IFF-WAV2VEC: Noise Robust Low-Resource Speech Recognition Based on Self-supervised Learning and Interactive Feature Fusion

Jing Cao
School of Computer and Artificial Intelligence of Beijing Technology and Business University
2130062081@st.btbu.edu.cn

Zhaopeng Qian*
School of Computer and Artificial Intelligence of Beijing Technology and Business University
qianzhaopeng@btbu.edu.cn

Chongchong Yu
School of Computer and Artificial Intelligence of Beijing Technology and Business University
chongzhy@vip.sina.com

Tao Xie
School of Computer and Artificial Intelligence of Beijing Technology and Business University
xietao@btbu.edu.cn

ABSTRACT

In recent years, self-supervised learning representation (SSLR) has shown remarkable performance in low-resource speech recognition. However, it lacks consideration for the robustness of low-resource models in noisy environments, making it crucial to enhance their noise robustness. Speech enhancement is a commonly used denoising method, but it suffers from information over-suppression during training, leading to reduced accuracy in automatic speech recognition (ASR). To address this issue, this paper proposes an innovative Iff-wav2vec network architecture. Firstly, the network architecture integrates voice enhancement, SSLR, and ASR into one network. Secondly, this article uses interactive feature fusion methods to fuse noise features and enhanced features to compensate for the lack of information in the enhanced features. Finally, experimental results on Tujia and Shui languages show that the proposed method can effectively improve low resource ASR performance under various noise settings, resulting in stronger noise robustness.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Natural language processing**; • **Speech recognition**;

KEYWORDS

low-resource speech recognition, speech enhancement, self-supervised pre-training, interactive feature fusion

ACM Reference Format:
Jing Cao, Zhaopeng Qian, Chongchong Yu, and Tao Xie. 2023. IFF-WAV2VEC: Noise Robust Low-Resource Speech Recognition Based on Self-supervised Learning and Interactive Feature Fusion. In *2023 6th Artificial Intelligence and Cloud Computing Conference (AICCC) (AICCC 2023), December 16–18, 2023, Kyoto, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3639592.3639624

*Corresponding Author

AICCC 2023, December 16–18, 2023, Kyoto, Japan

ACM ISBN 979-8-4007-1622-5/23/12
https://doi.org/10.1145/3639592.3639624

1 INTRODUCTION

With the advancement of deep neural networks, automatic speech recognition (ASR) systems have achieved remarkable performance. Previous research has shown that training ASR systems using large-scale labeled audio data is highly effective. However, many languages around the world face language resource scarcity, with only a limited amount of speech data available, which can lead to overfitting when training models. Recently, researchers have acknowledged that self-supervised learning (SSL) can effectively tackle the problem of overfitting caused by insufficient data. SSL architectures can be broadly categorized into two types: generative learning [1–3] and contrastive learning [4–6]. Generative learning automatically learns the representation of data through an analysis of its internal structure and distribution. On the other handontrastive learning, on the other hand, extracts good feature representations for labelled data by identifying relationships between sample representations and their transformations. Although these SSL methods have demonstrated impressive results on various low-resource ASR tests, there has been limited research on low-resource ASR systems in complex environments, which is a crucial aspect for practical applications.

To improve the noise robustness of ASR models, speech enhancement techniques are commonly used to enhance the quality and intelligibility of speech signals. Traditional methods such as spectral subtraction [7] and Wiener filtering [8] often require specific assumptions and have limited effectiveness in non-stationary conditions. In recent years, deep learning-based speech enhancement algorithms [9–11] have gained popularity and are being extensively researched. One such example is the use of noise suppression techniques that employ Bidirectional Long Short-Term Memory (BiLSTM) in [12]. For low-resource speech recognition in complex environments, HLGAN [14] is proposed to input noisy and clean audio signals in parallel to fully utilize information from the clean audio. Another method, SpeechStew [15], trains the model with mixed data and fine-tunes it with low-resource noisy data to reduce overfitting problems and quickly learn significant features from noisy data. Finally, a transformer-based SE model [16] is proposed and fine-tuned through a two-stage training scheme.

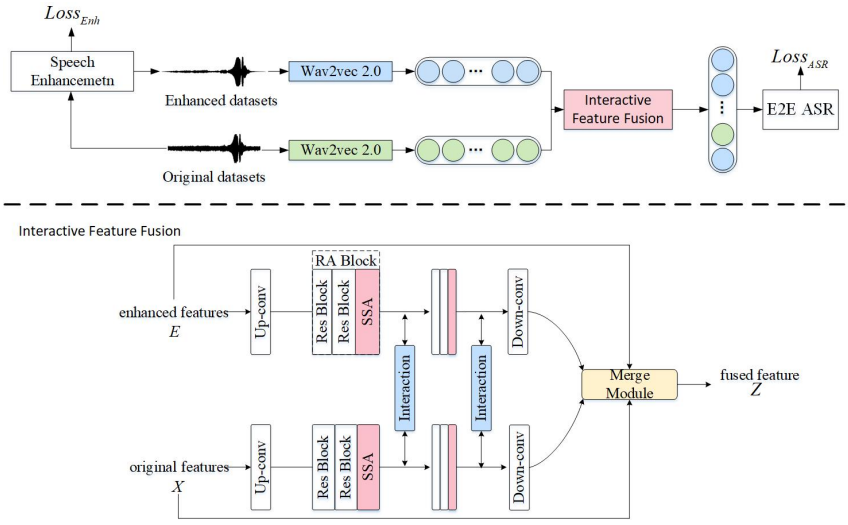

Figure 1: System architecture of the proposed Iff-wav2vec framework.

It has been observed that speech enhancement does not always lead to good performance for ASR systems [17, 18] due to the over-suppression of speech signals. Although scholars have proposed cascaded frameworks [19] to optimize SE modules and ASR modules with ASR objectives, this strategy increases the complexity of the ASR system. To enhance the noise robustness of self-supervised models, various methods have been proposed. For instance, SPIRAL [20] utilizes a teacher-student framework to learn denoised representations from noisy data. Wav2vec-switch [21] learns context representations with noise robustness by inputting pairs of original and noisy speech into a network and performing contrastive learning tasks. In [22], a reconstruction module is combined with the contrastive learning framework of wav2vec 2.0, and multi-task continuous pretraining is performed on noisy data to improve the noise robustness of learning speech representations during the pretraining stage.

This paper aims to enhance the performance of low-resource ASR systems in complex environments by proposing an ASR model called Iff-wav2vec. It integrates SE modules, self-supervised learning representation modules (SSLR), interactive feature fusion modules (IFF-NET), and ASR modules into a single end-to-end model. The contributions of this paper are as follows: Firstly, it proposes an interactive feature fusion framework to address the problem of over-suppression of speech signals during speech enhancement. Secondly, it integrates SE, SSLR, and end-to-end ASR into a single neural network based on joint optimization. Finally, it compares the effectiveness of the model under different noise environments to evaluate its robustness.

2 METHOD

This section describes the proposed Iff-wav2vec model architecture, as shown in Figure 1. The model consists of four components: speech enhancement, wav2vec 2.0, interactive feature fusion, and speech recognition. The speech enhancement module is designed to extract clean speech from noisy and polluted audio while maintaining both quality and intelligibility. The wav2vec 2.0 module serves as the feature extraction module for the overall architecture, aiming to obtain more robust representations. The interactive feature fusion module is intended to complement the missing information in the enhanced speech.

2.1 Feature Extraction by Wav2vec 2.0

In this paper, wav2vec 2.0 is employed as the feature extraction module in the overall architecture. The wav2vec 2.0 system consists of three primary components: convolutional feature encoder, context network, and quantization block. wav2Vec 2.0 processes the original audio signal X through a CNN to obtain a latent speech representation Z, which is then fed into a Transformers model after applying a random mask to acquire contextual feature representations for subsequent tasks. Simultaneously, the model transforms the latent speech representation Z into a discrete vector using the Gumble softmax operation in the quantization module. Finally, contrastive loss is computed between the contextual feature representations and quantization embeddings, enabling the context network to identify accurate quantization representations even in the presence of interference.

The loss during the training process of Wav2vec 2.0 is composed of two components: contrastive loss L_m and diversity loss L_d. This loss is defined as follows:

$$L = L_m + \alpha L_d \tag{1}$$

Where α is a hyperparameter that controls the diversity loss.

During the contrastive learning process, the model needs to select the correct quantized latent representation $\tilde{q} \in Q$ from a set of $K+1$ candidate quantized representations, where the false quantized representations $\tilde{q}/q_t$ are obtained by uniformly sampling from the same time step. The contrastive loss is defined as:

$$L_m = -\log \frac{\exp\left(\frac{sim(c_t, q_t)}{k}\right)}{\sum_{\tilde{q} \sim Q_t} \exp\left(\frac{sim(c_t, q_t)}{k}\right)} \tag{2}$$

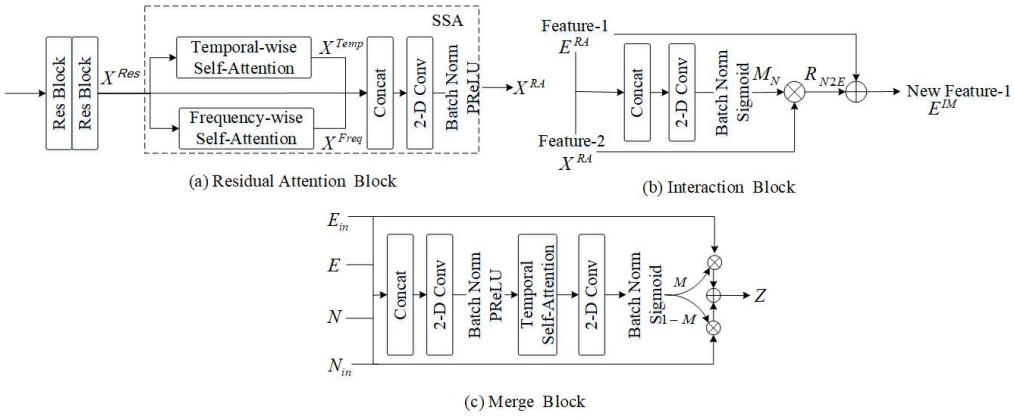

Figure 2: IFF-NET Network Architecture Diagram

Where $sim(a, b)$ represents the cosine similarity between the contextual representation and the quantized latent representation, and k represents the number of interfering items.

To effectively supervise the clustering process during quantiz tion, the model utilizes the diversity loss. The goal of the diversity loss is to maximize the entropy of the average $softmax$ probability of the fully occupied entries in each codeword group. This loss is defined as:

$$L_d = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \bar{p}_{g,v} log \bar{p}_{g,v} \tag{3}$$

Where G represents the number of entries and V represents the number of cluster centers. $p_{g,v}$ is the calculation formula for Gumbel softmax.

2.2 Interactive Feature Fusion Network

The Interactive Feature Fusion Network (IFF-NET) is employed to address the problem of information loss in speech enhancement. The network comprises upsampling convolutional blocks, residual attention (RA) blocks, interactive modules, downsampling modules, and merging modules.

Upsampling convolution and downsampling convolution: The Up-conv block and Down-conv block consist of 2D convolutional layers, layer normalization, and ReLU activation functions. We input the enhanced features and original features into the Up-conv block for feature extraction. At the end of the model, the Down-conv block is used to ensure that the channel dimensions of the interaction features E_{in} and X_{in} match the channel dimensions of the original inputs E and X.

Residual Attention (RA) block: The role of the residual attention module is to capture both local and global dependencies in the features. This module is composed of residual blocks, temporal self-attention blocks, frequency self-attention blocks, and convolutional layers, as shown in Figure 2 (a). Each residual block contains 2D convolutions to extract deep local features X^{Res}. The features X^{Res} obtained from the residual blocks are then fed into the temporal attention module and frequency attention module separately to obtain global dependencies in both time and frequency. Since the temporal attention and frequency attention mechanisms are similar, we will only present the formula for frequency attention, which is as follows:

$$X_f^i = Reshape^f \left(X^{Res} \right), \quad i \in \{q, k, v\} \tag{4}$$

$$SA^f = \frac{Softmax\left(X_f^q * \left(X_f^k \right)^T \right)}{\sqrt{C \times F}} * X_f^v \tag{5}$$

$$X^{Freq} = X^{Res} + Reshape^{f\_inv} \left(SA^f \right) \tag{6}$$

Where C represents the filter index, T represents the frame index, and F represents the frequency index. $Reshape^f$ refers to reshaping the tensor from $R^{C \times T \times F}$ to $R^{F \times (C \times T)}$ along the F dimension, and $Reshape^{f\_inv}$ represents the inverse operation.

Finally, the generated deep features X^{Temp} and X^{Freq} are concatenated with X^{Res}, and then fed into a 2D convolutional layer to obtain the output X^{RA}.

Interaction Module: The introduction of the interaction module is intended to learn complementary information from enhanced features and original features, as illustrated in Figure 2 (b). This module consists of two directions: enhanced to noise (e2n) and noise to enhanced (n2e). The computation process in these two directions is similar, with the only difference being the exchange of "Feature-1" and "Feature-2" as shown in Figure 2 (b).

The process depicted in Figure 2 (b) is the n2e flow.To begain with, the enhanced (E^{RA}) and original audio features (X^{RA}) are concatenated and fed into a 2D convolutional layer. Then, a ge nerated mask M_N is used to determine whether the information in X^{RA} is to be removed or preserved. Next, the residual features are obtained by synthesizing X^{RA} and M_N. Finally, R_{N2E} and E^{RA} are concatenated to obtain an enhanced version of the enhanced features E^{IM}.

Merge Module: The merge module is utilized to further integrate the interaction features of the enhanced and original branches, as illustrated in Figure 2 (c). To begin, the initial inputs X and E, along with the interaction features, are concatenated and fed into the merge module. Following this, a 2D convolutional layer and temporal attention module are applied to obtain a mask M, which controls the retention of the interaction features. The final fused feature Z is represented as follows:

$$Z = E_{in} * M + X_{in} * (1 - M) \tag{7}$$

Table 1: Dataset Statistics

Language	Audio data	Labeled data
Tujiayu	7h8m59s	7h8m59s
Shuiyu	8h40m23s	7h2m29s

2.3 E2E ASR by Joint CTC/Attention

This paper presents a speech recognition model that combines CTC and attention-based encoder-decoder [27]. The model architecture includes Conformer encoder, CTC, and Transformer decoder, as shown in the figure below:

$$Q = Conformer\,(Z) \tag{8}$$

$$C_{CTC} = CTC\,(Q) \tag{9}$$

$$C_{Att} = TransformerDec\,(Q) \tag{10}$$

Where C_{CTC} and C_{Att} denote the estimates derived from the CTC and Transformer decoder, respectively. During decoding, a combination of posteriors from both decoders is employed in conjunction with beam search.

The ASR model is optimized based on the sum of the following two objective functions:

$$L_{ASR} = \alpha Loss_{CTC} + (1-\alpha)\,Loss_{Att} \tag{11}$$

Where α is a hyperparameter, and $Loss_{CTC}$ and $Loss_{Att}$ are the posterior distributions from CTC and decoder, respectively. The CTC objective function enforces alignment between features and transcription during training, mitigating mislocalization in attention-based encoder-decoder.

3 EXPERIMENTS

3.1 Dataset

For low-resource speech recognition of clean corpora, we used Tujia and Shui languages. Both of the two low-resource languages used in this paper belong to the Sino-Tibetan language family, and both suffer from limited audio data due to the relatively large age of most native speakers and small number of users. The detailed information about the dataset is shown in Table 1. Among them, the Tujia language includes 300 core spoken vocabulary words, 2000 major spoken vocabulary words, and 27 spoken phrases, with a total duration of 7 hours, 8 minutes, and 59 seconds. The Shui language includes 2474 sentences, 7514 vocabulary items, and 1171 exemplary characters, with a total duration of 8 hours, 40 minutes, and 23 seconds.

For noisy speech recognition, we selected the MUSAN dataset [23] to synthesize noisy audio by mixing it with clean audio. The MUSAN dataset contains three categories of noise: 1) Music data, including various types such as jazz and rap; 2) Noise data, including sounds like car horns and thunder; 3) Speech data, including recordings of hearings and debates. Noisy audio is obtained by mixing clean speech data with any noise from the MUSAN dataset.

3.2 Experimental Setup

All experiments were implemented using the fairseq and ESPnet toolkits. The proposed IFF-wav2vec model consists of four modules: SE module, IFF-NET module, wav2vec 2.0, and decoder. The SE module consists of 3 layers of bidirectional long short-term memory (BLSTM), a linear layer, and ReLU activation function to predict noise magnitude feature masks. The IFF-NET module contains 4 RA blocks and 64 filters. The Transformer used is based on the configuration of BASE in fairseq, and the decoder consists of 6 transformer layers.

During the pre-training phase, we trained the wav2vec 2.0 models separately for the Tujia and Shui languages using the fairseq toolkit. We used the Adam optimizer with a learning rate of . The diversity loss function was set to 0.1 during the computation of the loss function. In the fine-tuning phase, our model was implemented using the espnet toolkit. In this phase, we mixed noise into the clean data to obtain the noisy audio. The same Adam optimizer was used with a learning rate of 5. Additionally, for multitask learning, we set the weight of the enhancement loss to 0.3.

3.3 Experimental Results and Analysis

To better evaluate the effectiveness of the proposed method in this paper, we selected three models from pre-vious works as our baseline models for comparison. The SE module and ASR module in our model use the same architecture. The baseline models chosen in this paper are as follows: 1) an end to end speech recognition system that combines the SE module and the conformer module; 2) the IFF-NET model that integrates interactive fusion of enhanced speech and noisy speech; 3) the wav2vec 2.0 model based on contrastive learning.

Table 2 presents the ASR performance under noisy environments, which is evaluated by mixing clean test sets with audio segments of various types of noise at different signal to noise ratios (SNRs) and then assessing the model.

From Table 2, it can be observed that compared to the cascade SE and ASR method, IFF-NET compensates for missing information during the speech enhancement process by introducing interactive feature fusion, resulting in a relative reduction of 3.49 and 10.45 in CER for Tujia and Shui languages, respectively. The contrastive self-supervised pretraining model, wav2vec 2.0, without using a speech enhancement module, achieved a relative reduction of 4.77 and 13.73 in CER for the two languages, respectively. The experimental results indicate that wav2vec 2.0, trained on a large amount of unlabeled data through self-supervised learning, can obtain more robust speech representations for improved ASR performance. It is observed that the proposed IFF-wav2vec achieves the best results, with a respective CER improvement of 2.11 and 0.35 compared to the best performing wav2vec 2.0.

To understand the impact of different types of noise on ASR performance, we conducted experiments using three types of noise from the MUSAN dataset. During the experiments, the noise was mixed with clean audio at the same signal to noise ratio (0-10dB).

Table 2: Performance Comparison of All Models at Different SNRs

	Noisy(0dB)		Noisy(10dB)		Noisy(20dB)		Original	
	Tujiayu	Shuiyu	Tujiayu	Shuiyu	Tujiayu	Shuiyu	Tujiayu	Shuiyu
SE+Conformer[24]	51.6	45.3	44.8	38.9	37.8	33.9	33.3	21.2
IFF-NET[22]	45.4	32.7	38.3	27.6	33.8	23.9	26.6	16.3
Wav2vec 2.0[14]	36.5	19.1	35.5	15.3	26.2	10.9	18.5	9.3
IFF-wav2vec(ours)	33.1	16.2	26.1	14.7	22.6	10.7	18.1	8.6

Table 3: Performance Comparison of Various Models in Different Noise Environments

	Music		Noisy		Speech		Original	
	Tujiayu	Shuiyu	Tujiayu	Shuiyu	Tujiayu	Shuiyu	Tujiayu	Shuiyu
SE+Conformer[24]	47.6	40.3	40.8	31.9	63.4	59.1	33.3	21.2
IFF-NET[22]	41.4	28.7	33.9	25.6	53.8	51.2	26.6	16.3
Wav2vec 2.0[14]	33.5	13.1	26.4	13.6	41.2	30.7	18.5	9.3
IFF-wav2vec(ours)	27.1	12.2	23.1	10.4	36.5	18.3	18.1	8.6

Table 4: CER% Results for Different Data Mixtures

	Different Types of Data Fusion	IFF-NET	IFF-wav2vec(Ours)
Tujiayu	SE+Noisy	31.8	25.6
	SE+Clean	26.6	17.5
Shuiyu	SE+Noisy	23.9	12.5
	SE+Clean	10.7	9.2

From Table 3, it can be observed that IFF-wav2vec effectively improves ASR performance under different types of noise. However, the CER is highest when the noise type is "speech" with values of 36.5 and 18.3 for Tujia and Shui languages, respectively. The reason for this is that the presence of speech in the noise confuses the model, making it difficult to distinguish the true speaker's voice, leading to a decrease in recognition accuracy.

From Tables 2 and 3, it can be observed that in the case of the same total duration, the recognition accuracy of Shui language is higher than that of Tujia language. The reason may be that the duration of each audio in Tujia language is longer than that of Shui language, resulting in more noise features in the fused features of Tujia language, which leads to a decrease in recognition accuracy. Therefore, in this study, the original audio in the fusion process was replaced with clean audio for experimentation. The test set consisted of various noise types mixed with clean audio at SNRs ranging from 0 to 20 dB. The experimental results are shown in Table 4.

From Table 4, it can be seen that when replacing the noise audio features with clean features in the fused features, IFF-NET and the proposed method in this paper showed improvements in CER for Tujia language by 5.2 and 8.1, respectively, and for Shui language by 12.2 and 3.3, respectively. This indicates that although the original audio in the feature fusion process helps to supplement some missing information during enhancement, it also reintroduces noise into the features, ultimately leading to a decrease in recognition accuracy.

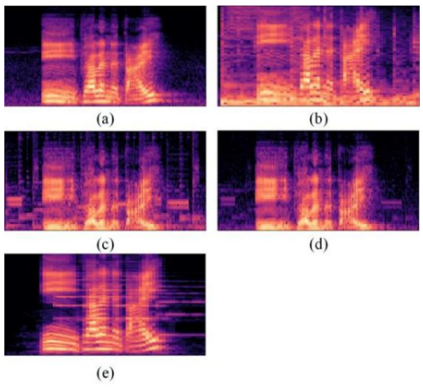

Figure 3: Spectrums of (a) clean , (b) noisy, and ASR input of (c) Cascaded SE and ASR System, (d)IFF-NET,(e)IFF-wav2vec.

To further demonstrate the contribution of the proposed IFF-wav2vec method in handling noise, we present the mel spectrograms of ASR inputs for different methods, as shown in Figure 3 From (a) and (b), it can be observed that there is a significant amount of noise in the features. Then, comparing with methods (c-d), it is observed that the proposed IFF-wav2vec method can effectively reduce more background noise while preserving richer clean information.

4 CONCLUSIONS

In this paper, a new framework called IFF-wav2vec is proposed to enhance the performance of low-resource ASR systems in complex environments. The model's performance is analyzed by simulating different noise environments. The results demonstrate that the proposed mod-el architecture can effectively improve the recognition accuracy of low-resource languages in complex environments. The proposed method achieves a CER improvement of 8.6 on low-resource settings.

REFERENCES

[1] Chung, Yu-An, Wei-Ning Hsu, Hao Tang, and James Glass. "An unsupervised autoregressive model for speech representation learning." arXiv preprint arXiv: 1904.03240, 2019.

[2] Liu, Andy T., Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6419-6423. IEEE, 2020.

[3] Liu, Andy T., Shang-Wen Li, and Hung-yi Lee. "Tera: Self-supervised learning of transformer encoder representation for speech." IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, pp.2351-2366, 2021.

[4] Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. "wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in neural information processing systems, 33, pp.12449-12460, 2020.

[5] Hsu, Wei-Ning, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." IEEE/ACM Transactions on Audio, Speech, and Language Processing , 29, pp.3451-3460, 2020.

[6] Chen, Sanyuan, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li *et al.* "Wavlm: Large-scale self-supervised pre-training for full stack speech processing." IEEE Journal of Selected Topics in Signal Processing , 16(6), pp.1505-1518, 2022.

[7] Boll, Steven. "Suppression of acoustic noise in speech using spectral subtraction". *IEEE Transactions on acoustics, speech, and signal processing*, 27(2), pp.113-120, 1979.

[8] Loizou, Philipos C. *Speech enhancement: theory and practice*. CRC press, 2013.

[9] Cheng, Jiaming, Ruiyu Liang, and Li Zhao. "DNN-based speech enhancement with self-attention on feature dimension." *Multimedia Tools and Applications* 79, 32449-32470, 2020.

[10] Juan Manuel Martin-Doñas, , Angel Manuel Gomez, Jose A. Gonzalez, and Antonio M. Peinado. "A deep learning loss function based on the perceptual evaluation of the speech quality." *IEEE Signal processing letters*, 25(11), 1680-1684, 2018.

[11] Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin. "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. " In *International Conference on Machine Learning*, pp. 2031-2041, 2019.

[12] Li, Xiaoqi, Yaxing Li, Yuanjie Dong, Shan Xu, Zhihui Zhang, Dan Wang, and Shengwu Xiong. "Bidirectional LSTM Network with Ordered Neurons for Speech Enhancement." In *Interspeech*, pp. 2702-2706, 2020.

[13] Tan, Ke, Jitong Chen, and DeLiang Wang. "Gated residual networks with dilated convolutions for monaural speech enhancement." *IEEE/ACM transactions on audio, speech, and language processing*, 27(1), pp.189-198, 2018.

[14] Yang, Fan, Ziteng Wang, Junfeng Li, Risheng Xia, and Yonghong Yan. "Improving generative adversarial networks for speech enhancement through regularization of latent representations." Speech Communication, 118, pp.1-9, 2020.

[15] Chan, William, Daniel Park, Chris Lee, Yu Zhang, Quoc Le, and Mohammad Norouzi. "Speechstew: Simply mix all available speech recognition data to train one large neural network." arXiv preprint arXiv:2104.02133, 2021.

[16] Noor, Md Mahbub E., Yen-Ju Lu, Syu-Siang Wang, Supratip Ghose, Chia-Yu Chang, Ryandhimas E. Zezario, Shafique Ahmed, Wei-Ho Chung, Yu Tsao, and Hsin-Min Wang. "Investigation of a single-channel frequency-domain speech enhancement network to improve end-to-end Bengali automatic speech recognition under unseen noisy conditions." In 2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), pp. 7-12. IEEE, 2021.

[17] Loizou, Philipos C., and Gibak Kim. "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions." *IEEE transactions on audio, speech, and language processing* 19.1 (2010): 47-56.

[18] Kazuma Iwamoto , Tsubasa Ochiai, Marc Delcroix , Rintaro Ikeshita , Hiroshi Sato , Shoko Araki and Shigeru Katagiri. "How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR, " in *Interspeech*, pp. 5418–5422, 2022.

[19] Aswin Shanmugam Subramanian , Xiaofei Wang , Murali Karthick Baskar , Shinji Watanabe , Toru Taniguchi, Dung Tran and Yuya Fujita. "Speech enhancement using end-to-end speech recognition objectives. "*IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 234-238,2019.

[20] Huang, Wenyong, Zhenhe Zhang, Yu Ting Yeung, Xin Jiang, and Qun Liu. "SPIRAL: Self-supervised perturbation-invariant representation learning for speech pre-training." *arXiv preprint arXiv:2201.10207*, 2022.

[21] Wang, Yiming, Jinyu Li, Heming Wang, Yao Qian, Chengyi Wang, and Yu Wu. "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.

[22] Wang, Heming, Yao Qian, Xiaofei Wang, Yiming Wang, Chengyi Wang, Shujie Liu, Takuya Yoshioka, Jinyu Li, and DeLiang Wang. "Improving noise robustness of contrastive speech representation learning with speech reconstruction." In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6062-6066. IEEE, 2022.

[23] Chang, Xuankai, Takashi Maekaku, Yuya Fujita, and Shinji Watanabe. "End-to-end integration of speech recognition, speech enhancement, and self-supervised learning representation." *arXiv preprint arXiv:2204.00540*, 2022.

[24] Ma, Duo, Nana Hou, Haihua Xu, and Eng Siong Chng. "Multitask-based joint learning approach to robust asr for radio communication speech." In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 497-502. IEEE, 2021.

[25] Hu, Yuchen, Nana Hou, Chen Chen, and Eng Siong 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)*, pp. 6292-6296. IEEE, 2022.

[26] Snyder, David, Guoguo Chen, and Daniel Povey. "Musan: A music, speech, and noise corpus." *arXiv preprint arXiv:1510.08484*, 2015.

[27] Gulati, Anmol, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han *et al.* "Conformer: Convolution-augmented transformer for speech recognition." arXiv preprint arXiv:2005.08100, 2020.