A Repetitive Spectrum Learning Framework for Monaural Speech Enhancement in Extremely Low SNR Environments (Student Abstract)

Wenxin Tai

University of Electronic Science and Technology of China
Chengdu, CN 610054
wxtai@std.uestc.edu.cn

Abstract

Monaural speech enhancement (SE) at an extremely low signal-to-noise ratio (SNR) condition is a challenging problem and rarely investigated in previous studies. Most SE methods experience failures in this situation due to three major factors: overwhelmed vocals, expanded SNR range, and short-sighted feature processing modules. In this paper, we present a novel and general training paradigm dubbed repetitive learning (RL). Unlike curriculum learning that focuses on learning multiple different tasks sequentially, RL is more inclined to learn the same content repeatedly where the knowledge acquired in previous stages can be used to facilitate calibrating feature representations. We further propose an RL-based end-to-end SE method named SERL. Experimental results on TIMIT dataset validate the superior performance of our method.

Introduction

Reducing background noise and improving the quality and intelligibility of degraded speech has been a long-standing topic in speech processing applications (Loizou 2013). Recently, significant progress on this research topic has been made with the involvement of deep learning paradigms (Tan and Wang 2018). However, according to our knowledge, monaural speech enhancement (SE) at extremely low SNR conditions is rarely investigated in previous works.

Presently, three challenges restrict the performance of existing SE algorithms. First, for challenging acoustic scenarios as low SNR conditions, current SE systems usually suffer from performance bottlenecks in recovering clean speech from mixtures (Li et al. 2021). Second, noise intensities in real-world scenes change dynamically, which requires SE systems to accommodate wide expansion of the SNR range and raises the difficulty of the network design (Hao et al. 2020). The third problem is caused by the limited kernel size of convolution layers, which often results in a short-sighted feature extractor. It might severely decrease the performance because the correlations of harmonics in the spectrogram are mostly non-local, that is, the value at a base frequency is strongly correlated with the values at its surrounding overtones.

Methodology

SERL

The basic idea behind SERL is to take the advantage of multi-stage learning to obtain deep-seated information, and use these advanced features to provide guidance for shallow information processing. SERL consists of two stages, and the network architecture of each stage is identical. The diagram of our proposed system is presented in Fig.1.

The input to the network is the magnitude spectrum after the STFT operation, denoted by X_m. Here $X_m \in \mathcal{R}^{T \times F \times 1}$ is a real-valued spectrogram, where T represents the number of time steps and F represents the number of frequency bands.

Feature Extractor In a regular convolutional layer, a fixed convolutional kernel is used for all speech spectra, making it difficult for the model to cope with diverse environments. In view of this, we introduce dynamic convolution (Yang et al. 2019) to meet the requirement as well as ease the computational burden. Since dynamic convolution kernels are different in the same mini-batch, we need to fuse kernels before composing a mini-batch. Compared to mix multi-branch results at the feature map level, dynamic convolution is more efficient because the convolution is computed only once per sample.

Bottleneck Layer Considering the importance of sequence modeling, we cascade 18 squeezed temporal convolutional modules (S-TCMs) (Li et al. 2021) as bottleneck layer. Compared with LSTMs, S-TCMs can obtain better performance in temporal sequence processing. Fig.1 (d) presents the architecture of S-TCM.

Spectrum Reconstructor We adopt the GLU formats and use 2-D deconvolution operation to construct Deconv-GLU, where the compressed features can be gradually interpolated and restored to the original size.

Experiments

Dataset

Our experiments are conducted on the TIMIT corpus (Garofolo et al. 1993). Clean speeches are mixed with a large number of non-stationary noises from two public noise datasets. We generate approximately 30 hours of data for training, 1.5 hours for validation, and 1 hour for testing.

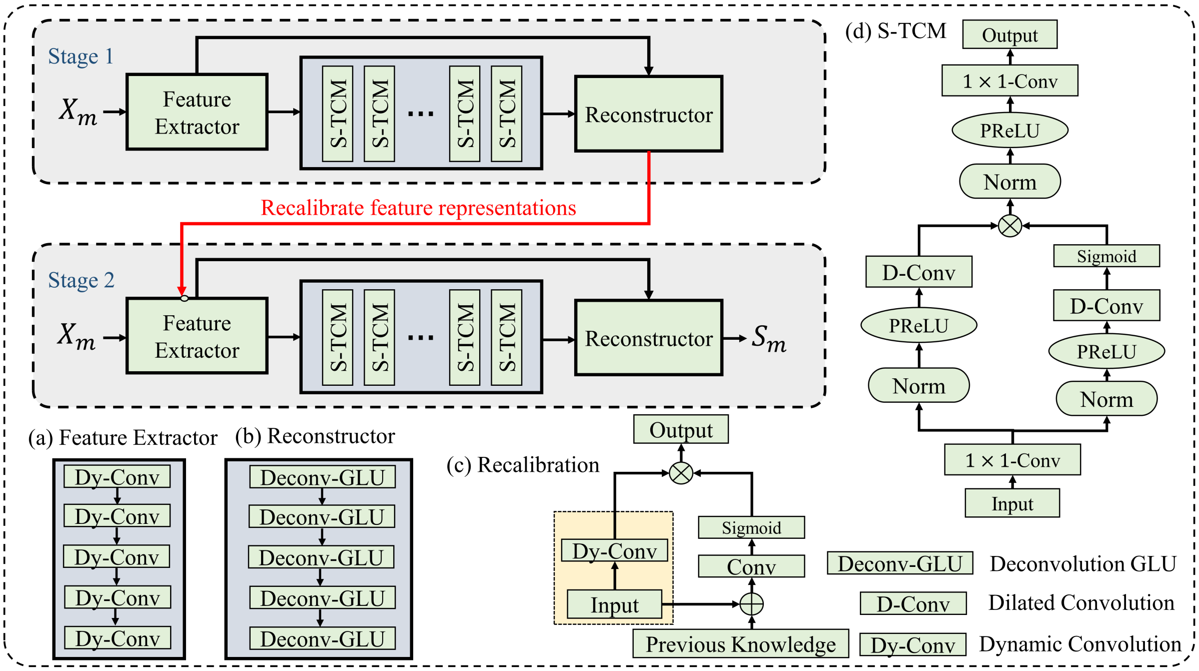

Figure 1: Workflow of SERL. The channel size of each layer is fixed to 64 except for S-TCMs (details in (Li et al. 2021)).

Performance Comparison

Method	SEEN		UNSEEN	
	STOI(%)	PESQ	STOI(%)	PESQ
Noisy	59.80	1.30	52.19	1.24
AECNN	77.86	2.07	59.23	1.53
PHASEN	78.03	1.99	54.40	1.44
CTS-Net	77.10	2.01	56.58	1.53
w/o RL	77.84	2.18	58.20	1.63
SERL	**81.11**	**2.33**	**61.55**	**1.71**

Table 1: Comparison of different methods.

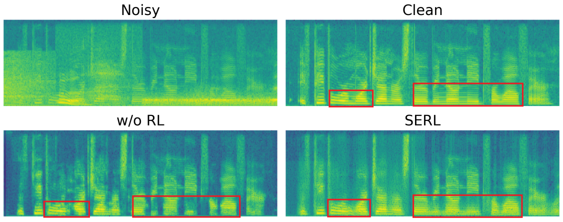

Figure 2: Spectrum visualization under -15 dB.

We report the results of different methods in Tables 1. For both seen and unseen noise cases, SERL achieves the best results compared to recent state-of-the-art algorithms.

As shown in Fig.2, the basic skeleton without RL is comparable with SERL in local information processing, but becomes worse in terms of long-term spectrum recovery (red box in the figure). This finding supports our design of using RL to provide global macro guidance.

References

Garofolo, J. S.; Lamel, L. F.; Fisher, W. M.; Fiscus, J. G.; and Pallett, D. S. 1993. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n*, 93: 27403.

Hao, X.; Su, X.; Wang, Z.; Zhang, Q.; Xu, H.; and Gao, G. 2020. SNR-based teachers-student technique for speech enhancement. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Li, A.; Liu, W.; Zheng, C.; Fan, C.; and Li, X. 2021. Two Heads are Better Than One: A Two-Stage Complex Spectral Mapping Approach for Monaural Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 1829–1843.

Loizou, P. C. 2013. *Speech Enhancement: Theory and Practice*. CRC Press.

Tan, K.; and Wang, D. 2018. A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement. In *Interspeech*, 3229–3233.

Yang, B.; Bender, G.; Le, Q. V.; and Ngiam, J. 2019. CondConv: conditionally parameterized convolutions for efficient inference. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 1307–1318.