# MIXUP-BREAKDOWN: A CONSISTENCY TRAINING METHOD FOR IMPROVING GENERALIZATION OF SPEECH SEPARATION MODELS

Max W. Y. Lam, Jun Wang, Dan Su, Dong Yu

# Conventional Supervised Learning

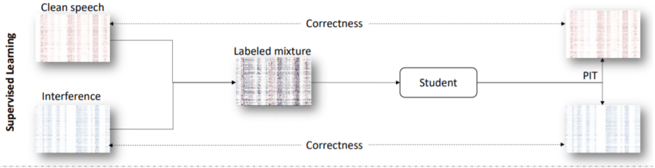
We have a labeled training set of $N_L$ input-output pairs $D_L = \{x_i, y_i\}_{i=1}^{N_L}$, where $y = (s, e)$, $x = s + e$, $s$ - clean speech signal, $e$ -interference signal.And unlabeled data $D_U = \{x_j\}_{j=1}^{N=N_L+N_u}$

In a supervised learning framework, given a speech separation model $f_\theta$ with parameters $\theta$, an objective function $\mathcal{L}(f_\theta(x), y)$ is usually defined as the divergence between the predicted outputs $f_\theta(x) = (\hat{s}, \hat{e})$ and the original clean sources $y$.

$$\mathcal{L}(f_\theta(x), y) = \min_{u \in \{\hat{s}, \hat{e}\}} \mathcal{L}_{SI-SNR}(s, u) + \min_{v \in \{\hat{s}, \hat{e}\}} \mathcal{L}_{SI-SNR}(e, v)$$

$$\mathrm{L}_{SI-SNRI}(a, b) = -10 log_{10} \frac{||\Pi_a(b)||_2^2}{||b - \Pi_a(b)||_2^2}$$

where $\Pi_a(b) = a^T b / ||a||_2^2 \cdot a$ is a projection of b onto a.



# Conventional Supervised Learning

Assuming that the input-output pairs follow a joint distribution $P(x, y)$, which is usually unknown, we minimize the average of the objective function over the joint distribution, i.e., the expected risk, to find an optimal set of parameters $\theta^*$ :

$$\theta^* \approx arg \min_\theta \int \mathcal{L}(f_\theta(x), y) dP_{EMP}(x, y; D_L) = arg \min_\theta \frac{1}{N_L} \sum_{i=1}^{N_l} \mathcal{L}(f_\theta(x_i), y_i)$$

We approximate the unknown joint data distribution P(x, y), an empirical distribution is used:

$$P_{EMP}(x, y; D_L) = \frac{1}{N_L} \sum_{i=1}^{N_l} \delta(x = x_i, y = y_i)$$

is also known as **Empirical Risk Minimization (ERM)**.

# Mixup approach[1]

In the Vicinal Risk Minimization (VRM) principle (Chapelle et al., 2000), the distribution P is approximated by:

$$P_\nu(\tilde{x}, \tilde{y}) = \frac{1}{n} \sum_{i=1}^{n} \nu(\tilde{x}, \tilde{y} | x_i, y_i)$$

To learn using VRM, we sample the vicinal distribution to construct a dataset $D_\nu := \{\hat{x}_i, \hat{y}_i\}_{i=1}^{m}$ and minimize the empirical vicinal risk:

$$R_\nu(f) = \frac{1}{m} \sum_{i=1}^{m} \ell(f(\tilde{x}_i), \tilde{y}_i)$$

[1] mixup: BEYOND EMPIRICAL RISK MINIMIZATION,Hongyi Zhang Moustapha Cisse, Yann N. Dauphin, David Lopez-Paz 2018

# Mixup approach

We get a generic vicinal distribution called *mixup*:

$$\mu(\tilde{x}, \tilde{y}|x_i, y_i) = \frac{1}{n}\sum_{j}^{n} E_\lambda[\delta(\tilde{x} = \lambda \cdot x_i + (1-\lambda)\cdot x_j, \tilde{y} = \lambda \cdot y_i + (1-\lambda)\cdot y_i)]$$

where $\lambda \sim Beta(\alpha, \alpha)$ for $\alpha \in (0, \infty)$

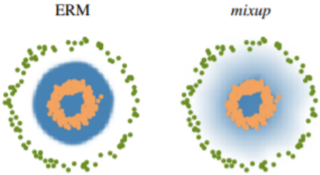



Рис.: Effect of mixup ($\alpha = 1$) on a toy problem. Green: Class 0. Orange: Class 1. Blue shading indicates $p(y = 1|x)$.

# Mixup-Breakdown

Let's introduce Mixup and Breakdown operations:

$$Mix_\lambda(a, b) \triangleq \lambda \cdot a + (1 - \lambda) \cdot b$$

$$Break_\lambda(a, b) \triangleq (\lambda \cdot a, (1 - \lambda) \cdot b)$$

where a and b two arbitrary signals and $\lambda \sim Beta(\alpha, \alpha)$ for $\alpha \in (0, \infty)$ is inherited from the mixup approach. The Mixup-Breakdown (MB) strategy trains a student model $f_{\theta_S}$ to provide consistent predictions with the teacher model $f_{\theta_T}$ of the same network structure at perturbations of predicted separations from the input mixtures (either labeled or unlabeled):

$$f_{\theta_S}(Mix_\lambda(f_{\theta_T}(x_j))) \approx Break_\lambda(f_{\theta_T}(x_j))$$

Mathematically, the MB operation can view as a generic augmentation of the empirical distribution:

$$dP_{EMP}(\tilde{x}, \tilde{y}; D) = \frac{1}{N} \sum_{i=1}^{N} v(\tilde{x}, \tilde{y}|x_i)$$

$$v(\tilde{x}, \tilde{y}|x_i) = E_\lambda[\delta(\tilde{x} = Mix_\lambda(f_{\theta_T}(x_i)), \tilde{y} = Break_\lambda(f_{\theta_T}(x_i)))]$$

# Mixup Breadown Training

In this way we present a new consistency-based training method, namely, Mixup-Breakdown Training (MBT):

$$\theta_S^* \approx \underbrace{\Big[\int \mathcal{L}(f_{\theta_S}(x), y) dP_{EMP}(x, y; D_L)}_{Correctnes} +$$

$$r(t) \underbrace{\int \mathcal{L}(f_{\theta_S}(\tilde{x}), \tilde{y}) dP_{MBT}(\tilde{x}, y \tilde{;} D)\Big]}_{Consistensy} =$$

$$= arg \min_{\theta_S} \Big[ \frac{1}{N_L} \sum_{i=1}^{n_L} \mathcal{L}(f_{\theta_S}(x_i), y_i) +$$

$$+ \frac{r(t)}{N} \sum_{j=1}^{N} \mathcal{L}(f_{\theta s}(Mix_\lambda(f_{\theta_T}(x_j))), Break_\lambda(f_{\theta_T}(x_j))) \Big]$$

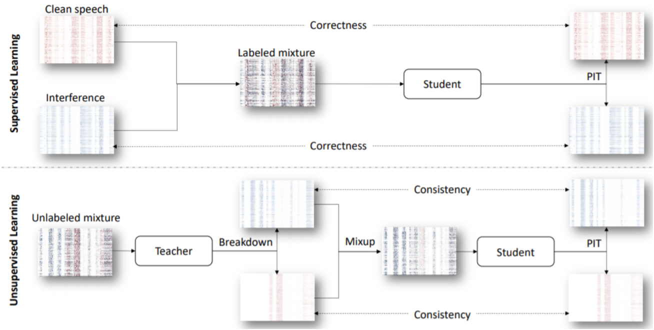



Рис.: Mixup-Breakdown Training

# Experiments

- **Data**
  - WSJ0-Libri: using clean speech drawn from the publicly available Librispeech 100h training corpus.
  - WSJ0-music: using music clips drawn from a 43-hour music dataset that contains various classical and popular music genres, e.g., baroque, classical, romantic, jazz,country, and hip-hop.
  - WSJ0-noise: using noise clips drawn from a 4-hour recording collected in various daily life scenarios such as office, restaurant, supermarket, and construction place.
- **Implementation Details** Authors implemented the mixup, MT, ICT, and our proposed MBT to train Conv-TasNet for comparative performance analysis. In all SSL settings, we set the same decay coefficient for the mean-teacher to 0.999, and the same ramp function r(t) = exp(t/Tmax 1) for t 1, ..., Tmax, where Tmax = 100 was the maximum number of epochs. Besides, we set $\alpha = 1$ , so that $\lambda$ becames uniformly distributed in [0, 1].

# “online” data augmentation for purely supervised learning

| Method | Params. | Trained | SI-SNRi |
|---|---|---|---|
| DPCL++[1] | 13.6M | WSJ0-2mix | 10.8 |
| DANet [2] | 9.1M | | 10.5 |
| ADANet [4] | 9.1M | | 10.4 |
| Chimera++ [30] | 32.9M | | 11.5 |
| WA-MISI-5 [31] | 32.9M | | 12.6 |
| BLSTM-TasNet [32] | 23.6M | | 13.2 |
| *Conv-TasNet | 8.8M | | 15.3 |
| *MBT | 8.8M | WSJ0-2mix+ “online” data augmentation | **15.5** |
| *MBT | 8.8M | WSJ0-2mix+ Unlabeled WSJ0-multi | **15.6** |

Рис.: Comparison of performances on the WSJ0-2mix dataset

# Generalization Capability.Mismatch Speech Interference

| Method | Trained on | Tested on | SI-SNRi |
|---|---|---|---|
| ERM | WSJ0-2mix | WSJ0-Libri | 13.56 |
| mixup | | | 13.58 |
| MBT | | | **13.75** |
| MT | WSJ0-2mix+ Unlabeled WSJ0-Libri | | 13.81 |
| ICT | | | 13.78 |
| MBT | | | **13.95** |
| MBT | WSJ0-2mix+ Unlabeled WSJ0-multi | | 13.88 |

Рис.: Separation performance of different training approaches in the presence of mismatch speech interference

# Generalization Capability. Mismatch Background Noise Interference

| Method | Trained on | Tested on | SI-SNRi |
|---|---|---|---|
| ERM | WSJ0-2mix | WSJ0-noise | 1.86 |
| mixup | | | 1.91 |
| MBT | | | **2.10** |
| MT | WSJ0-2mix + Unlabeled WSJ0-noise | | 12.51 |
| ICT | | | 12.36 |
| MBT | | | **13.21** |
| MBT | WSJ0-2mix + Unlabeled WSJ0-multi | | **13.52** |

Рис.: Separation performance of different training approaches in the presence of mismatch background noise interference

# Generalization Capability.Mismatch Music Interference

| Method | Trained on | Tested on | SI-SNRi |
|---|---|---|---|
| ERM | WSJ0-2mix | WSJ0-music | 1.93 |
| mixup | | | 1.94 |
| MBT | | | **1.99** |
| MT | WSJ0-2mix + Unlabeled WSJ0-music | | 14.12 |
| ICT | | | 14.02 |
| MBT | | | **15.95** |
| MBT | WSJ0-2mix + Unlabeled WSJ0-multi | | 15.67 |

Рис.: Separation performance of different training approaches in the presence of mismatch music interference

Thank you for your attention!