Speech Enhancement Based on Lightweight Neural Networks

Name: Wenye Zhang
SID:540473182
Github Username: imsb226
GitHub Project Link: https://github.com/imsb226/elec5305_project_540473182

Project Overview

This work aims for single-channel speech enhancement deployable on laptop CPU under hard budget constraint (parameters <1M) under delivering reasonable perceptual quality (PESQ(wb) ≥2.95, STOI ≥0.94, ΔSI-SDR ≥6 dB). We extend DPCRN (encoder → dual-path time/frequency RNN → decoder with skip connections), adding a lightweight Transformer encoder to learn global relations and predicting a Complex Ratio Mask (CRM); training adopts a multi-objective loss function (waveform MSE + complex-spectrum MSE), efficiency is made by depthwise/group convolutions together with structured/channel pruning to allow CPU-friendly inference. Tests are conducted under unified protocol on VoiceBank-DEMAND (train: 28 speakers, ~9.4 h, SNR 0/5/10/15 dB; test: 2 unseen speakers, ~0.6 h, SNR 2.5/7.5/12.5/17.5 dB; resampled to 16 kHz), reporting mean ± std results together with CPU RTF/latency. Deliverables include source code/doc, trained weights, reproducible CPU demo script, and concise technical report with ablations on attention, loss design, pruning.

Background and motivation

Speech enhancement aims to squeeze intelligible talk out of noisy blends for far-field calls, hearables, online meetings, and embedded IoT. Executing models on laptops locally on the CPU decreases end-to-end delays compared to cloud inference over bandwidth-constrained links. Efficient methods take two paths. Time-domain models process waveforms directly, with short reconstruction paths and high-temporal continuity, but usually require wider channels and wider receptive fields; slim designs tend to use an all-U-Net backbone with narrow BiLSTM bottleneck to capture cross-frame context (Défossez et al., 2020). Time-frequency methods take input STFs with Narrow U-Net/CRN derivatives, allowing consistencies by complex-domain estimate or mask learning. Two-path CRNs such as DPCRN integrate convolutional subsampling with time–frequency recurrent modeling for strong small-model performance (Le, Chen, Chen, & Lu, 2021). Following work reveals that gating,

temporal refinement, attention, and multi-loss training enhance real-time quality and efficiency.

Based on DPCRN, this project injects an lightweight Transformer into the encoder–dual-path–decoder framework and generates a complex ratio mask (CRM). Using multi-objective losses and compression (compression/pruning/channel pruning), this project aims for <1 M parameters and CPU-friendly inference with perceptual quality (e.g., PESQ/STOI) and robustness preserved.

Proposed Methodology

This project will use the Dual-Path Convolutional Recurrent Network (DPCRN) as the basic framework (Le, Chen, Chen, & Lu, 2021), which has been proven effective in speech enhancement tasks (Lee & Kang, 2023; Rong et al., 2024; Wan et al., 2023). The basic framework of this project is shown in Figure 1. This framework puts the spectrogram of the speech signal after Short-Time Fourier Transform (STFT) into a U-net style encoder-decoder, uses deep convolutional neural networks to extract local features of the signal while downsampling, then sequentially uses recurrent neural networks on frequency and time dimensions to obtain their short-term and long-term dependencies, and uses fully connected layers to map high-dimensional channels back to the original dimensions, finally uses transposed convolutional neural networks (ConvTranspose) to decode the features back to the original size, and constructs the complex ratio mask of the spectrogram with the structure of skip connections.

This project will add an attention mechanism on the basis of DPRNN and use a transformer encoder to enhance the model's ability to capture global information.

The loss function of this project is as follows:

$$\mathcal{L} = \lambda_{\text{wave}}\,\mathcal{L}_{\text{wave}} + \lambda_{\text{spec}}\,\mathcal{L}_{\text{spec}}.$$

where $\mathcal{L}_{\text{wave}}$ is the mean square error of the waveform, and $\mathcal{L}_{\text{spec}}$ is the mean square error of the complex spectrum.

Finally, while minimizing the impact on the results, this project will reduce the model's parameter count through pruning operations.

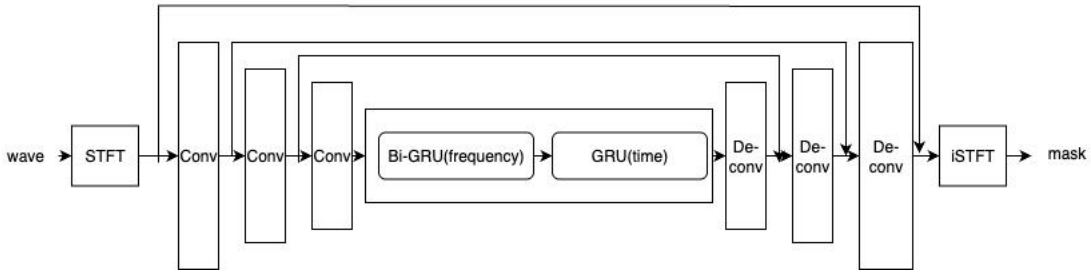

Figure 1: The basic framework

The experiment will adopt the most popular VOICEBANK + DEMAND dataset (Valentini-Botinhao, 2017) for experiment. The training set involves 28 speakers, whose total speech time is around 9.4 hours, the sampling rate is set to 48kHz. The noise is injected into the speech at signal-to-noise ratios of 0, 5, 10, and 15dB. The

test set is comprised by two invisible speakers, whose total speech time is around 0.6 hours, signal-to-noise ratios are set to 2.5, 7.5, 12.5, 17.5, respectively.

Expected Outcomes

On VoiceBank-DEMAND (16 kHz), a lightweight model with < 1M parameters was constructed, achieving PESQ(wb) ≥ 2.95, STOI ≥ 0.94, and an improvement of ΔSI-SDR ≥ 6.0 dB relative to the original noisy speech. The evaluation was performed using a unified setting, and the results are given as mean ± standard deviation.

Expected Metric	Value
PESQ	≥ 2.95
STOI	≥ 0.94,
ΔSI-SDR(dB)	≥ 6
Parameters(M)	< 1

Timeline

Week	Task
9	Building the basic framework and training scripts
10-12	Implement the experiment
13	Complete the report

Renference

Défossez, A., Synnaeve, G., & Adi, Y. (2020). Real time speech enhancement in the waveform domain. Proceedings of Interspeech 2020, 3291–3295. ISCA. https://doi.org/10.21437/Interspeech.2020-2409

X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-Path Convolution Recurrent Network for single-channel speech enhancement," in *Proc. Interspeech*, 2021. https://arxiv.org/abs/2107.05429

Lee, J., & Kang, H.-G. (2023). Real-time neural speech enhancement based on temporal refinement network and channel-wise gating methods. Digital Signal Processing, 133, 103879. https://doi.org/10.1016/j.dsp.2022.103879

Wan, L., Liu, H., Zhou, Y., & Jia, J. (2023). Multi-loss convolutional network with time-frequency attention for speech enhancement. arXiv preprint arXiv:2306.08956.

Rong, X., Sun, T., Zhang, X., Hu, Y., Zhu, C., & Lu, J. (2024). GTCRN: A speech enhancement model requiring ultralow computational resources. ICASSP 2024 (pp. 971–975). https://doi.org/10.1109/ICASSP48485.2024.10448310