Application of Mixup Breakdown algorithm to improve speaker diarization

Svetlana Kuchuganova

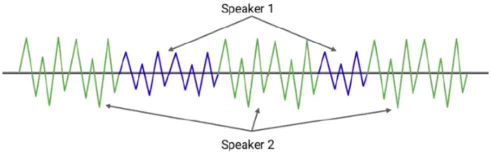
NSU, Department of Mathematics and Mechanics

Scientific adviser: E. N. Pavlovsky, PhD in Physics and Mathematics

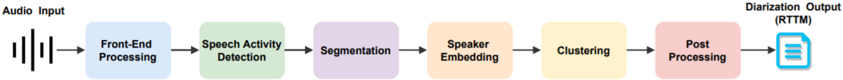
Novosibirsk
2021

What is diarization?

1. Diarization is the process of dividing the incoming stream into homogeneous segments in accordance with the belonging of the stream to one or another speaker.
2. Diaitarization answers the question "Who spoke when?"



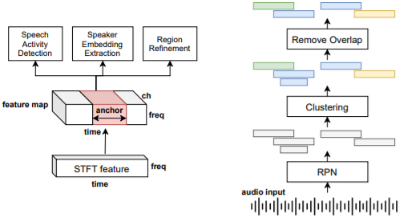
Recent Advantages. Modular speaker diarization systems



1. Deep Learning approaches(e. g. LSTM based)
2. Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs) or DNNs
3. Uniform segmentation
4. d-vectors
5. DOVER

Recent Advantages. Joint optimization for speaker diarization

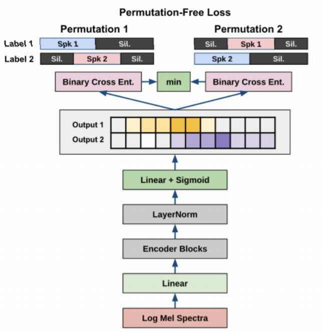
- **Joint segmentation and clustering.**A model called Unbounded Interleaved-State Recurrent Neural Networks (UIS-RNN) was proposed.
- **Joint segmentation, embedding extraction, and resegmentation.** Region Proposal Networks (RPN).



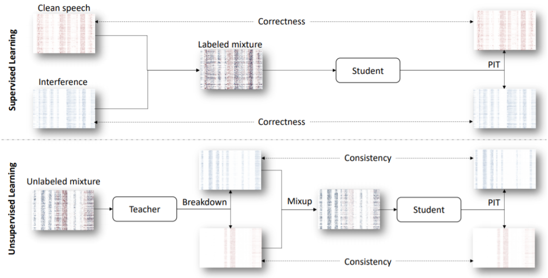
(a) Region proposal network (RPN) (b) Diarization by RPN

Recent Advantages. Joint optimization for speaker diarization

- **Joint speech separation and diarization.** Kounades-Bastian proposed to incorporate a speech activity model into speech separation based on the spatial covariance model with non-negative matrix factorization. Neumann later proposed a trainable model, called online Recurrent Selective Attention Network (online RSAN).
- **Fully end-to-end neural diarization.**



Mixup-Breakdown Training



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

The Mixup-Breakdown (MB) strategy trains a student model f_{θ_S} to provide consistent predictions with the teacher model f_{θ_T} of the same network structure at perturbations of predicted separations from the input mixtures (either labeled or unlabeled):

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

Mixup-Breakdown Training

The semi-supervised learning mode with a labeled dataset and an unlabeled dataset looks like this:

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

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

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

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

Goals and objectives

Goal: improve the quality of diarization with Mixup Breakdown Training.
Objectives:

1. implement the Mixup Breakdown Training
2. adapt the Mixup Breakdown Training to the diarization task
3. analyze and select a suitable backbone

Results

- Generated datasets: speaker + speaker, noise + speaker, 2 speakers + noise.
- MBT v0.1 implemented
- The model was trained with the Mixup-Breakdown algorithm and the ConvTasNet network as a student and a teacher model on two data sets: speaker + speaker, speaker + noise.

Further work

- Tests on AMI dataset: MBT, TasNet, SpectralCluster, compare results
- Consider options for replacing TasNet with newer models
- Genralize the results by 3, 4, etc. speaker
- Publish the code on GitHub