Motivation/ Setting

- Blind source separation
- Leverage Deep Attractor Network (DAN)
  - Problematic with long mixtures/ sessions
- Multiple microphones available
  \[ y_{tf} = \sum_k h_{fk} s_{tfk} + n_{tf} = \sum_k x_{tfk} + n_{tf} \]
  \( t \): time frame index
  \( f \): frequency bin index
- DAN generates embedding vectors \( e_{tf} \) indicative of which time frequency bin belongs to the same speaker.
- Embedding vectors \( e_{tf} \) can then be clustered.
- DAN not directly applicable to streaming data
  - Uses BLSTM
  - Split signal into blocks
  - Embedding space not fixed, centroid \( \mu_{fk} \) of each speaker changes from block to block (block permutation problem)
- Resort to spatial model (i.e. time variant complex GMM (TV-cGMM))
  - Independent solution on each frequency bin
  - Solutions not aligned (frequency permutation problem)

Problem statement

- DAN not directly applicable to streaming data
  - Uses BLSTM
  - Split signal into blocks
  - Embedding space not fixed, centroid \( \mu_{fk} \) of each speaker changes from block to block (block permutation problem)
- Resort to spatial model (i.e. time variant complex GMM (TV-cGMM))
  - Independent solution on each frequency bin
  - Solutions not aligned (frequency permutation problem)

Idea

- GeV beamforming
  - Source extraction by generalized eigenvalue decomposition of target and non-target covariance matrix:
  \[ \Phi_{fk}^{\text{target}} = \frac{1}{\Gamma_{fk}} \sum_{t \in T_k} \gamma_{tfk} y_{tf}^H, \quad \Phi_{fk}^{\text{non-target}} = \sum_{k' \neq k} \Phi_{fk}^{\text{target}} \]

Results

Obtain spatial covariance matrices with incremental update.

Block-online algorithm

1. Split into \( N \) blocks and run model on the first block.
2. Apply GEV beamforming to the first block.
3. for \( n \) from 1 to \( N \) do
   4. Forget all parameters but \( R_{n-1,k} \) and \( \Phi_{n-1,k} \).
   5. Initialize \( \sigma_{nk} \) with using \( R_{n-1,k} \).
   6. Initialize \( \gamma_{nk} \) only with spatial observation model.
   7. while not converged do
      8. Obtain \( \mu_{nk} \) and \( \Sigma_{nk} \).
      9. Incremental update for \( R_{nk} \).
     10. Calculate variance \( \sigma_{nk} \).
     11. E-step with permutation alignment yields \( \gamma_{nk} \) and \( \Pi_n \).
   12. Obtain spatial covariance matrices with incremental update.
   13. Apply beamforming on current block.

GEV beamforming

Source extraction by generalized eigenvalue decomposition of target and non-target covariance matrix:

$$
\Phi_{fk}^{\text{target}} = \frac{1}{\Gamma_{fk}} \sum_{t \in T_k} \gamma_{tfk} y_{tf}^H, \quad \Phi_{fk}^{\text{non-target}} = \sum_{k' \neq k} \Phi_{fk}^{\text{target}}
$$

Dual frequency- and block-permutation alignment for deep learning based block-online blind source separation

Lukas Drude\(^1\), Takuya Higuchi\(^2\), Keisuke Kinoshita\(^2\), Tomohiro Nakatani\(^2\), Reinhold Haeb-Umbach\(^1\)

\(^1\)Department of Communications Engineering, Paderborn University, Paderborn, Germany
\(^2\)NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan

ICASSP 2018, Calgary

{drude, haeb}@nt.upb.de, {higuchi.takuya, kinoshita.k, nakatani.tomohiro}@lab.ntt.co.jp