# Separation matrix optimization using associative memory model for blind source separation

Motoi Omachi, Tetsuji Ogawa and Tetsunori Kobayashi (Waseda University, Japan); Masaru Fujieda and Kazuhiro Katagiri (Oki Electric Industry Co., Ltd., Japan)

## Abstract

**Objective:**

- **Linear blind source separation (BSS)** yielding high quality and less distorted speech

**Approach:**

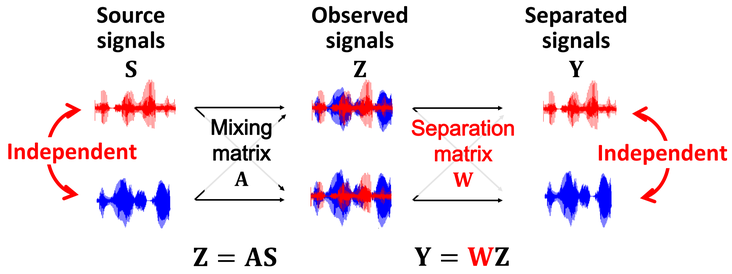
- Optimization of a linear separation matrix using **neural network-based associative memory model** (AMM)

**Result (Simultaneous speech separation):**

- Residual distortion caused by Independent vector Analysis (IVA) can be reduced.

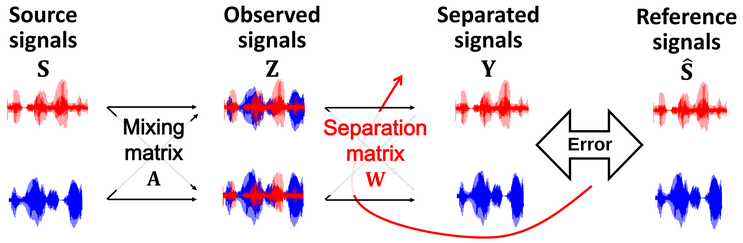
## Conventional linear BSS (e.g., ICA, IVA)

- Estimate linear separation matrix defeating effect of mixing matrix.
- Assume that source signals are statistically Independent.
- **Does not take account of property of source signals.**



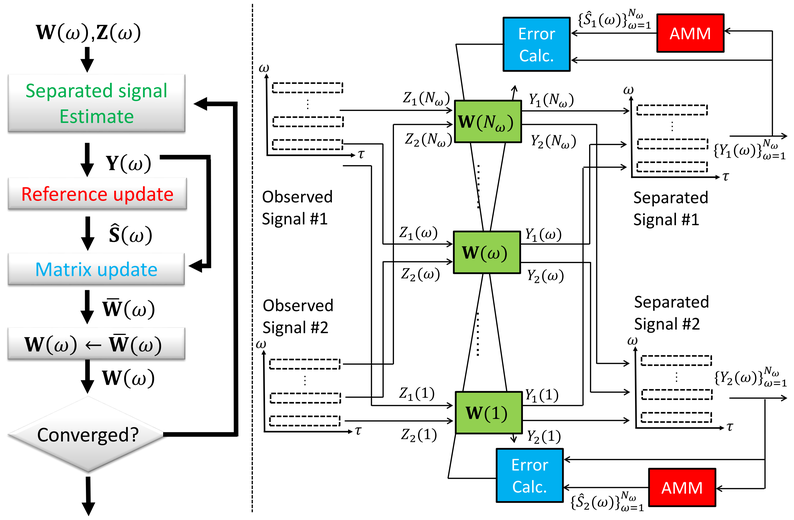

## Proposed linear BSS

- Separation matrix is estimated by iterating following two steps:
  - *Reference signal estimation*: Source signal (referred to ``reference signal'') is estimated from separated signal using AMM.
  - *Separation matrix optimization*: Separation matrix is optimized by minimizing error between separated signal and reference signal.



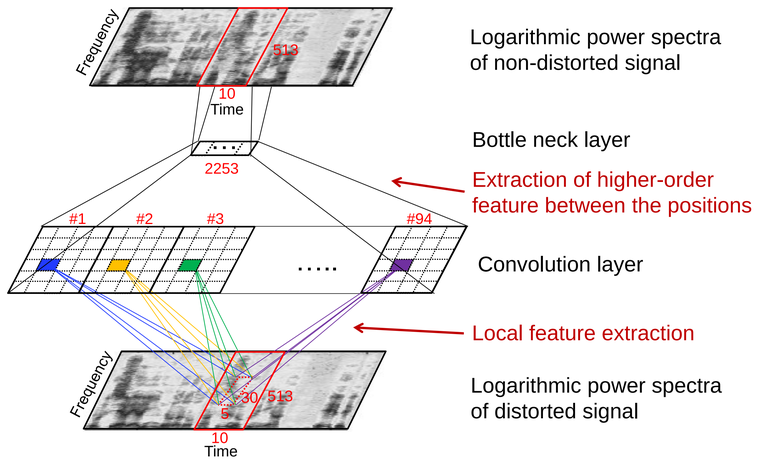

**Property of source signals is explicitly incorporated in separation matrix optimization.**

## Algorithm




### 1. *Reference signal estimation using AMM*

- **Convolutional neural network** can remove residual distortion included in spectra of separated signal.




### 2. *Separation Matrix optimization*

- Separation matrix is modified using a **linear projection matrix**.

$$\begin{bmatrix}\bar{Y}_1(\omega,\tau)\\ \bar{Y}_2(\omega,\tau)\end{bmatrix} = \begin{bmatrix}M_{11}(\omega) & M_{12}(\omega)\\ M_{21}(\omega) & M_{22}(\omega)\end{bmatrix}\begin{bmatrix}Y_1(\omega,\tau)\\ Y_2(\omega,\tau)\end{bmatrix}$$

Modified signal — Projection matrix — Separated signal (IVA)

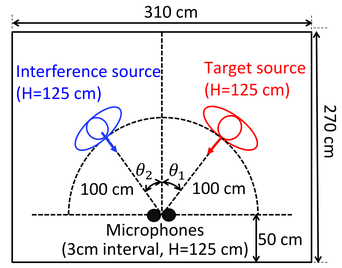
- Projection matrix is optimized with a gradient descend method.
  - $N_s, N_\tau$ and $N_m$: # of sources, frames and microphones

$$J(\omega) = \sum_{n=1}^{N_s}\sum_{\tau=1}^{N_\tau}\left|\log\left|\hat{S}_n(\omega,\tau)\right|^2 - \log\left|\sum_{j=1}^{N_m} M_{nj}Y_j(\omega,\tau)\right|^2\right|^2$$

Reference signal — Modified signal

## Source separation experiment

### *Acoustic Environment*




| Dataset | Source directions $(\theta_1, \theta_2)$ |
|---|---|
| Training | (-15,15),(-45,45), (-75,75),(-90,90) |
| Development | (-60,60) |
| Testing | (-30,30),(-30,0),(0,-30), (0,30),(30,0),(30,-30) |

### *Speech materials*

Training and Development set (for AMM training):

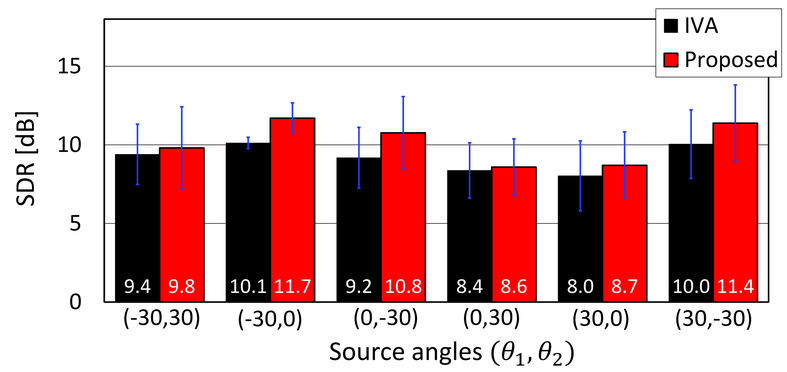
- Japanese phoneme balanced sentence database [4 females]
- Train. : 9 speaker pairs, 50 utterances for each pair
- Dev. : 1 speaker pair, 52 utterances
- Mixed signals are obtained by delay-based approximation (with SNR = 0dB).
- Dev. Set is used for early stopping

Test set (for BSS experiment)

- Japanese newspaper article sentence database [20 females]
- 10 speaker pairs, 3 utterances for each pair
- Mixed signals are obtained by convoluting impulse response (with SNR = 0dB).

### *Signal-to-distortion ratio (SDR)*

- Our method can **reduce residual distortions** caused by IVA




### *Phoneme error rate (DNN/ triphone + bigram)*

- Our method can **improve speech recognition** performance