Title
Multi-speaker Speech Separation with Permutation Invariant Training Conv-Tasnet for Hearing assistive devices

Abstract
These days, speech enhancement achieves remarkable speech clarity for hearing-aid devices using large recurrent neural networks(RNNs). However, only some models focus on both multi-speaker and multi-channel speech separation, and most deep learning models, such as convolutional recurrent neural networks(CRNN), have limitations to expand. In this work, we use Conv-Tasnet model architecture having utterance-based Permutation Invariant Training(PIT) to solve the speaker tracing issue, which is feasible to use several source separations.

Introduction
This research is for the ICASSP 2023 Clarity Challenge focusing on the problems of speech clarity in a hearing aid. The healthy ear is a complex, nonlinear system capable of operating over a large dynamic range. When the ear is damaged, a hearing aid(HA) supports this auditory system, which performs some of the amplification and compression of the sound dynamic range. In this environment, Speech enhancement(SE) can help listeners to hear clearly, especially speech.

Recent SE approaches focus on monaural speech enhancement, which means single-channel and single-speaker environments. The speech separation method aims to approximate the clean signal from the noisy signal. This process can be performed using nonlinear regression techniques, in which the training target usually uses a clean signal. Those works achieve significant work with several Deep Learning(DL) models such as Recurrent Neural Network(RNN) type [1, 2], Convolutional Recurrent Neural Network(CRN)[3], and Transformer[4]. However, the single-channel speech enhancement limits the number of speech sources. This mixture of sound from several sources leads to a cocktail party problem.

In multi-speaker speech separation works, those works approach source separation, which can distinguish several sources. For this task, several models proposed CRN[3], Conv-Tasnet[5], and Dual-Path Long short-term memory(LSTM)[6]. As we set a clean signal as a target single-speaker speech enhancement, in other ways, multi-speaker sources have several sources, which makes the output assume each source should have the same location. Kolbæk, et al 2017 [7] said this assumption had a problem[Detail] and improved utterance-based Permutation Invariant Training(uPIT)[7], which can apply to all frames. However, approaches in the frequency domain transmit phase information, which leads to the loss in the amplitude of the time-domain waveform.

This work presents Conv-Tasnet models with uPIT to resolve speaker tracing problems. This model can approach the non-linear regression in the time-domain waveform, which also relieves speaker tracing problems using uPIT technique.

Model Architecture

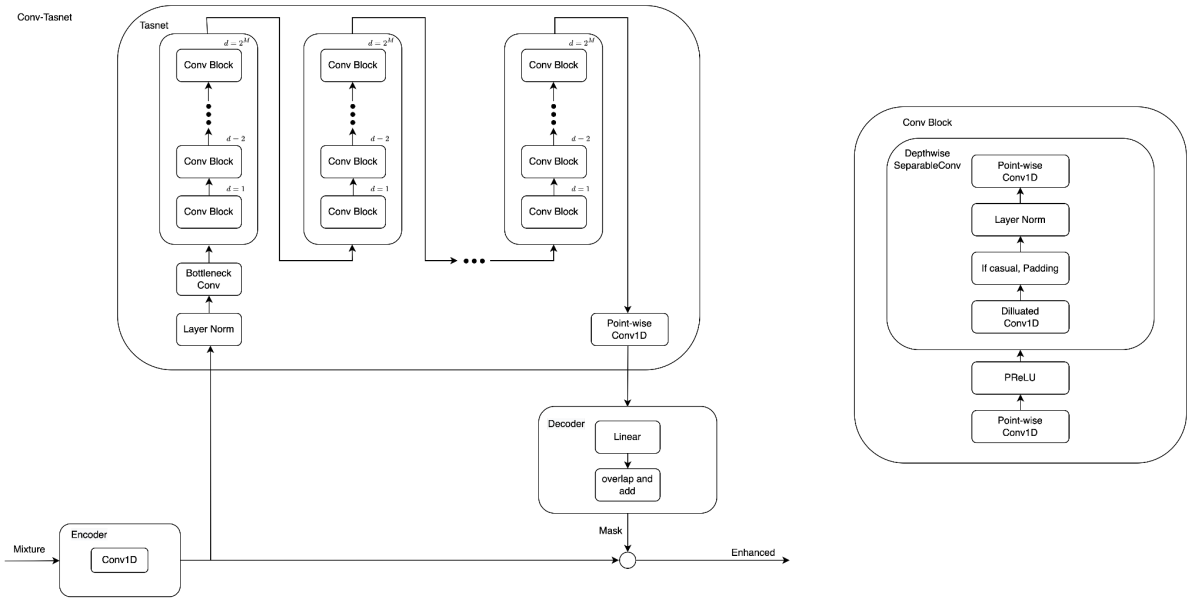

Figure 1. Model Structure of Conv-Tasnet

We have Conv-Tasnet as a baseline whose model structure is shown in [TODO] Figure 1. Conv-Tasnet has an encoder, time convolution network(TCN), and decoder, which can have segmented waveforms as input. As Fourier transforms have a windowing as frames, this model approaches 1D convolution with half of the kernel size. The second block, the time convolution network, has dilated convolutions preserving the waveform information by expanding channels. We used overlap-and-add operations as a last block decoder. The output from decoder masks encoded input, and we approximate the target signal.

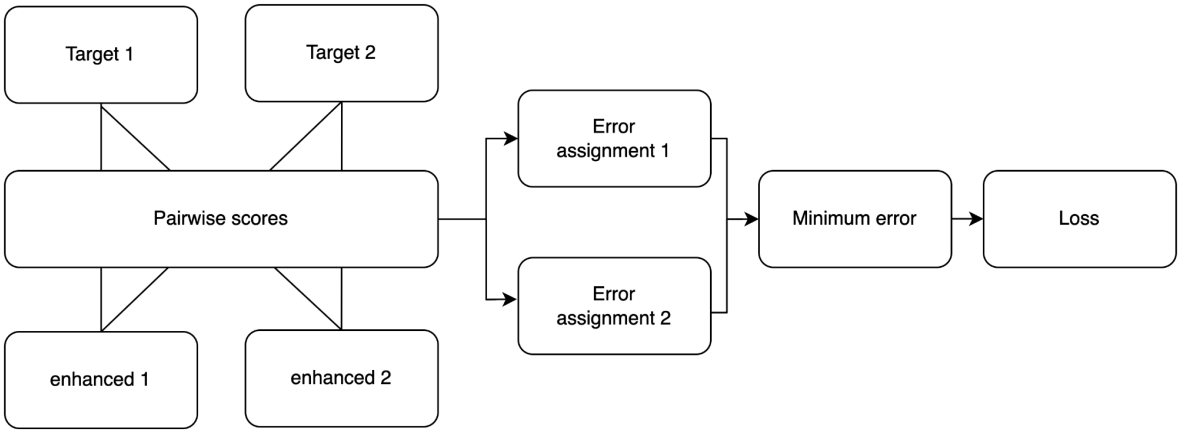

Figure 2. Utterance-based Permutation Invariant Training

As a loss function, we use uPIT because it can assume the approaches between the encoder and frequency domain as the extraction of similar features as input using windowing and convolution, and utterance is based on a frame connecting to the encoder hop length. uPIT uses Pairwise scores, which compute minimum loss with pairwise MSE.

Experiment Setup

We use the dataset Clarity Challenge 2023 provides. The dataset includes several sounds: target, interferer, and mixture. As a clean signal, we set two sources(target and interferer). As a noisy signal, we choose a mixture and keep all sources as stereo channels. These inputs randomly segment 1.024 seconds in every training step. The model parameter is as below.

N	L	B	H	P	X	R
128	40	128	256	3	7	2

N: Number of filters in encoder
L: Kernel size in encoder
B: Number of channels in bottleneck 1x1 convolution block
H: Number of channels in convolutional blocks
P: Kernel size in convolutional blocks
X: Number of convolutional blocks in each TCN
R: Number of TCN

We use global Layer Normalize as Layer Normalization, non-casual, and batch size is 4.

Currently, we have limited results for loss, Scale-Invariant Signal to Distortion Ratio, Short-Time Objective Intelligibility, HASPI, HASQI because a model is still training. The submission includes the current trained-model, and we shared the source code model[1] and evaluation[2] for HASPI and HASQI.

Reference

[1] Takeuchi, Daiki, et al. "Real-time speech enhancement using equilibriated RNN." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[2] Fedorov, Igor, et al. "TinyLSTMs: Efficient neural speech enhancement for hearing aids." *arXiv preprint arXiv:2005.11138* (2020).

[3] CRN: Tan, Ke, and DeLiang Wang. "A convolutional recurrent neural network for real-time speech enhancement." *Interspeech*. Vol. 2018. 2018.

[4] Transformer Subakan, Cem, et al. "Attention is all you need in speech separation." *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

[5] Conv-Tasnet: Luo, Yi, and Nima Mesgarani. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation." *IEEE/ACM transactions on audio, speech, and language processing* 27.8 (2019): 1256-1266.

1. https://github.com/ooshyun/SpeechEnhancement-Pytorch/tree/dev
2. https://github.com/ooshyun/ClarityChallenge2023/tree/dev

[6] Dual-Path Pandey, Ashutosh, and DeLiang Wang. "Dual-path self-attention RNN for real-time speech enhancement." *arXiv preprint arXiv:2010.12713* (2020).

[7] uPIT Kolbæk, Morten, et al. "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.10 (2017): 1901-1913.