# Mixup Augmentation for Generalizable Speech Separation

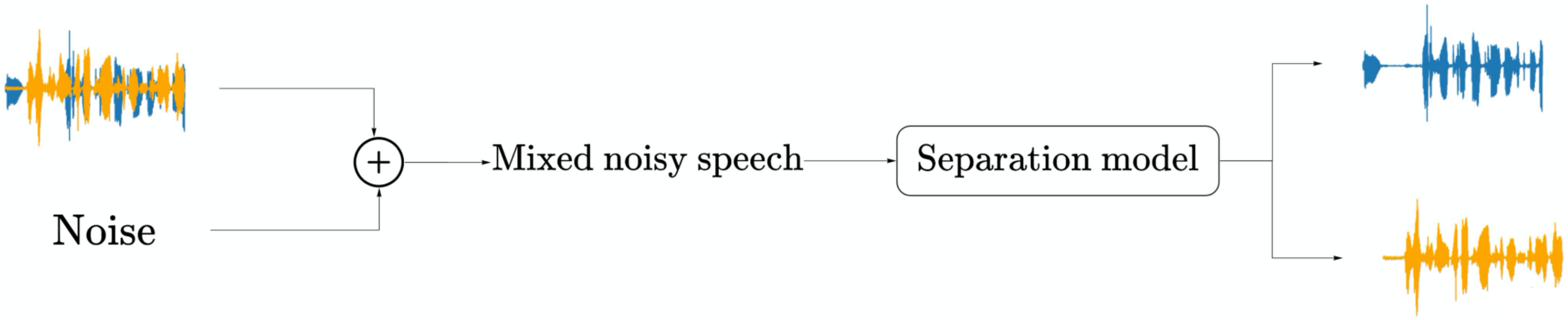
Ashish Alex[1], Lin Wang[1], Paolo Gastaldo[2], Andrea Cavallaro[1]

1. Centre of Intelligence Sensing, Queen Mary University of London
2. DITEN, University of Genoa

MMSP 2021

# Introduction

- Single channel speech separation in noisy environments




- Applications: Hearing aids, captioning & transcription (YouTube), human robotic interaction, automatic speech recognition

# Motivation

- Motivation
  - Improve generalization of separation models across datasets
  - Improve separation performance in unseen noisy conditions
  - Traditional regularization techniques, augmentations did not improve generalization

- Contributions
  - Extend Mixup augmentation and variations for time-domain speech separation
  - Proposed Data-only Mixup improves inter corpus separation performance

# Separation model architecture

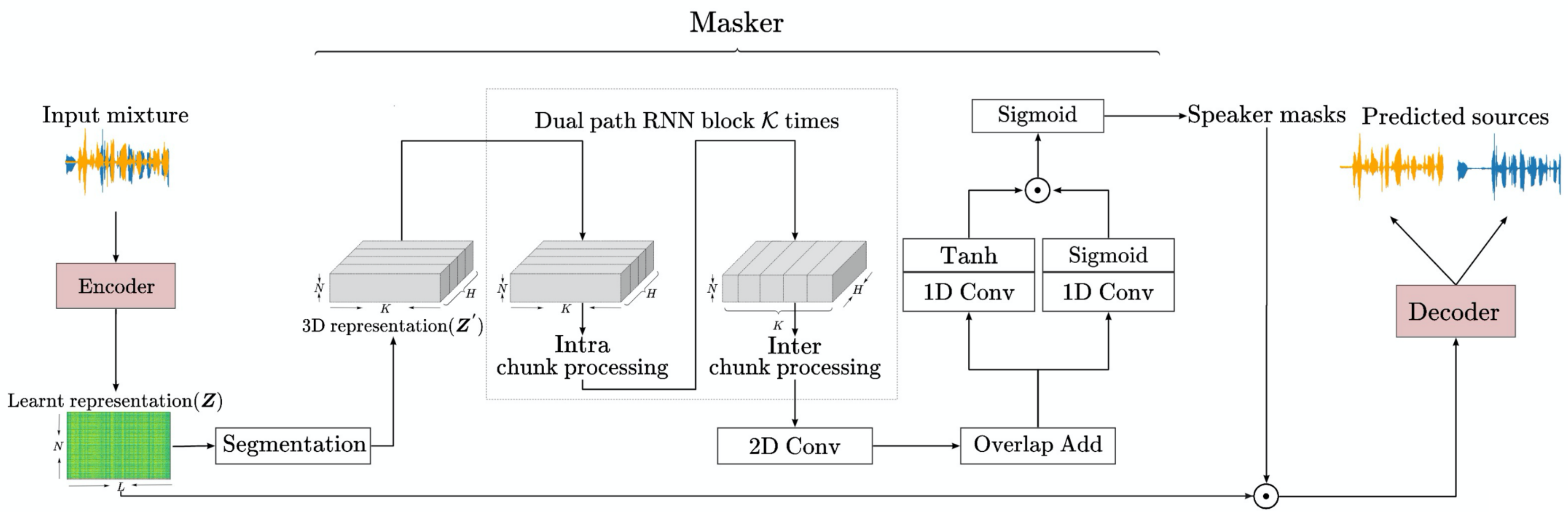
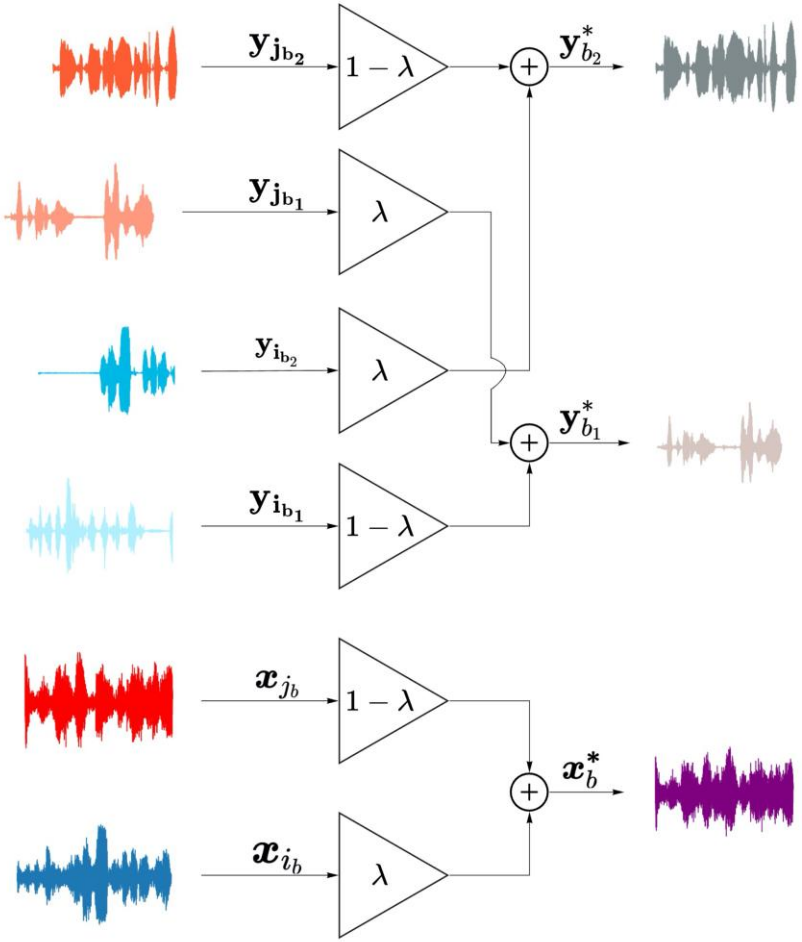



**Fig.** DPRNN [1] separation model

[1] Luo et al., Dual-path RNN: efficient long sequence modelling for time-domain single-channel speech Separation, in Proc. Interspeech, 2019.

# Mixup




$\boldsymbol{x}_{i_b}, \boldsymbol{x}_{j_b}$ : Two distinct mixtures from the mini-batch
$\boldsymbol{y}_{i_{b_1}}, \boldsymbol{y}_{i_{b_2}}$ and $\boldsymbol{y}_{j_{b_1}}, \boldsymbol{y}_{j_{b_2}}$ : ground-truth speech

$\boldsymbol{x}^*_b$ : Augmented mixture
$\boldsymbol{y}^*_{b_1}$ and $\boldsymbol{y}^*_{b_2}$ : Augmented ground truth speech

$\lambda = beta(\alpha, \beta)$

# Mixup variants

- Complete Mixup: Augment training data using Mixup for all epochs

$$\mathcal{L}_{\text{CP}} = \mathcal{L}_{\text{augment}}, \quad 0 < e < E_{\text{max}}$$

$E_{max}$ : Maximum number of epochs for training

- Partial Mixup: Regular training in initial epochs followed by Mixup Augmentation in subsequent epochs

$$\mathcal{L}_{\text{PA}} = \begin{cases} \mathcal{L}_{\text{regular}}, & 0 < e \leq E_{\text{early}} \\ \mathcal{L}_{\text{regular}}, & (E_{\text{early}} < e < E_{\text{max}}) \wedge (e|Q \neq 0) \\ \mathcal{L}_{\text{augment}}, & (E_{\text{early}} < e < E_{\text{max}}) \wedge (e|Q = 0) \end{cases}$$

$E_{max}$ : Maximum number of epochs for training

$E_{early}$ : Number of epochs until which Augmentation is applied for

# Mixup variants

- Pretrained Mixup: Fine tune a pretrained model using Mixup Augmentation

$$\mathcal{L}_{\text{PT}} = \begin{cases} \mathcal{L}_{\text{regular}}, & 0 < e \leq E_{\max} \\ \mathcal{L}_{\text{augment}}, & E_{\max} < e < E_{\text{pt}} \end{cases}$$

$E_{pt}$ : Maximum number of epochs pre-trained model is finetuned for

- Data only Mixup
  - Apply Mixup on mixtures only
  - Keep ground truth as most dominant sources in Mixup augmented mixture

$$\begin{cases} \boldsymbol{x}_b^\circ = \lambda \boldsymbol{x}_{i_b} + (1 - \lambda)\boldsymbol{x}_{j_b} \\ \boldsymbol{Y}_b^\circ = \boldsymbol{Y}_{i_b} \end{cases}$$

$$\mathcal{L}_{\text{DO}} = \mathcal{L}(\boldsymbol{Y}^\circ, \hat{\boldsymbol{Y}}^\circ), \quad 0 < e < E_{\max}$$

# Experiments & datasets

| Dataset | Split | Hours | Speakers | Noise corpus |
|---|---|---|---|---|
| LibriMix[2] | train-100 | 58 | 251 | WHAM[5] |
| LibriMix[2] | test | 11 | 40 | WHAM[5] |
| VCTK[3] | test | 9 | 109 | WHAM[5] |
| TIMIT[4] | test | 10 | 630 | Env noise corpus[6] |

- Intra corpus - Train on LibriMix (train-100) & test on LibriMix (test)
- Inter corpus – Train on LibriMix (train-100) & test on TIMIT (test) and VCTK (test)

[2] Cosentino et al., LibriMix: An Open-Source Dataset for Generalizable Speech Separation, arXiv preprint arXiv:2005.11262
[3] C. Veaux et al., Superseded-CSTR VCTK corpus: English multispeaker corpus for cstr voice cloning toolkit, 2016
[4] J. S. Garofolo, TIMIT acoustic phonetic continuous speech corpus, Linguistic data consortium, 1993
[5] Wichern et al., WHAM!: Extending speech separation to noisy environments, in Proc. Interspeech, 2019
[6] Xu et al. A regression approach to speech enhancement based on deep neural networks, IEEE/ACM Trans. Audio, Speech, Lang. Process., 2014

# Ablation of scalars for Mixup

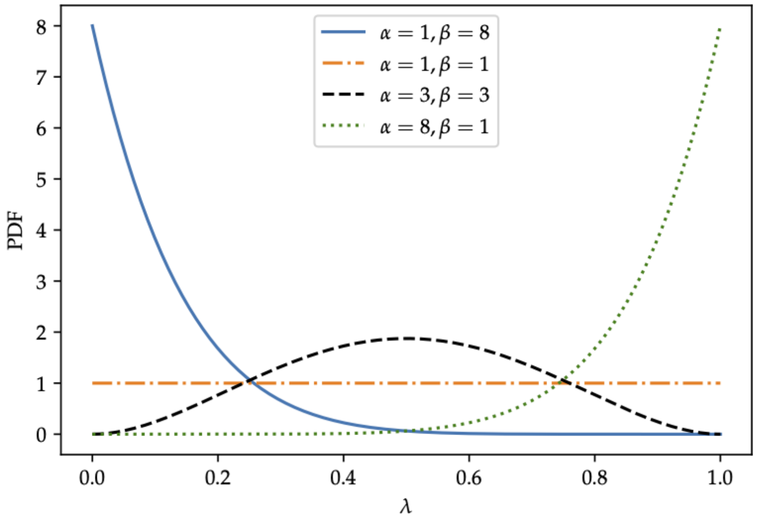



**Fig.** Probability density function of beta distribution with different $\alpha$ and $\beta$. $\lambda = beta(\alpha, \beta)$

| $\alpha \backslash \beta$ | 1 | 3 | 8 |
|---|---|---|---|
| 1 | 11.69 | 11.70 | 11.51 |
| 3 | 11.64 | 11.70 | 11.64 |
| 8 | **11.97** | 11.70 | 11.57 |

(a) Complete Mixup

| $\alpha \backslash \beta$ | 1 | 3 | 8 |
|---|---|---|---|
| 1 | 11.17 | 7.60 | 5.89 |
| 3 | 11.55 | 11.31 | 7.12 |
| 8 | **12.00** | 11.51 | 11.26 |

(b) Data Only Mixup

**Table.** SI-SNRi (dB) of data-augmented DPRNN for various values of $\alpha$ and $\beta$

# Intra corpus results

| Augmentation type | Augmentation variation | SI-SNRi |
|---|---|---|
| None | - | 12.00 |
| SpecAugment[7] | Frequency masking | 11.63 |
| | Time masking | 12.04 |
| | T-F masking | **12.05** |
| Mixup | Complete | 11.97 |
| | Data-only | 12.00 |
| | Partial | 11.50 |
| | Pre-trained | 12.00 |

**Table:** Model trained and tested on LibriMix dataset

- None of the augmented models significantly outperform non-augmented model

[7] Park et al., SpecAugment: A simple data augmentation method for automatic speech recognition, in Proc. Interspeech, 2019

# Inter corpus results (1)

| SNR | UAUG | SpecAugment | | | Mixup | | | |
|---|---|---|---|---|---|---|---|---|
| | | TM | FM | T-F | PA | PT | CP | DO |
| -5 | 4.95 | 5.09 | 4.99 | 4.53 | 4.86 | 5.19 | 4.95 | 5.61 |
| 0 | 5.41 | 5.76 | 5.94 | 4.84 | 5.38 | 5.69 | 5.85 | 6.60 |
| 5 | 6.52 | 6.62 | 6.59 | 6.10 | 6.34 | 6.95 | 6.87 | 8.48 |
| 10 | 8.24 | 8.32 | 8.18 | 8.18 | 8.39 | 8.81 | 8.84 | 10.25 |
| 15 | 9.80 | 10.22 | 9.85 | 9.82 | 10.21 | 10.64 | 10.33 | 11.42 |
| 20 | 10.93 | 11.24 | 11.08 | 11.30 | 10.92 | 11.84 | 10.94 | 11.97 |
| Avg | 7.64 | 7.87 | 7.77 | 7.46 | 7.68 | 8.13 | 7.96 | **9.06** |

**Table:** Model trained on LibriMix and tested on TIMIT dataset

- Data-only Mixup improves separation performance on TIMIT dataset
- Noise types & speakers in TIMIT are different from LibriMix

# Inter corpus results (2)

| Augmentation type | Augmentation variation | SI-SNRi |
|---|---|---|
| None | - | 11.07 |
| SpecAugment[7] | Frequency masking | 10.79 |
| | Time masking | 11.09 |
| | T-F | 11.04 |
| Mixup | Complete | 11.11 |
| | Data-only | **11.43** |
| | Partial | 10.93 |
| | Pre-trained | 11.06 |

**Table:** Model trained on LibriMix and tested on VCTK dataset

- Data-only Mixup slightly improves separation performance on VCTK dataset
- Speakers in VCTK are different from LibriMix
- Noise samples in VCTK dataset is the same as LibriMix dataset

[7] Park et al., SpecAugment: A simple data augmentation method for automatic speech recognition, in Proc. Interspeech, 2019

# Conclusion & future work

- Data-only Mixup augmentation improves cross-corpus performance for speech separation model

- Data augmentation approach doesn’t incur additional in network parameters

- Future work – Finding optimal augmentation combinations using learnt augmentation strategies