Speech Enhancement via Time and Time-Frequency Domain Feature Fusion

Name: Wenye Zhang
SID:540473182
Github Username: imsb226
GitHub Project Link: https://github.com/imsb226/ELEC5305_Project

Project Overview

The purpose of this project is to propose a feature fusion method to more effectively extract information from 1D audio signals and spectrograms, thereby improving the effect of speech enhancement.

Background and Motivation

Deep learning methods have achieved remarkable results in current noise removal and speech enhancement tasks. Many studies have used time-domain signals or spectrograms as input.

In the time domain, the 1D U-Net framework (Kim, E., & Seo, H., 2021, Saleem, N., 2024, Qu, Q., 2025) has become the go-to backbone for extracting temporal features. This approach, through skip connections, enables multi-scale fusion, preserving more details. Building on this, Park et al. (2022) introduced a multi-view attention mechanism, surpassing the baseline at the time in terms of performance.

Meanwhile, in the time-frequency domain, current research involves first performing a Short-time Fourier transform (STFT) on the time signal and then using the spectrogram to generate a noise mask or directly generate a denoised image using a generative network (Dang et al., 2022, Cao et al., 2022). This approach in the time-frequency domain avoids the difficulty of learning features due to the length of the speech signal and allows the use of two-dimensional convolution operations for speech enhancement, significantly improving task efficiency. However, this approach will lose phase information, which affects the model's performance.

On this basis, this project proposes a method that fuses time-domain signal and spectrogram features for speech enhancement. This method allows the network to capture time-frequency information while retaining phase information, and encodes audio from multiple dimensions. In addition, considering that the model can respond to input quickly to meet industrial needs, this project does not intend to use time-consuming networks like Transformer, although studies have

shown that such models perform well in speech enhancement tasks.

Proposed Methodology

This project will use Python as a tool for experimentation. The project refers to methods such as STFT, U-Net, convolutional neural networks (CNN), long short-term memory (LSTM), Gated recurrent unit (GRU), and attention mechanisms. This project uses models such as 1D U-Net, 1D convolution, and LSTM to extract time domain information, and 2D U-Net or 2D convolution to obtain time-frequency information. Furthermore, this project will try using generative models such as GAN (Fu at el., 2021) or Diffusion (Richter at el., 2023) to directly generate more realistic speech signals, thereby avoiding the loss of speech details in regression tasks. The project will use the most common VOICEBANK + DEMAND (Valentini-Botinhao, 2017) dataset for experiments.

The training set consists of 28 subjects, with a total speech duration of approximately 9.4 hours and a sampling frequency of 48kHz. Noise is added to the speech at signal-to-noise ratios of 0, 5, 10, and 15dB. The test set consists of two unseen subjects, with a total speech duration of approximately 0.6 hours and signal-to-noise ratios of 2.5, 7.5, 12.5 and 17.5, respectively.

Expected Outcome

The final project submission will include a project file containing the scripts for the project framework and the dataset. This report will use the ForkNet model (Dang at el., 2023), also for the feature fusion task, as a baseline, and aim to surpass its performance as the final task. The expected performance metrics for this report are shown in the table below.

Expected Metric	Value
PESQ	3.2
CSIG	4.4
CBAK	3.65
COVL	3.85
STOI (%)	95

Timeline

week	Task
6	Literature review
7-10	Experiment
11	Analysis and evaluation
12-13	Final report and document

Reference:

Kim, E., & Seo, H. (2021). SE-Conformer: Time-Domain Speech Enhancement Using Conformer. In INTERSPEECH 2021 (pp. 2736–2740).

Saleem, N., Gunawan, T. S., Dhahbi, S., & Bourouis, S. (2024). Time domain speech enhancement with CNN and time-attention transformer. Digital Signal Processing, 146, 104408.

Qu, Q., Song, J., Zhang, Y., & Yuan, W. (2025). A Time-Domain Speech Enhancement Model with Controllable Output Based on Conditional Network. Circuits, Systems, and Signal Processing, 44, 5260–5278.

Park, H. J., Kang, B. H., Shin, W., Kim, J. S., & Han, S. W. (2022, May). MANNER: Multi-view attention network for noise erasure. In 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7842–7846). IEEE.

Dang, F., Hu, Q., Zhang, P., & Yan, Y. (2023). ForkNet: Simultaneous time and time-frequency domain modeling for speech enhancement. arXiv.

Dang, F., Chen, H., & Zhang, P. (2022). DPT-FSNet: Dual-Path Transformer Based Full-Band and Sub-Band Fusion Network for Speech Enhancement. In ICASSP 2022 (pp. 6857–6861).

Cao, R., Abdulatif, S., & Yang, B. (2022). CMGAN: Conformer-Based Metric GAN for Monaural Speech Enhancement. In INTERSPEECH 2022.

Fu, S.-W., Liao, C.-F., Tsao, Y., & Lin, S.-D. (2021). MetricGAN+: An improved version of MetricGAN for speech enhancement. In Proc. Interspeech 2021.

Richter, J., Welker, S., Lemercier, J.-M., Lay, B., & Gerkmann, T. (2023). Speech enhancement and dereverberation with diffusion-based generative models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 2351–2364.

Valentini-Botinhao, C. (2017). Noisy speech database for training speech enhancement algorithms and TTS models [Data set]. University of Edinburgh, School of Informatics, Centre for Speech Technology Research (CSTR).