# MULTI-DECODER DPRNN: SOURCE SEPARATION FOR VARIABLE NUMBER OF SPEAKERS

Junzhe Zhu  Raymond A. Yeh  Mark Hasegawa-Johnson

University of Illinois at Urbana-Champaign

## MOTIVATION

**Task: Source separation for variable number of speakers. Given audio mixture $M = \sum S_1, S_2, S_3, ...$ we want to find the sources $S_n$, without prior knowledge of how many sources there are**

## INTRODUCTION

Due to the recent progress in deep learning, supervised methods have received a lot of interest. For example, Luo et al. presented a dual-path RNN architecture to better capture both short and long-term features. However, these works have focused on the setting where the number of speakers is a priori known.

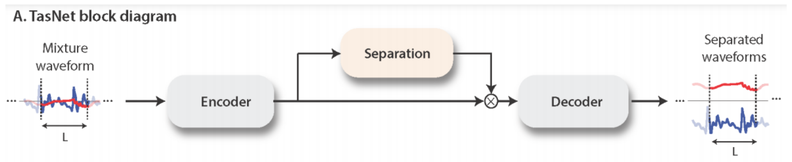



**Figure 1:** Block diagram of DPRNN. The decoder part includes a projection layer whose output size is the product of hidden size and number of output sources

## EVALUATION METRIC

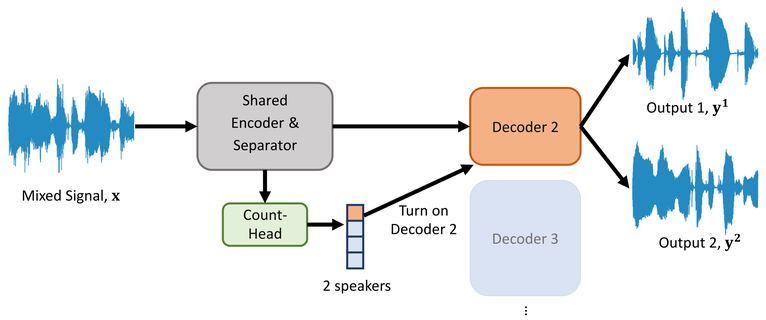
To deal with the situation when estimation and target have different number of sources, we add a penalty term to the error metric based on number of mismatching sources.

$$\frac{1}{|\mathcal{D}|}\sum_{(\mathbf{x},\mathcal{Y})\in\mathcal{D}}\frac{1}{\max(|\mathcal{Y}|,|\hat{\mathcal{Y}}|)}(\mathcal{L}_{\text{match}}+\mathcal{L}_{\text{pad}})$$

where $\mathcal{L}_{\text{match}}$ is the Si-SNR between matched sources, $\mathcal{L}_{\text{pad}} = \mathcal{P}_{\text{ref}} \cdot \Big||\mathcal{Y}| - |\hat{\mathcal{Y}}|\Big|$. is the penalty term for mismatching sources, and we set $\mathcal{P}_{\text{ref}}$ to either be -30 or average Oracle SNR of the measured system.

## OUR APPROACH

We use the same encoder and separation(LSTM) network, but a list of decoders instead of just one. Each decoder corresponds to a different number of sources. For example, a decoder for 2 sources, another for 3 sources, and so on.




Each decoder takes the same input, but has a different number of output channels (achieved by a different projection layer) We also train a classifier(count-head) that selects which decoder to use. During training, we select the decoder based on the ground-truth. During test-time, we select the decoder based on the classifier output.

Training steps:
$M \leftarrow$ Mixture Signal
$S \leftarrow$ Ground Truth Sources$\{S_1, S_2, S_3, ...\}$
$\alpha \leftarrow$ Balancing Factor
$Z \leftarrow$ Separator(Encoder($M$))
$P \leftarrow$ Count_Head($Z$)
$\hat{S} \leftarrow$ Decoders_List$[|S|](Z)$
$Loss = (1-\alpha) \times$ Si-SNR$(S, \hat{S}) + \alpha \times$ Cross-Entropy$(P, |S|)$

Inference steps:
$M \leftarrow$ Mixture Signal
$Z \leftarrow$ Separator(Encoder($M$))
$P \leftarrow$ Count_Head($Z$)
$\hat{S} \leftarrow$ Decoders_List[argmax($P$)]($Z$)
Project page: `https://junzhejosephzhu.github.io/Multi-Decoder-DPRNN/`

## RESULTS

| Pred \ True | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| **2** | 2998 | 17 | 1 | 0 |
| **3** | 2 | 2977 | 27 | 0 |
| **4** | 0 | 6 | 2928 | 80 |
| **5** | 0 | 0 | 44 | 2920 |

**Table 1:** Speaker counting confusion matrix for the proposed count-head method

| Model | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Conv-Tasnet* | 15.3 | 12.7 | - | - |
| DPRNN* | 18.8 | - | - | - |
| DPRNN* | 18.21 | 14.71 | 10.37 | 8.65 |
| Mulcat* | 20.12 | 16.85 | 12.88 | 10.56 |
| Attractor Network | 15.3 | 14.5 | - | - |
| OR-PIT | 14.8 | 12.6 | 10.2 | - |
| Ours | 19.1 | 14.1 | 9.3 | 5.9 |

**Table 2:** Oracle SNR; Each column shows results averaged from all mixtures with corresponding number of speakers. *models above double-line are models with fixed number of speakers.

| $\mathcal{P}_{\text{ref}} = -30dB$ | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Model-Select(DPRNN)* | 15.2 | 10.7 | 6.0 | 7.7 |
| Model-Select(Mulcat)* | 17.5 | 13.21 | 8.4 | 10.0 |
| Attractor Network | 14.7 | 14.2 | - | - |
| OR-PIT | 13.1 | | - | - |
| Ours | 19.1 | 14.0 | 9.2 | 5.8 |

**Table 3:** P-SI-SNR of each model; For OR-PIT, result is computed by averaging the P-SI-SNR for both 2 and 3 speakers computed with 95.7% recall. Note that models with lower max speaker count generally have higher accuracy, since fewer classes implies a higher P-SI-SNR. * denotes models trained on fixed number of speakers.