On Permutation Invariant Training for Speech Source Separation

Paper ID: 3140

Xiaoyu Liu, Jordi Pons

Dolby Laboratories

Introduction

Permutation ambiguity and utterance level PIT (uPIT)

- C! prediction to target loss pairs for training a speaker independent network.
- uPIT minimizes the smallest separation error of all utterance-level permutations but causes local speaker swaps and leakage between separated signals.

Frame level PIT (tPIT) + clustering (Deep CASA)

- Optimize separation for each frame independently in the STFT domain by tPIT.
- Followed by a clustering model to track permutation across frames.

uPIT + speaker-ID loss

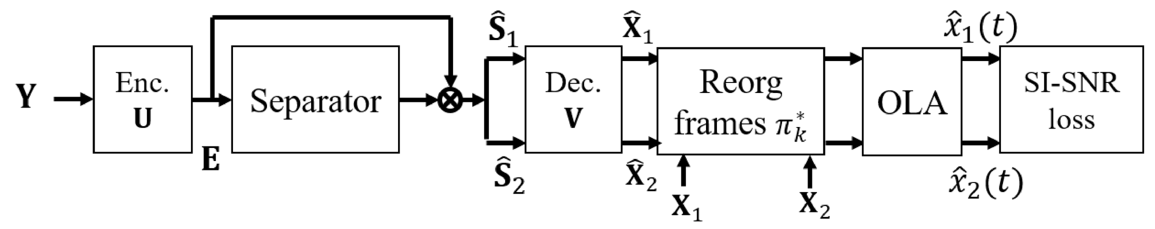
- uPIT aligns the order of the separation signals with the input.
- A loss in the speaker embedding space strengthens speaker consistency.

Our work

- Extend tPIT + clustering to waveform-based models (Conv-TasNet).
- Propose an efficient loss for the clustering stage in waveform-based models.
- Study three domains for tPIT: waveform, latent space, STFT (Deep CASA).
- Extend uPIT + speaker-ID to uPIT + PASE and compare with tPIT + clustering.

tPIT for Conv-TasNet

tPIT-time

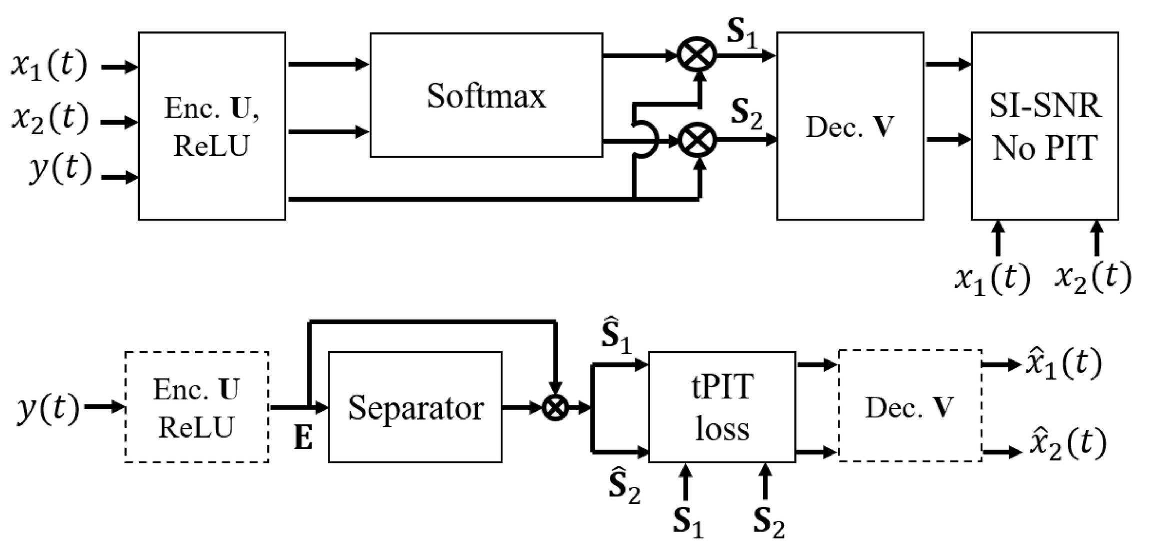


- $\hat{X}_1, \hat{X}_2$: matrices containing separated short waveform frames (2 ms/frame)
- tPIT finds the best permutation π_k^* for the kth frame, and reorders frames

$$\pi_k^* = \arg\min_{\pi_k \in P} \sum_{c=1}^{C} |\hat{\mathbf{x}}_{c,k} - \mathbf{x}_{\pi_k(c),k}|$$

- After overlap-and-add, SI-SNR loss is minimized to optimize the model

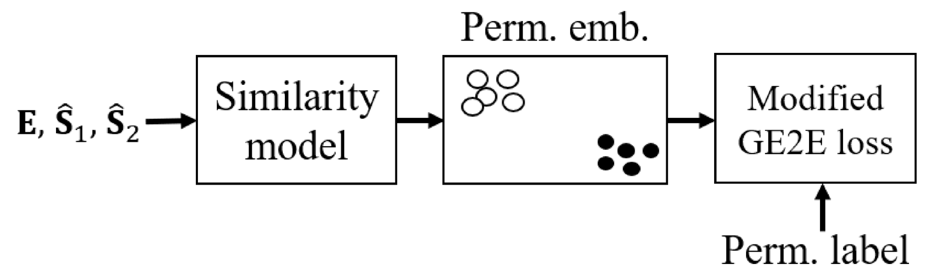
tPIT-latent



- First train the enc./dec. to generate the N-dim ideal latent features S_1, S_2.
- Then train the separator by tPIT in the latent space for each frame k:

$$loss_{tPIT} = \frac{1}{KNC} \sum_{k=1}^{K} \min_{\pi_k \in P} \sum_{c=1}^{C} |\hat{\mathbf{s}}_{c,k} - \mathbf{s}_{\pi_k(c),k}|$$

An efficient loss for training the clustering model

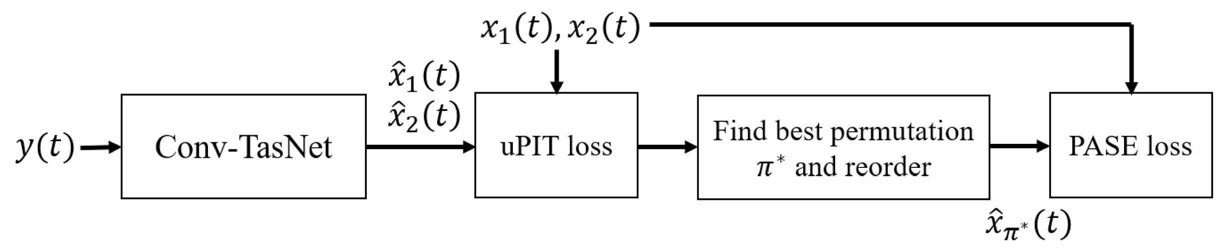


- A clustering model is needed to track permutations across frames at test time.
- The clustering model maps the separated frames to a permutation embedding space, such that each cluster contains frames with the same permutation.
- Deep CASA optimizes pairwise distance between every pair of frames, which is too expensive for waveform-based models, due to very short frame shift.
- We propose to use the generalized end-to-end (GE2E) loss, which only compares each frame with the centroids of the clusters
- For the kth frame that belongs to permutation p, we optimize

$$loss_{GE2E} = \sum_{k=1}^{K} -\log \frac{\exp -d(\mathbf{h}_{k,p}, \mathbf{m}_p)}{\sum_{i=1}^{C!} \exp\left(-d(\mathbf{h}_{k,p}, \mathbf{m}_i)\right)}$$

where $\boldsymbol{h}_{k,p}$ is the permutation embedding, $\boldsymbol{m}_p$ is the pth cluster center, and $d(\boldsymbol{x}, \boldsymbol{y}) = w||\boldsymbol{x} - \boldsymbol{y}||^2 + b$ is the Euclidean distance with learnable w and b.

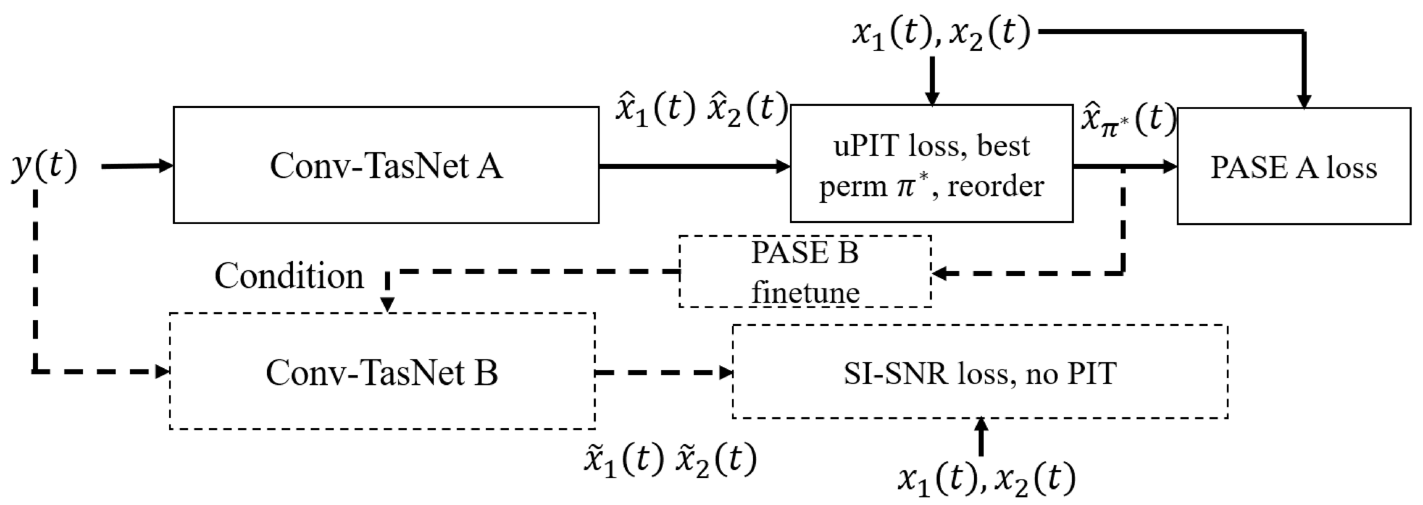
uPIT + PASE



- PASE is a pretrained problem agnostic speech encoder, that generates features with various speech information, such as pitch, speaker-ID, phoneme.
- uPIT finds the best utterance permutation π^*, and reorders outputs to align with the reference signals.
- The PASE loss implicitly enforces permutation consistency across frames.

$$loss = uPIT + \sum_{c=1}^{C} \|PASE(\hat{x}_c(t)) - PASE(x_{\pi_u^*(c)}(t))\|^2$$

Conditioning Conv-TasNet on PASE



- Investigate if model conditioning could further reduce permutation errors.
- First train a Conv-TasNet by uPIT + PASE loss.
- Then train another Conv-TasNet conditioned on PASE features of the separated signals (reordered by the best permutation) from the first Conv-TasNet.
- The PASE encoder in the second stage is finetuned with the Conv-TasNet.

Experiment setup

- Trained models on the WSJ0-2mix dataset.
- Evaluated models on the test sets of WSJ0-2mix, Libri-2mix, and VCTK-2mix.
- 8 kHz data, except for the uPIT+PASE experiments, which uses 16 kHz version.
- PASE pretrained on 50 hours of LibriSpeech data (2338 speakers).

Evaluation metrics

- SI-SNRi (dB): measures separation quality.
- Frame error rate, FER (%): percentage of frame level permutation errors.
- Hard sample rate, HSR (%): percentage of test samples with SI-SNRi < 5 dB.

SI-SNRi results of tPIT + clustering

	WSJ0	Libri	VCTK
uPIT-waveform	15.9	10.4	9.4
uPIT-STFT	15.5	11.4	12.7
tPIT-STFT + optimal clusters	18.5	16.0	15.5
tPIT-STFT + clustering	17.5	13.9	13.6
tPIT-time + optimal clusters	16.7	12.1	13.0
tPIT-time + clustering	15.5	9.8	9.9
tPIT-latent: enc/dec (Fig. 2, top)	55.5	54.9	53.9
tPIT-latent + optimal clusters	17.6	12.9	13.7
tPIT-latent + clustering	16.5	11.0	11.0
tPIT-latent + clustering: clustering loss variants			
pairwise similarity loss	16.2	10.7	10.8
GE2E loss	16.5	11.0	11.0

FER and HSR of tPIT + clustering

	uPIT (waveform)		tPIT-latent + clustering		tPIT-STFT + clustering	
	FER	HSR	FER	HSR	FER	HSR
WSJ0	6.1	6.0	5.4	1.8	4.9	2.2
Libri	9.4	14.8	8.5	9.1	6.6	7.4
VTCK	12.3	22.8	9.4	10.7	7.8	7.2

- uPIT trained Deep CASA (uPIT-STFT) generalizes better than Conv-TasNet .
- tPIT-latent performs better than tPIT-time, for both the ground truth (optimal) clusters and the predicted clusters.
- tPIT-STFT + clustering (Deep CASA) performs the best, indicating the advantage of STFT in terms of reducing permutation errors and improving generalization.
- GE2E loss is more effective than the pairwise loss.

uPIT + PASE results

	uPIT-waveform		uPIT+PASE		uPIT+PASE cascaded		tPIT-latent+clustering	
	SI-SNRi	FER	SI-SNRi	FER	SI-SNRi	FER	SI-SNRi	FER
WSJ0	15.5	5.2	15.9	4.5	17.5	4.6	16.0	4.3
Libri	10.7	9.0	10.8	8.0	11.9	7.6	11.1	7.8
VTCK	9.5	12.4	9.9	11.2	10.9	11.3	10.9	9.5

- uPIT + PASE improves uPIT, but not by as much as tPIT-latent + clustering
- The additional conditioning improves SI-SNRi, but does not further reduce permutation errors.