Spectrograms Fusion-based End-to-end Robust Automatic Speech Recognition

Hao Shi*, Longbiao Wang†, Sheng Li‡, Cunhang Fan§, Jianwu Dang†¶, Tatsuya Kawahara*
* Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto, Japan
† Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
‡ National Institute of Information and Communications Technology (NICT), Kyoto, Japan
§ Anhui Province Key Laboratory of Multimodal Cognitive Computation,
School of Computer Science and Technology, Anhui University, 230601
¶ Japan Advanced Institute of Science and Technology, Ishikawa, Japan
E-mail: shi@sap.ist.i.kyoto-u.ac.jp

***Abstract*—To improve the robustness of automatic speech recognition (ASR), speech enhancement (SE) is often used as a front-end noise-removal process. Although there is complementarity between the mapping-based and the mask-based SE system, one of the SE systems has been conventionally used as the front-end of ASR. We propose a spectrogram fusion (SF)-based end-to-end (E2E) robust ASR system, in which the mapping-based and masking-based SE are used as the front-end simultaneously. We adopt SF to combine the advantages of mapping-based and masking-based SE systems. SF and ASR modules are connected in an E2E manner, and joint training is conducted to finetune the front-end and the back-end. We compared the performance of different front-ends after joint training. From the experiments using Aishell and PNL 100 Nonspeech Sounds datasets, we found that the fusion of two SEs are beneficial for ASR, especially under low signal-to-noise ratio, where a relative improvement of more than 7% is achieved.**

Index Terms: robust automatic speech recognition, speech enhancement, spectrograms fusion

I. Introduction

The performance of automatic speech recognition (ASR) [1] in a clean environment is very high [2]. But when noise is present, it will drop sharply [3]. For example, the word error rate (WER) drops from 1% to over 80% as the signal-noise ratio (SNR) transitions from clean to 0 dB [2]. To address this problem, many approaches have been investigated. Speech enhancement (SE) [4] is one of the approaches for improving the performance of robust ASR [5], [6]. First, an SE front-end is adopted to enhance the noisy speech signal, and then the enhanced speech signal is input into the ASR back-end to obtain the final recognition results [5], [7].

With increases in computing resources and available data, deep learning-based SE [8], [9] systems have attracted considerable attention. Because these systems have few, if any, assumptions, they can often achieve better performance than traditional SE systems [9], [10]. Mapping [9] and masking [11] are two major targets for training deep learning-based systems. Mapping-based SE systems [9], [11] use the strong nonlinear mapping capabilities of deep neural networks to directly obtain the mapping relationship between noisy features and clean features. On the other hand, masking-based SE systems [11], [12] first use a deep neural network to obtain a mask between the speech and the noisy speech. This mask is then used to extract the clean speech features from the noisy features. Although some studies have shown the complementarities [13], [14] between the mapping-based and masking-based SE systems, only one of them is still used as a front-end system in ASR [15], [16].

Meanwhile, the ASR back-end has also adopted end-to-end (E2E) models [17], [18] because of its conveniecnce. The joint training of the front- and back-end can optimize the entire pipeline [15], [19], [20]. Moreover, the joint training makes front-end more suitable for ASR [19], [21]. This will also lead to changes in the SE front-end [15] that were trained by mean squared error or mean absolute error. However, there is no work to explore the joint use of mapping-based and masking-based SE front-ends.

In this paper, we propose a spectrograms fusion [22], [23] (SF)-based E2E robust ASR system. The mapping-based and masking-based SE are combined for the front-end. They are connected to ASR in an E2E manner. Joint training [20], [24], [25] is adopted to finetune the front-end and back-end. Furthermore, we compare the performance of different front-end systems after joint training, and domenstrate the effect of joint training of mapping-based and masking-based front-end.

The remainder of this paper is structured as follows. Section 2 reviews the conventional systems. Section 3 presents the proposed system. Then, we summarize the results of this study in Section 4.

II. Conventional Approaches

Currently, there are three major approaches for improving the robustness of ASR: ASR training with noisy data, use of mapping-based and masking-based SE.

A. ASR training with noisy data

One straightforward way of improving the robustness of ASR performance is training with noisy data. Although this

Tatsuya Kawahara and Longbiao Wang are corresponding authors.

approach can boost the robustness of ASR to some degree, the complexity and computing costs are also increased. Additionally, when the training data are mismatched with the test data, the system performance will be greatly reduced. A flowchart of the approach is shown in Fig. 1(a).

B. ASR with mapping-based SE

Mapping-based SE systems use the strong nonlinear mapping capabilities of deep neural networks to directly obtain the mapping relationship between noisy features and clean features. The loss is defined as follows:

$$\mathcal{L}_{Mapping} = \frac{1}{TF}\sum ||\widetilde{|X|} - |X|||_F^2 \tag{1}$$

where T and F represent the time and frequency, respectively, $\widetilde{|X|}$ is the mapping-based speech magnitude spectrogram, and $|X|$ is the target clean speech magnitude spectrogram. With the mapping-based front-end, the ASR system uses the enhanced features as the input for the recognition process. A flowchart of this approach is shown in Fig. 1(b).

C. ASR with masking-based SE

Masking-based SE systems aim to use deep neural networks to obtain a mask between the speech and the noisy speech. This mask is intended to enable the extraction of speech signals from the noisy speech signals. The loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{Masking} &= \frac{1}{TF}\sum ||\widetilde{M} \odot |Y| - |X|||_F^2 \\ &= \frac{1}{TF}\sum |||\ddot{X}| - |X|||_F^2\end{aligned} \tag{2}$$

where $\widetilde{M}$ is the estimated mask, $\ddot{X}$ is the masking-based speech magnitude spectrogram, $|Y|$ is the noisy input magnitude spectrogram, and $\odot$ denotes point-wise matrix multiplication. With the masking-based front-end, the ASR system uses the enhanced features as the input for the recognition process. A flowchart of this approach is shown in Fig. 1(c).

III. Proposed Approach

Our proposed SF-based E2E robust ASR method is composed of three modules: an enhancement module, a fusion module, and a recognition module. A flowchart of the proposed approach is shown in Fig. 1(d).

A. Enhancement module

To obtain mapping- and masking-based spectrograms simultaneously, we train the enhancement module in a multi-target learning manner:

$$\mathcal{L}_{SE} = \alpha\mathcal{L}_{Mapping} + (1-\alpha)\mathcal{L}_{Masking} \tag{3}$$

Here, α is a hyperparameter for adjusting the loss from the two outputs.

B. Fusion module

SF [23] is an effective approach for exploiting the complementarities between mapping- and masking-based SE systems. It fuses the T-F bins from the mapping and masking spectrograms that are closest to the true labels to a single spectrogram. We use a deep neural network to estimate the minimum

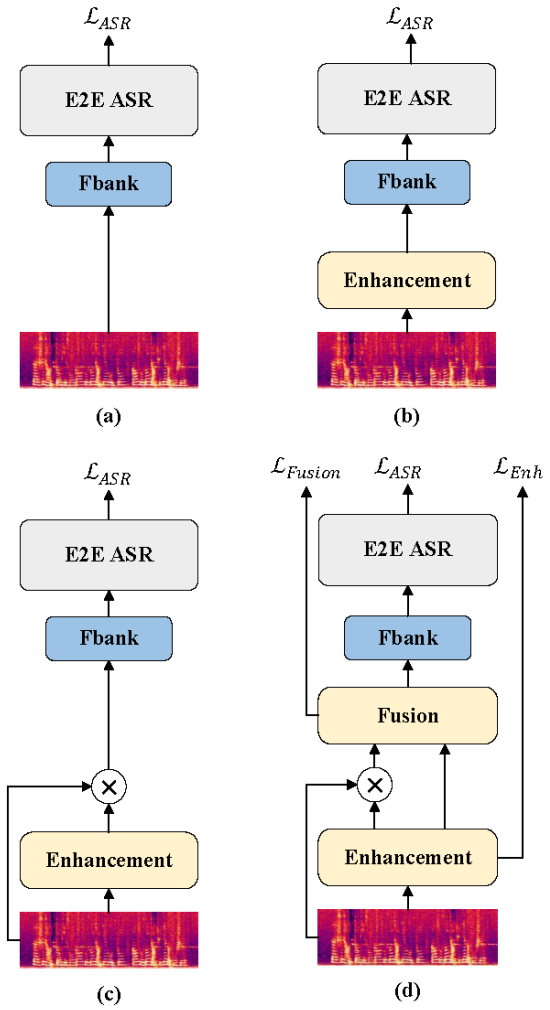

Fig. 1. Overview of robust ASR systems.

difference masks (MDMs) between $\widetilde{X}$ and $\ddot{X}$ [23]. The labels to train the MDM estimator are obtained from the enhanced and clean spectrogram. By comparing the Euclidean distance between the two spectrograms and the clean spectrogram, we set the corresponding MDM with a closer distance to 1, otherwise 0. Thus, each enhanced magnitude spectrogram has a corresponding MDM for each T-F bin, which gives a smaller absolute distance from the target magnitude spectrograms. In this paper, $\widetilde{MDM}$ and $M\ddot{D}M$ are used to extract the better parts of $\widetilde{X}$ and $\ddot{X}$, respectively.

Because the spectrogram is continuous, the MDMs in the testing stage are real values in (0, 1). The loss is defined as follows:

$$\mathcal{L}_{SF} = \frac{1}{TF}\sum_i ||\widetilde{MDM}_i - MDM_i||_F^2 \tag{4}$$

After predicting the MDMs, nonlinear selection and fusion processing are conducted to obtain the fusion speech magnitude spectrogram:

$$\widehat{X} = \widetilde{MDM} \odot \widetilde{X} + M\ddot{D}M \odot \ddot{X} \tag{5}$$

C. Recognition module

A speech transformer [17] with self-attention [26] is used for the E2E ASR component. Except for the different input features, the structure of the model is not changed. We use Fbank as the input feature of the recognition module. We can easily obtain Fbank from the enhanced features using the log Mel filterbank. Based on the cross-entropy criterion, the loss function of ASR is defined as follows:

$$\mathcal{L}_{ASR} = -\ln P(S^*|\widehat{X}) \tag{6}$$

where S^* is the ground-truth of the whole sequence of output labels.

D. Joint training

We propose a robust E2E ASR system that transforms noisy speech signals into text using a single network. The SE networks, the SF network, and ASR based on speech transformer are implemented with a single neural network. The parameters are updated by the stochastic gradient descent. SE, SF and ASR network are finetuned with joint training. The loss function of the joint training is defined as follows:

$$\mathcal{L}_{Joint} = \beta\mathcal{L}_{ASR} + (1-\beta)\mathcal{L}_{SE} + \gamma\mathcal{L}_{SF} \tag{7}$$

The hyperparameter β, γ control the loss between $\mathcal{L}_{ASR}$, $\mathcal{L}_{SF}$ and $\mathcal{L}_{SE}$.

IV. Experimental Evaluations

The enhancement module has three bidirectional long short-term memory (BLSTM) hidden layers, each having 512 nodes. The input and output were both 257-dimensional magnitude spectrograms. We used a short-time Fourier transform with a 32-ms Hamming window and a 16-ms window shift to obtain the 257-dimensional magnitude spectrograms for feature extraction. The Fbank was 80-dimensional. This module was trained in a multi-target learning manner to obtain mapping- and masking-based SE systems. The fusion module has three fully connected layers, with each hidden layer having 512 nodes. The input can be noisy, mapping-based, and masking-based magnitude spectrograms. Moreover, the output has two MDMs and two enhanced spectrograms. For the recognition module, we used the speech transformer with self-attention, under the same settings as described in [27]. Specifically, we used six self-attention blocks as encoders and six self-attention blocks as the prediction network. For each module, we performed pre-training. For pre-training enhancement module and fusion module, we used the data described in Section 3.1. For pre-training the recognition module, we used the clean data. The hyperparameter α was set to 0.5, the hyperparameter β was set to 1, and the hyperparameter γ was set to 0, meaning the ASR loss is primarily used. All model training ran for 60 epochs.

A. Dataset

We used the Aishell ASR corpus [28] and the PNL 100 Nonspeech dataset [1] to synthesize the experimental dataset. For the training set, we randomly selected 70 kinds of noise and randomly synthesized the training set from the Aishell corpus with SNR values of 0, 5, 10, 15, and 20. We did not use the development set to tune or select the system. On the other hand, we used the development and the test set from the Aishell corpus to synthesize the test sets. For test set 1, we randomly selected 15 kinds of noise, different from those in the training set, and the whole development set from the Aishell corpus with same SNR values as training set. For test set 2, we used the remaining 15 kinds of noise and randomly synthesized the test set from the Aishell corpus according to SNR values of -5, 2.5, 7.5, 12.5, and 17.5. In summary, test set 1 contained some unknown noise, while all kinds of noise of test set 2 was unknown.

[1] http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html

B. Evaluation metrics

We evaluated the performance of SE and ASR separately. To evaluate the performance of SE, we used the following evaluation metrics: Signal distortion (Csig) [29], background intrusiveness (Cbak) [29], overall quality (Covl) [29], perceptual evaluation of speech quality (PESQ). For the ASR backend, we used the character error rate (CER) as an evaluation metric.

C. Model abbreviation

In the following discussion, "Mapping_separate", "Masking_separate", and "Fusion_separate" denote the systems that directly used pre-trained modules without joint training. "Noisy", "Mapping_Joint", and "Masking_Joint" denote conventional systems with joint training, as described in Section 2. "Fusion_Joint" denotes our proposed system with joint training, as described in Section 3. "Fusion_Joint_Mapping" and "Fusion_Joint_Masking" denote two outputs after enhancement module of "Fusion_Joint".

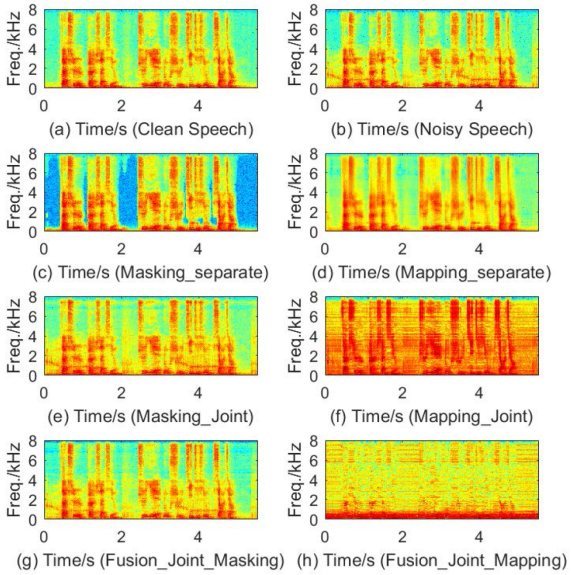

Fig. 2. Spectrograms of different SE systems: SNR is 15.

TABLE I
THE PERFORMANCE OF SE IN THE TEST SETS.

Systems	Test set 1				Test set 2			
	Csig	Cbak	Covl	PESQ	Csig	Cbak	Covl	PESQ
Original noisy	2.800	**2.642**	2.115	1.499	2.365	**2.280**	1.827	1.408
Mapping_separate	3.456	1.817	2.734	2.040	2.805	1.595	2.187	1.646
Masking_separate	3.092	1.664	2.428	1.830	2.552	1.498	1.999	1.549
Fusion_separate	**3.568**	1.825	**2.799**	**2.065**	**2.868**	1.583	**2.215**	**1.650**
Mapping_Joint	1.243	1.191	1.133	1.149	1.124	1.211	1.084	1.165
Masking_Joint	2.737	1.487	2.070	1.481	2.290	1.412	1.786	1.404
Fusion_Joint_Mapping	1.531	1.251	1.207	1.074	1.496	1.277	1.210	1.098
Fusion_Joint_Masking	2.634	1.462	2.033	1.521	2.116	1.325	1.688	1.407
Fusion_Joint	1.053	1.091	1.024	1.083	1.011	1.080	1.014	1.124

TABLE II
CER RESULTS (%) OF DIFFERENT E2E SYSTEMS WITH TEST SET 1: THE SNRS OF THE TEST SET ARE KNOWN; THE NOISE OF THE TEST SET IS UNKNOWN.

β	System	CER Results(%) on Test Set 1 (Seen SNRs, Unseen Noise)					
		20dB	15dB	10dB	5dB	0dB	AVG.
-	Noisy	12.51	14.60	17.31	24.62	39.65	21.78
	Mapping_separate	14.58	19.40	29.33	51.65	81.14	39.32
	Masking_separate	19.96	33.22	53.33	83.04	111.28	60.27
	Fusion_separate	13.34	18.74	29.69	55.14	89.79	41.45
0.5	Mapping_Joint	11.67	12.97	16.37	25.49	43.63	22.08
	Masking_Joint	10.18	11.56	13.94	20.17	36.39	18.49
	Fusion_Joint	9.82	11.59	15.43	26.76	49.89	22.76
1	Mapping_Joint	11.08	12.33	15.58	22.83	40.21	20.45
	Masking_Joint	9.83	11.38	13.78	20.18	36.05	18.29
	Fusion_Joint	**9.77**	**11.00**	**12.93**	**19.04**	**33.45**	**17.28**

TABLE III
CER RESULTS (%) OF DIFFERENT E2E SYSTEMS WITH TEST SET 2: BOTH THE SNRS AND THE NOISE OF THE TEST SET ARE UNKNOWN.

β	System	CER Results(%) on Test Set 1 (Unseen SNRs, Unseen Noise)					
		17.5dB	12.5dB	7.5dB	2.5dB	-5dB	AVG.
-	Noisy	20.31	24.54	30.45	45.41	76.66	39.99
	Mapping_separate	25.14	35.75	58.11	91.47	125.36	68.43
	Masking_separate	31.13	48.81	78.56	113.53	157.40	87.42
	Fusion_separate	23.77	36.33	61.57	97.53	139.44	73.13
0.5	Mapping_Joint	16.53	20.71	30.96	54.86	102.55	45.93
	Masking_Joint	14.19	17.10	23.77	38.03	83.29	35.81
	Fusion_Joint	14.86	19.88	33.26	57.94	106.59	47.40
1	Mapping_Joint	16.41	19.63	27.87	45.70	96.42	41.85
	Masking_Joint	14.04	**16.68**	24.01	37.69	75.91	34.18
	Fusion_Joint	**13.75**	16.75	**23.03**	**36.81**	**73.86**	**33.34**

D. Impact of SF-based ASR without joint training

Comparing the spectrograms in Figure 2, we find that "Mapping_separate" retains some non-speech components as speech, and "Masking_separate" causes loss of information; "Fusion_separate" was similar to "Mapping_separate", but some high-frequency information is lost. Although "Fusion_separate" achieves the best performance in SE tasks, "Mapping_separate" achieves better performance on ASR. The ASR back-ends of "Mapping_separate", "Masking_separate", and "Fusion_separate" are all trained on clean data. Comparing the results of Table 2 and Table 3, high SNR is beneficial to the "Fusion_separate". However, when the SNR is low, the poor recognition performance of "Masking_separate" affects the "Fusion_separate". The performance of the fusion system is between the two fused systems.

E. Impact of SF-based ASR with joint training

Joint training has different effects on "Fusion_Joint", "Mapping_Joint" and "Masking_Joint". From Figure 2 we find that "Masking_Joint" restores the previously lost information with blurred speech information . "Mapping_Joint" will introduce a lot of new noise, but the energy of the speech is very obvious. During joint training with β=1.0, the system did not introduce enhanced loss. So the mapping module did not work for enhancement but became a part of recognition network. On the other hand, the masking module still works for enhancement because of the architecture of masking. It provides better enhanced spectrogram.

In "Fusion_Joint", the fusion module fuses "Fusion_Joint_Masking" and "Fusion_Joint_Masking". "Fusion Joint Masking" restores spectrogram from "Noisy Speech". The loss in speech signal "Fusion Joint Mapping" becomes serious, as it does no longer work for enhancement, but focuses on feature extraction for recogition.

The ASR performance of the systems obtained by using joint training is greatly improved. We explored whether it is necessary to introduce an enhancement loss ($\beta = 0.5$) during joint training. The results indicate that it does not improve the ASR performance, but instead dramatically degrades it. We compared the two parts of the loss and found that the

enhancement loss is larger than the ASR loss, which may affect the convergence of the ASR part. With β set to 1, the recognition rate constantly improved as the SNR improves. "Fusion_Joint" gives improved results in almost all cases, especially when the SNR is low. In the case of 0 dB SNR, "Fusion_Joint" gives a relative improvement of more than 7% compared with "Masking_Joint", and close to 17% compared with "Mapping_Joint". This shows that leveraging the complementarities between mapping- and masking-based SE systems is effective for robust ASR, especially when the noise is large (i.e., low SNR).

When the noise and SNR are both unknown, the performance of "Fusion_Joint" is improved compared with "Mapping_Joint" and "Masking_Joint", though the improvement is not large. This may be because "Fusion_Joint" benefits from better enhancement systems. Improving robustness of deep-learning-based SE for unseen noise is important.

V. Conclusions and Future Work

We proposed an SF-based E2E robust ASR system. From a series of experiments, we showed that the joint training is very important for robust ASR. Joint training had different effect on front-ends; masking-based front-end blurs some speech details and mapping-based front-end introduces some noise, but the energy of the speech are kept. Fusion-based front-end will highlight the low-frequency and some high-frequency components. The introduction of a front-end gives only a slight improvement, and may even degrade the performance. This was solved by SF, which improves the performance of robust ASR at low SNRs. The proposed SF-based E2E ASR system demonstrates that the combination of mapping- and masking-based front-ends improves the robustness of ASR. In future work, we will examine more fusion approaches to improve the robustness of ASR.

VI. Acknowledgements

This work was supported partially by the National Natural Science Foundation of China under Grant 61771333.

References

[1] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007.

[2] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Proc. Interspeech*, 2012, pp. 22–25.

[3] A. Acero and R. M. Stern, "Environmental robustness in automatic speech recognition," in *Proc. ICASSP*, 1990, pp. 849–852.

[4] P. C. Loizou, *Speech enhancement: theory and practice*, 2013.

[5] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR," in *Proc. LVA/ICA*. Springer, 2015, pp. 91–99.

[6] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Unsupervised speech enhancement based on multichannel nmf-informed beamforming for noise-robust automatic speech recognition," *IEEE/ACM TASLP*, vol. 27, no. 5, pp. 960–971, 2019.

[7] 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.

[8] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, 2013, pp. 436–440.

[9] Y. Xu, J. Du, L. Dai, and C. Lee, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[10] R. Gomez and T. Kawahara, "Optimizing spectral subtraction and wiener filtering for robust speech recognition in reverberant and noisy conditions," in *Proc. ICASSP*, 2010, pp. 4566–4569.

[11] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM TASLP*, vol. 26, no. 10, pp. 1702–1726, 2018.

[12] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, "Improving Noise Robust Automatic Speech Recognition with Single-Channel Time-Domain Enhancement Network," in *Proc. ICASSP*, 2020, pp. 7009–7013.

[13] Y. Xu, J. Du, Z. Huang, L. Dai, and C. Lee, "Multi-objective Learning and Mask-based Post-processing for Deep Neural Network based Speech Enhancement," in *Proc. Interspeech*, 2015, pp. 1508–1512.

[14] L. Sun, J. Du, L. Dai, and C. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Proc. HSCMA*, 2017, pp. 136–140.

[15] T. Menne, R. Schlüter, and H. Ney, "Investigation into Joint Optimization of Single Channel Speech Enhancement and Acoustic Modeling for Robust ASR," in *Proc. ICASSP*, 2019, pp. 6660–6664.

[16] D. Bagchi, P. Plantinga, A. Stiff, and E. Fosler-Lussier, "Spectral Feature Mapping with MIMIC Loss for Robust Speech Recognition," in *Proc. ICASSP*, 2018, pp. 5609–5613.

[17] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 5884–5888.

[18] A. Gulati, J. Qin, C. Chiu, 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 *Proc. Interspeech*, 2020, pp. 5036–5040.

[19] M. Mimura, S. Sakai, and T. Kawahara, "Joint optimization of denoising autoencoder and dnn acoustic model based on multi-target learning for noisy speech recognition," in *Proc. Interspeech*, 2016, pp. 3803–3807.

[20] Y. Xu, C. Weng, L. Hui, J. Liu, M. Yu, D. Su, and D. Yu, "Joint training of complex ratio mask based beamformer and acoustic model for noise robust asr," in *Proc. ICASSP*, 2019, pp. 6745–6749.

[21] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Joint optimization of neural network-based wpe dereverberation and acoustic model for robust online asr," in *Proc. ICASSP*, 2019, pp. 6655–6659.

[22] H. Shi, L. Wang, S. Li, C. Ding, M. Ge, N. Li, J. Dang, and H. Seki, "Singing Voice Extraction with Attention-Based Spectrograms Fusion," in *Proc. Interspeech*, 2020, pp. 2412–2416.

[23] H. Shi, L. Wang, M. Ge, S. Li, and J. Dang, "Spectrograms Fusion with Minimum Difference Masks Estimation for Monaural Speech Dereverberation," in *Proc. ICASSP*, 2020, pp. 7544–7548.

[24] T. Gao, J. Du, L. Dai, and C. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Proc. ICASSP*, 2015, pp. 4375–4379.

[25] Z. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM TASLP*, vol. 24, no. 4, pp. 796–806, 2016.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[27] Z. Tian, J. Yi, J. Tao, Y. Bai, and Z. Wen, "Self-Attention Transducers for End-to-End Speech Recognition," in *Proc. Interspeech*, 2019, pp. 4395–4399. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2203

[28] X. N. B. W. H. Z. Hui Bu, Jiayu Du, "AIShell-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline," in *Proc. Oriental COCOSDA*, 2017, p. Submitted.

[29] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE TASLP*, vol. 16, no. 1, pp. 229–238, 2008.