Learning Consistent Embedding Distribution for Robust ASR

Zuyao Ma[1(✉)], Yulong Wang[1,2], Tongcun Liu[2], Lei Zhang[2], and Wei Li[2]

[1] Beijing University of Posts and Telecommunications, Beijing, China
mazuyao233@bupt.edu.cn
[2] E-Byte Technologies, Chengdu, China
{liutongcun,zhanglei,liwei,wangyulong}@ebupt.com

Abstract. Despite the success achieved by existing Automatic Speech Recognition (ASR) models, they are highly dependent on the sufficiency of labeled clean training data, which is unrealistic in practice due to expensive labeling costs and unpredictable noise. To address this challenge, we propose a novel Distribution Transformation network (DT-net), which attempts to refine the pre-trained embeddings to mitigate the influence brought by noise. The proposed DT-net consists of a front-end Speech Enhancement (SE) module and a back-end Automatic Speech Recognition (ASR) module. Besides, two types of novel distribution transformations are introduced into the SE and ASR module respectively to adapt to the distributions of clean and noisy pre-trained embeddings. Extensive experiments conducted in public datasets, CHiME-4, reveal that the proposed DT-net outperforms other baselines in terms of both recognition performance and robustness.

Keywords: Robust ASR · Distribution transformation · Speech enhancement

1 Introduction

Automatic speech recognition (ASR) [2] is a technology that can automatically convert spoken language into written text. ASR has become increasingly important in recent years, because it automates many tasks that require human intervention, which includes controlling smart homes, providing intelligent customer service, and enabling voice assistants and voice input. As an active research topic in computer science, linguistics, and cognitive psychology, ASR has appealed to more and more researchers to improve the performance of ASR and applications.

Given sufficient clean training data, existing models have achieved satisfactory performance on various ASR tasks. However, high-quality training data is often unavailable due to expensive manpower and material resources. In real scenarios, researchers are confronted with audio data containing different types of noises. Previous literature has proven that existing models' recognition performance will dramatically degrade when trained on noisy data. Recently, many researchers have developed various approaches to address this issue. For example,

W. Yu and L. Xuan (Eds.): MLICOM 2023, LNICST 518, pp. 3–14, 2024.
https://doi.org/10.1007/978-3-031-71716-1_1

from the pre-processing perspective, speech enhancement (SE) [3–6] attempts to remove the noise of data in the time domain or frequency domain; From the model perspective, [7–9] considers redesigning models (e.g., adding a denoiser module) to improve their robustness. Some other researchers even introduce side information such as visual and positional information, to mitigate the negative effects brought by noise [10,11].

Despite the success achieved by existing SE modules, their outputs are not necessarily suitable for downstream ASR tasks, because some potential information in the original audio signal [12–14] may be removed, which results in the over-suppression phenomenon. That is, ASR models show robustness to noisy data after enhancement while observing an obvious performance degradation in clean data. Although the over-suppression phenomenon can be alleviated by training SE and ASR models jointly with a cascaded framework, as a whole, the recognition performance will be degraded on both noisy and clean data.

Recently, pre-trained models have witnessed significant progress and can provide high-quality speech embeddings to facilitate various downstream tasks [16–18]. To fully utilize the superior generalization capacity, instead of improving the raw signal quality, [15] directly performed speech enhancement on the pre-trained speech embeddings, which provides a promising direction for robust ASR. Compared to time-domain or frequency-domain encodings, encoding by the pre-trained model can capture more comprehensive features. However, it is often ignored that the pre-trained model is usually trained by clean data, which inevitably becomes sensitive to noise, i.e., the output embedding of clean data and its noisy counterpart may show a significant difference, leading to erroneous results for downstream tasks. Therefore, in this paper, we are focused on the enhancement of the output embedding of pre-trained speech models.

To improve the recognition performance of existing end-to-end ASR systems, we propose a novel Distribution Transformation network (DT-net), which attempts to refine the pre-trained embeddings to mitigate the influence brought by noisy speech. To be specific, we first feed the clean and noisy pre-training embedding pairs to a front-end Speech Enhancement (SE) module to obtain corresponding advanced features. Then, two types of novel Distribution Transformations (DTs) are introduced to make the SE module adapt to the distributions of clean and noisy pre-trained embedding. Afterward, to avoid information loss brought by the SE module, we utilize the optimal transport (OT) [19] loss to maximize the mutual information between pre-trained embeddings and corresponding extracted features. Finally, an MSE loss is adopted to pull close the clean features and their noisy counterparts. The main contributions of this work are listed as follows:

– A speech enhancement framework, DT-net, is proposed to improve the recognition performance of existing ASR systems. Two types of novel distribution transformation strategies are introduced to adapt to the distributions of clean and noisy pre-trained embedding.

– We optimize the enhancement model and ASR model collaboratively by combining the optimal transport theory and dynamic scaling strategy, which can ensure the utilization of information adequately.
– Extensive experiments were conducted on the CHiME-4 dataset, and the results shown that DT-net has lower word error rate and better generalization compared with all baseline approaches.

2 Related Work

With well-processed clean data, existing ASR models such as RNN-T [20] and T-T [21] can be trained to take good recognition performance, demonstrating its impressive capacity. In fact, some models such as Hubert [17] and XLS-R [22] have even surpassed human-level performance in large-scale speech recognition tasks. However, in noisy environments, the quality of speech signals deteriorates, which affects the feature extraction and pattern-matching processes, leading to significant decreases in the performance of these models. Additionally, recognizing speech in noisy environments requires more complex algorithms and computational resources, which increases the computational complexity and cost of the system. To address this issue, researchers have proposed various speech recognition techniques for noisy environments, such as adjusting the acoustic model [24], optimizing feature extraction [18] and knowledge distillation algorithms [23] to improve the system's robustness and accuracy, using filtering and noise reduction techniques to improve the quality of speech signals before they are fed into the speech recognition system [25], and combining speech recognition technology with other sensor information, such as image and location data [10,11], to improve recognition accuracy and robustness and reduce the impact of noise on the system.

To improve the recognition performance on noisy speech, considerable research has been conducted to integrate front-end speech enhancement modules as a preprocessing stage into speech-processing systems. For instance, frequency-domain-based enhancement methods, such as traditional spectral subtraction [26] and short-time noise gating, have been replaced by deep learning methods. The basic paradigm involves using a network to output similar clean signals and corresponding noise signals. However, frequency-domain-based speech enhancement methods process speech signals in the frequency domain, which may not effectively handle some time-domain features, resulting in inaccurate processing of some time-correlated noise. To address this issue, researchers have explored deep speech enhancement methods in the time domain, which directly process the original audio signal to generate clearer speech signals [27–29]. These methods have the advantage of handling nonlinear distortion and retaining more time-domain information to generate higher-quality speech signals. However, time-domain speech enhancement methods may introduce some degree of time-domain distortion and have a strong dependency on noise, which may affect the quality and clarity of speech signals and require more resources.

3 Proposed Method

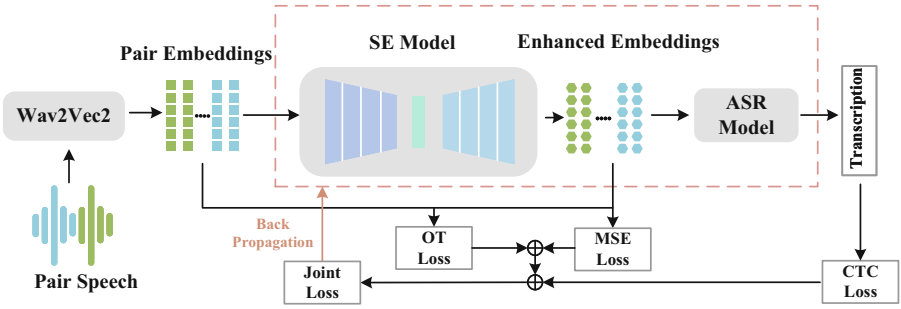

Fig. 1. Overview of our model.

In this section, we will introduce our DT-net, which consists of a front-end Speech Enhancement (SE) module and a back-end Automatic Speech Recognition (ASR) module. To further enhance the denoising effect on the embedded features, we propose Generalization Distribution Transformation (GDT) and Focus Distribution Transformation (FDT) to constrain the distribution consistency between clean speech embedding and noisy speech embedding. Additionally, we use the OT theory to supervise the generation of enhanced embedding and fully utilize the information in the source. Then the generated enhanced embedding is used as input for the subsequent ASR module. During the training phase, we adjust the loss function to improve the results. The overall structure of the model is shown in Fig. 1.

3.1 Front-End Speech Enhancement

Given a clean speech signal $x[n]$ and an additive noise signal $v[n]$, the noisy speech mixture signal obtained by microphone collection is $z[n] = x[n] + v[n]$, where n represents the time index, we aim to extract informative signals from the mixture signal. Firstly, we use the pre-training model, Wav2Vec2 [16], to encode $x[n]$ and $z[n]$ for extracting the corresponding clean speech vector $e_x[n]$ and noisy speech vector $e_z[n]$. Next, we introduce a front-end SE module to learn a speech estimate $\widetilde{x}$ from the mixture signal z with a generation function $f(\cdot)$, making this speech estimate $\widetilde{x}$ as close as possible to the clean speech signal x. For simplicity, we use e_x, z_x to denote $e_x[n], z_x[n]$ respectively.

$$e_x \approx f(e_z, \Theta_{se}) \tag{1}$$

where Θ_{se} is the trainable parameter of SE. Specifically, our SE module uses the U-net [30] as the underlying architecture for enhancement, because the U-net has a simple and efficient network architecture, demonstrating a remarkable capacity

for extraction and restoration of semantic information. The U-net architecture consists of a contracting path, a bottleneck, and an expansive path. The contracting path follows the standard architecture of a convolutional network, which consists of the repeated application of two convolutions with a kernel size of 3, and each one is followed by a BatchNorm and a rectified linear unit (ReLU). The number of feature channels is doubled at each downsampling step. Each step in the expansive path involves an upsampling of the feature map, which halves the number of feature channels, followed by a concatenation with the corresponding cropped feature map from the contracting path, and then two convolutions, each followed by a ReLU activation function. The intermediate results and outputs of each layer can be represented as follows:

$$\begin{aligned} e_z^{mid} &= ReLU\Big(BatchNorm\big(Conv2d(e_z^{in})\big)\Big) \\ e_z^{out} &= ReLU\Big(BatchNorm\big(Conv2d(GDT(e_z^{mid}))\big)\Big) \end{aligned} \tag{2}$$

The bottleneck layer is a dense layer with the GDT, which will be introduced in the next subsection. The structure of the SE was shown in Fig. 2.

3.2 Generalization Robustness Advancement

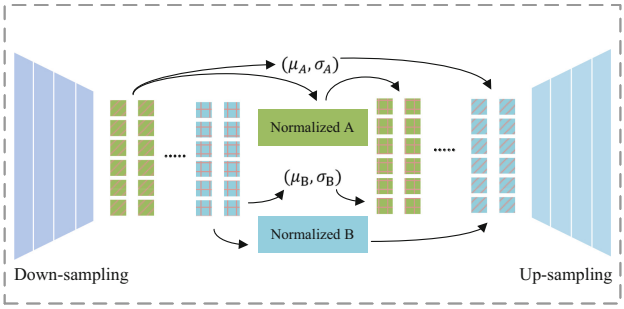

Fig. 2. speech enhancement model architecture.

Due to the distribution inconsistency between clean speech and noisy speech, the performance of existing ASR models well trained by clean speech will be dramatically degraded when confronted with noisy speech. Furthermore, different noise types and intensities also affect the model's generalization capacity. Therefore, it is important to train the SE module to extract shared features from clean speech embeddings and noisy speech embeddings, mitigating the issue of distribution inconsistency. To this end, we propose novel FDT and GDT to jointly transform clean data and noisy data into a shared vector space by data normalization, where FDT is used in the ASR module to replace the Batch Normalization (BN) in both training and testing phases, and GDT is used in the

bottleneck of SE module with a certain probability only in the training phase. In particular, FDT is designed to extract key information from noisy data and is trained to obtain appropriate affine parameters, while GDT is designed to enrich the distribution of data by swapping the mean and variance of clean data and noisy data, enhancing the robustness of the model. Given specific speech embeddings $A \in \mathbb{R}^{N \times M}$ and its instance normalization $\frac{A-\mu_A}{\sigma_A}$, where N is the length of the speech embedding, M is the feature dimension, and μ_A and σ_A are the mean and variance, an affine transformation (AT) based on the instance normalization is as follows:

$$f_{AT}(A) = \gamma \frac{A - \mu_A}{\sigma_A} + \beta \tag{3}$$

where γ and β are self-defined affine parameters. Different from the above transformation, FDT attempts to train a network g to obtain γ and β based on μ_A and σ_A, which can be expressed as:

$$f_{FDT}(A) = g_\gamma(\mu_A, \sigma_A) \frac{A - \mu_A}{\sigma_A} + g_\beta(\mu_A, \sigma_A) \tag{4}$$

FDT attempts to reconstruct the original speech data A. Furthermore, the mean and variance of data by FDT are accordingly adjusted to emphasize the crucial information of inputs. g_γ and g_β are implemented using fully connected networks. However, FDT is only applied to speech data of a single domain (clean data and noisy data), and the distribution gap still exists and damages the recognition performance. To bridge this gap, we introduce GDT to strengthen the generalization capacity by augmenting the shared parts of distributions of clean speech and noisy speech. Particularly, based on instance normalization, we swap the corresponding mean and variance of embeddings in two domains:

$$\begin{aligned} f_{GDT}^{A \to B} &= \sigma_A \frac{B - \mu_B}{\sigma_B} + \mu_A \\ f_{GDT}^{B \to A} &= \sigma_B \frac{A - \mu_A}{\sigma_A} + \mu_B \end{aligned} \tag{5}$$

3.3 Automatic Speech Recognition

In this section, we elaborate on the used ASR model, which is set as Squeezeformer [1] for its superior recognition performance. Concretely, the ASR model comprises the Temporal U-Net structure for Reducing the temporal redundancy in learning adjacent speech frame features, the standard Conformer-style block structure that only uses Post-Layer Normalization and the depthwise separable subsampling layer.

3.4 Training Strategy

During the training process, we employ a joint training approach with a warm-up strategy to optimize the model. We first train the SE model and then perform

joint fine-tuning of the front-end and back-end using the same data. Such a training strategy facilitates faster convergence while preventing severe penalties to the front-end due to the large size of the back-end ASR model during the training process. Next, we show more details for the two training processes.

SE Training: Previous research [12–14] has shown that the associated information of speech data can be lost after enhancement. Moreover, over-suppression is a common issue in enhancement models, where the model performs well on noisy data but poorly on clean data. To address these issues, we utilize optimal transport (OT) techniques to preserve the shared information between X and Y as much as possible. For the over-suppression issue, we aim to minimize the difference between the enhanced embeddings of clean data and the corresponding noisy data using mean squared error (MSE) loss. Finally, we optimize our model using both OT loss and MSE loss, as follows:

$$L_{se} = L_{mse} + L_{ot} \tag{6}$$

$$L_{mse} = \sum_{n} ||f(e_z[n], \Theta_{se}) - f(e_x[n], \Theta_{se})||^2 \tag{7}$$

$$L_{ot} = \min_{\mathbf{T} \in \Pi(\mu_s, \mu_t)} < \mathbf{C}, \mathbf{T} > + \epsilon H(\mathbf{T}) \tag{8}$$

where L_{mse} represents the mean square error loss between the enhanced clean speech embedding and the noise embedding, while L_{ot} represents the optimal transport loss. Here, L_{mse} is the mean squared error loss between the enhanced clean embedding and the noisy embedding, and L_{ot} represents the optimal transport loss.

The set $\Pi(\widetilde{e}_z, e_x) = \{\mathbf{T} \in \mathbb{R}^{N \times N} | \mathbf{T}\mathbf{1} = \mu_s, \mathbf{T}^\top \mathbf{1} = \mu_t\}$, where N is the length of the embedding vectors, and $< ... >$ denotes the Hadamard product of matrices. The term $H(\mathbf{T}) =< \mathbf{T}, log\mathbf{T} >$ is a regularization term, and ϵ is the hyper-parameter controlling the importance of the entropy term. The matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$ measures the cost of transporting from the source to the target. By using Sinkhorn's algorithm to solve the optimal transport problem, we can obtain the transportation plan $\mathbf{T}$ that minimizes the cost under $\mathbf{C}$. This allows us to obtain an OT loss that can learn the semantic similarity between the source and target.

Joint Training: Having optimized the parameters of SE model, we next jointly fine-tune the front-end and back-end using the same data used in SE training. The loss function for the joint training is as follows:

$$L_{joint} = \alpha L_{se} + (1 - \alpha) L_{asr}^{\gamma} \tag{9}$$

Here, L_{se} and L_{asr} denote the losses for the speech enhancement module and the Automatic Speech Recognition (ASR) module, respectively, and α is a weight hyperparameter to balance the two loss terms. Since it is more challenging to extract features from noisy speech than from clean speech, we use a modulation factor γ to control the learning difficulty for the easy samples and hard samples.

Table 1. Results presenting the % Word Error Rate (WER) of our proposed DT-net model on the superimposed CHiME-4 ASR task. Previously published papers and our reproduction of the baselines using WAVE-U-Net, IFF-net, Emb-Enh are shown for comparison.

Method	Param	WER		
		dt05_real	et05_real	clean
WAVE-U-Net	10 M	11.4	18.1	5.3
IFF-Net	1.49 M	10.4	16.8	4.9
Emb-Enh(CNN-4)	986 K	10.6	16.2	5.1
Emb-Enh(U-Net-2)	38 M	9.5	15.7	4.8
DT-net	3.76 M	**9.4**	**15.2**	**4.3**

4 Experiments

4.1 Experimental Setup

Dataset. We evaluated our method on the CHiME-4 dataset, which was collected by having volunteers read the speaker-independent medium (5k) vocabulary subset of the Wall Street Journal (WSJ0) corpus. The dataset includes both real and artificially simulated noisy speech data. **Baseline.** We compare our methods with three popular robust ASR models, WAVE-U-Net [27], IFF-Net [31], Emb-Enh [32], where the Emb-Enh (CNN-4) and Emb-Enh (U-net-2) denote Emb-Enh with two different backbones, CNN-4 and U-net2, respectively. In addition, we also take the Squeezeformer as the baseline. **Network Configurations.** The proposed DT-net consists of two modules SE and ASR. In particular, for the SE module, a U-net with 4 upsampling and downsampling modules is used as the backbone where the dimension of input is set as 512, the convolution kernel and maximum pooling kernel are set as 3 and 2 per upsampling module and downsampling module. As for ASR, the Squeezeformer with 16 layers is employed as the encoder and a single dense layer is used as the decoder. **Parameters.** In the joint training process, both the SE and ASR networks are optimized using the Adam algorithm with a warm-up phase, where the learning rate is linearly ramped up to 0.0001 in the first 10,000 steps, and then decreased by a factor of 0.985 per epoch. The weight α of the enhancement loss in multi-task learning, the modulation factor γ and batch size are set as 0.9, 2 and 16, respectively. To ensure fair comparison, the Squeezeformer is used as the backbone of ASR module and the training epoch is set as 150 for all models in all experiments. **Device.** All experiments are performed on 4 Quadro RTX 6000 GPUs.

4.2 Main Results

Table 1 displays the WER (%) performance of DT-net and other baseline models on the CHiME-4 dataset. Without a language model (LM), the WER for

WAVE-U-Net is 11.4/18.1/5.3, IFF-Net achieves a WER of 10.4/16.8/4.9, which is comparable to Emb-Enh(CNN-4) and slightly worse than Emb-Enh(U-Net-2) with a WER of 9.5/15.7/4.8. It is evident that the proposed method can further reduce the WER to 9.4/15.2/4.3 in the absence of an LM, with more pronounced improvements on et05_real and clean.

Table 2. Results presenting the % Word Error Rate (WER) of each control group in the superimposed CHiME-4 ASR task during the ablation experiments.

Method	WER		
	clean	dt05_real	et05_real
DT-net	4.3	9.4	15.2
WO/GDT	4.8	11.1	17.8
WO/FDT	5.4	10.5	17.1
WO/OT loss	4.7	10.4	16.7

4.3 Hyperparameter Experiment

In this section, we explore our model's sensitivity to balancing factor α and modulation factor γ on CHiME-4. As shown in Fig. 3, we give the WER results on different values of α: 0.1, 0.3, 0.5, 0.7, 0.9. It is clear that the optimal value resulted to be $\alpha = 0.9$. Therefore, we recommend using 0.9 as the default value of α. For parameter γ, we perform the DT-net with γ of values 1, 2, 3 and 4, where α is set as the default value 0.9. It is observed that DT-net achieved the best performance at $\gamma = 2$.

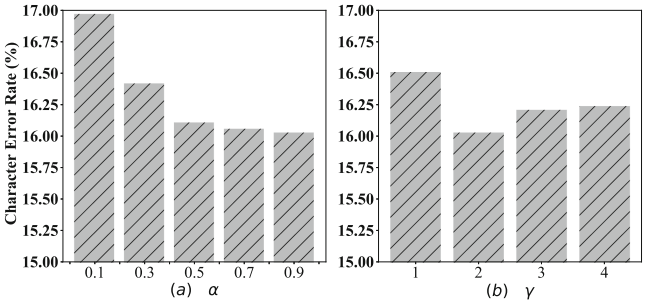

Fig. 3. Performance achieved with different values of hyperparameters α and γ. When experimenting with one hyperparameter, fix another hyperparameter to the default value ($\alpha = 0.9$, $\gamma = 2$).

4.4 Ablation Study

To demonstrate the effectiveness of each component in DT-net, we conducted ablation experiments on the noisy CHiME-4 dataset. Specifically, DT-net employs three components to capture robust speech embedding: (1) The Generalization Distribution Transformation (GDT) operator helps adapt to the clean and noisy pre-trained embedding distributions. (2) The Focus Distribution Transformation (FDT) operator aids in focusing on the more important embeddings. (3) The Optimal Transport (OT) loss maximizes the mutual information between inputs and outputs. Table 2 presents the performance after removing GDT, FDT, and OT loss from DT-net. Among all these components, GDT has the most significant impact on overall performance, emphasizing the importance of enriching data distributions in speech enhancement. The performance decrease with FDT is more pronounced on et05_real, indicating the role of learnable parameter distribution transformation in reducing the difference between test and training data distributions. Besides, the remaining OT loss also show the necessity of our expectations.

5 Conclusion

In this paper, we propose a joint SE and ASR system for robust speech recognition in noisy environments using an Distribution Transformation network (DT-net). Specifically, we enhance the features at the embedding level by using distribution transformation to exchange and reconstruct embedding distributions, to supplement the missing information for downstream ASR tasks. Experimental results on the Aishell1 corpus show that our proposed DT-net method achieves more effective noise-robust ASR compared to other competitive methods.

References

1. Kim, S., Gholami, A., Shaw, A., et al.: SqueezeFormer: an efficient transformer for automatic speech recognition. arXiv preprint arXiv:2206.00888 (2022)
2. Padmanabhan, J., Johnson Premkumar, M.J.: Machine learning in automatic speech recognition: a survey. IETE Tech. Rev. **32**(4), 240–251 (2015)
3. Ephraim, Y., Cohen, I.: Recent Advancements in Speech Enhancement. The Electrical Engineering Handbook, vol. 35 (2006)
4. Pascual, S., Bonafonte, A., Serra, J.: SEGAN: speech enhancement generative adversarial network. arXiv preprint arXiv:1703.09452 (2017)
5. Hou, N., Chenglin, X., Chng, E.S., Li, H.: Domain adversarial training for speech enhancement. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 667–672. IEEE (2019)
6. Fedorov, I., et al.: TinyLSTMs: efficient neural speech enhancement for hearing aids. arXiv preprint arXiv:2005.11138 (2020)
7. Zhang, Z., Geiger, J., Pohjalainen, J., et al.: Deep learning for environmentally robust speech recognition: an overview of recent developments. ACM Trans. Intell. Syst. Technol. (TIST) **9**(5), 1–28 (2018)

8. Hou, N., et al.: Speaker and phoneme-aware speech bandwidth extension with residual dual-path network. In: INTERSPEECH 2020, pp. 4064–4068 (2020)
9. Hou, N., Xu, C., Chng, E.S., Li, H.: Learning disentangled feature representations for speech enhancement via adversarial training. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 666–670. IEEE (2021)
10. Michelsanti, D., et al.: An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 1368–1396 (2021). https://doi.org/10.1109/TASLP.2021.3066303
11. Hou, N., Xu, C., Zhou, J.T., Chng, E.S., Li, H.: Multi-task learning for end-to-end noise-robust bandwidth extension. In: INTERSPEECH 2020, pp. 4069–4073 (2020)
12. Mporas, I., Ganchev, T., Kocsis, O., Fakotakis, N.: Speech enhancement for robust speech recognition in motorcycle environment. Int. J. Artif. Intell. Tools **19**, 159–173 (2010)
13. Loizou, P.C., Kim, G.: Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. IEEE Trans. Audio Speech Lang. Process. **19**(1), 47–56 (2011)
14. Wang, P., Tan, K., Wang, D.-L.: Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 39–48 (2020)
15. Ali, M.N., Brutti, A., Falavigna, D.: Direct enhancement of pre-trained speech embeddings for speech processing in noisy conditions. Comput. Speech Lang. **81**, 101501 (2023)
16. Baevski, A., Zhou, Y., Mohamed, A., et al.: wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. **33**, 12449–12460 (2020)
17. Hsu, W.N., Bolte, B., Tsai, Y.H.H., et al.: HuBERT: self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 3451–3460 (2021)
18. Wang, Y., Li, J., Wang, H., et al.: Wav2vec-switch: contrastive learning from original-noisy speech pairs for robust speech recognition. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7097–7101. IEEE (2022)
19. Cuturi, M.: Sinkhorn distances: lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems, vol. 26 (2013)
20. Graves, A.: Sequence transduction with recurrent neural networks. CoRR, vol. abs/1211.3711 (2012)
21. Zhang, Q., Lu, H., Sak, H., et al.: Transformer transducer: a streamable speech recognition model with transformer encoders and RNN-T loss. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7829–7833. IEEE (2020)
22. Babu, A., Wang, C., Tjandra, A., et al.: XLS-R: self-supervised cross-lingual speech representation learning at scale. arXiv preprint arXiv:2111.09296 (2021)
23. Mošner, L., Wu, M., Raju, A., et al.: Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6475–6479. IEEE (2019)
24. Li, B., Sainath, T.N., Narayanan, A., et al.: Acoustic modeling for Google home. In: Interspeech, pp. 399–403 (2017)

25. Pandey, A., Liu, C., Wang, Y., Saraf, Y.: Dual application of speech enhancement for automatic speech recognition. In: 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE (2021)
26. Martin, R.: Spectral subtraction based on minimum statistics. Power **6**(8), 1182–1185 (1994)
27. Stoller, D., Ewert, S., Dixon, S.: Wave-U-Net: a multi-scale neural network for end-to-end audio source separation. arXiv preprint arXiv:1806.03185 (2018)
28. Pandey, A., Wang, D.: Dense CNN with self-attention for time-domain speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 1270–1279 (2021)
29. Macartney, C., Weyde, T.: Improved speech enhancement with the Wave-U-Net. arXiv preprint arXiv:1811.11307 (2018)
30. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds.) Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015, Proceedings, Part III 18, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
31. Hu, Y., Hou, N., Chen, C., et al.: 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), pp. 6292–6296. IEEE (2022)
32. Ali, M.N., Falavigna, D., Brutti, A.: Enhancing embeddings for speech classification in noisy conditions. Proc. Interspeech **2022**, 2933–2937 (2022)