement for automatic speech recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021.

[16] I. Mporas, T. Ganchev, O. Kocsis, and N. Fakotakis, "Speech enhancement for robust speech recognition in motorcycle environment," *International Journal on Artificial Intelligence Tools*, vol. 19, pp. 159–173, 2010.

[17] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 47–56, 2011.

[18] P. Wang, K. Tan, and D.-L. Wang, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 39–48, 2020.

[19] Y. Hu, N. Hou, C. Chen, and E. S. Chng, "Interactive feature fusion for end-to-end noise-robust speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6292–6296.

[20] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[21] A. Gulati, J. Qin, C. Chung-Cheng, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*, 2020, pp. 5036–5040.

[22] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[23] D. Johnson and S. Sinanovic, "Symmetrizing the kullback-leibler distance," *IEEE Transactions on Information Theory*, 2001.

[24] D. Graff, K. Walker, S. M. Strassel, X. Ma, K. Jones, and A. Sawyer, "The rats collection: Supporting hlt research with degraded audio data," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), 2014, pp. 1970–1977.

[25] E. Vincent, S. Watanabe, J. Barker, and R. Marxer, "The 4th chime speech separation and recognition challenge," *URL: http://spandh. dcs. shef. ac. uk/chime_challenge/(last accessed on 1 August, 2018)*, 2016.

[26] D. B. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.

[29] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2018, pp. 66–71.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.