Similarity Search-based Blind Source Separation

Hiroshi Sawada

Kazuo Aoyama

NTT Communication Science Laboratories,
NTT Corporation
Background & Motivation

- **BSS methods**
  - Independence: ICA, IVA
  - Low-Rankness: ILRMA
  - Needs enough amount of observations ($\geq 3$ sec.)

- **Time-varying Environments**
  - Short observations ($\leq 2$ sec.)

- **Similarity search** on a clean source database
  - Human can separate mixtures if there is something familiar to us in the mixtures
Supervised learning?

- Source Separation
- Selection Optimization
- Similarity search
  - Controllable
  - Interpretable
- DNN models
  - NMF dictionaries
- Clean source database
- Supervised training
- Does not need a time consuming training phase
- Does not need a time consuming training phase
- Controllable
- Interpretable

- Selection Optimization
1. Existing BSS methods
   - Frequency-domain BSS
   - IVA, ICA, ILRMA

2. Proposed method
   - SSBSS: Similarity Search-based BSS
   - Differs in variance parameter updates

3. Experiments
   - Clean source databases: close and open
   - Convergence behavior
   - Computational time with a GPU
Frequency-domain BSS

Separation

\[ \mathbf{y}_{ft} = \mathbf{W}_f \mathbf{x}_{ft} \quad f = 1, \ldots, F \]

\[ \mathbf{y}_{ft} = \begin{bmatrix} y_{ft1} \\ y_{ft2} \end{bmatrix} \quad \mathbf{W}_f = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \quad \mathbf{x}_{ft} = \begin{bmatrix} x_{ft1} \\ x_{ft2} \end{bmatrix} \]

Separated signal
Separation matrix
Mixture
Objective function

\[ J(\{W_f\}_{f=1}^F) = \sum_{t=1}^T \sum_{n=1}^N G(\tilde{y}_{tn}) - 2T \sum_{f=1}^F \log |\det W_f| \]

- \( f = 1, \ldots, F \) Frequency bins
- \( n = 1, \ldots, N \) Separated signals
- \( t = 1, \ldots, T \) Time frames

Contrast function \( G \)

\[
\begin{align*}
\text{IVA} & \quad G(\tilde{y}_{tn}) = \sum_{f=1}^F \left( \frac{|y_{ftn}|^2}{v_{tn}} + \log v_{tn} \right) \\
\text{ICA} & \quad G(\tilde{y}_{tn}) = \sum_{f=1}^F \left( \frac{|y_{ftn}|^2}{v_{ftn}} + \log v_{ftn} \right) \\
\text{ILRMA} & \quad G(\tilde{y}_{tn}) = \sum_{f=1}^F \left( \frac{|y_{ftn}|^2}{v_{ftn}} + \log v_{ftn} \right), \quad v_{ftn} = \sum_{k=1}^K b_{fnt} a_{tnk}
\end{align*}
\]

Local Gaussian Model

\[ p(\tilde{y}_{tn}) = \prod_{f=1}^F \frac{1}{\pi v_{tn}} \exp \left( -\frac{|y_{ftn}|^2}{v_{tn}} \right) \]

variance parameter \( v \)
Variance parameters $v$

- Time varying activity
- Flat spectrum

Permutation problem

Low-rank model well estimated
IVA Optimization

Variance update

\[ v_{tn} \leftarrow \frac{1}{F} \sum_{f=1}^{F} |y_{ftn}|^2 \]

\[ y_{ft} = W_f x_{ft} \]

Frequency-wise separation matrix update \( W_f \)

Weighted covariance matrix

\[ U_{fn} = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{v_{tn}} x_{ft} x_{ft}^H \]

Solve HEAD: Hybrid Exact-Approximate Diagonalization \([\text{Yeredor 2009}]\) \([\text{Ono 2011}]\)

\[ w_{fk}^H U_{fn} w_{fn} = \delta_{kn} \]

N=2 case

\[ w_1^H U_1 w_1 = 1 \quad w_1^H U_2 w_2 = 0 \]

\[ w_2^H U_1 w_1 = 0 \quad w_2^H U_2 w_2 = 1 \]

\[ w_{fn} \leftarrow (W_f U_{fn})^{-1} e_n \]

\[ w_{fn} \leftarrow \frac{w_{fn}}{\sqrt{w_{fn}^H U_{fn} w_{fn}}} \]

\( n = 1, \ldots, N \)
Variance update
\[ v_{ftn} \leftarrow |y_{ftn}|^2 \]
\[ y_{ft} = W_f x_{ft} \]

Frequency-wise separation matrix update
\[ W_f \]
Weighted covariance matrix
\[ U_{fn} = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{v_{tn}} x_{ft} x_{ft}^H \]

Solve HEAD: Hybrid Exact-Approximate Diagonalization
\[ w_{fk}^H U_{fn} w_{fn} = \delta_{kn} \]

[Ono 2011]

N=2 case
\[ w_1^H U_1 w_1 = 1 \]
\[ w_1^H U_2 w_2 = 0 \]
\[ w_2^H U_1 w_1 = 0 \]
\[ w_2^H U_2 w_2 = 1 \]

[Yeredor 2009]

\[ w_{fn} \leftarrow \frac{w_{fn}}{\sqrt{w_{fn}^H U_{fn} w_{fn}}} \]
\[ n = 1, \ldots, N \]
**ILRMA Optimization**

**Variance update**

\[ v_{ftn} \approx |y_{ftn}|^2 \]

\[ v_{ftn} \leftarrow \sum_{k=1}^{K} b_{fnk} a_{tnk} \]

**Frequency-wise separation matrix update**

\[ y_{ft} = W_f x_{ft} \]

\[ W_f \]

Weighted covariance matrix

\[ U_{fn} = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{v_{tn}} x_{ft} x_{ft}^H \]

**Solve HEAD: Hybrid Exact-Approximate Diagonalization**

\[ w_{fk}^H U_{fn} w_{fn} = \delta_{kn} \]  

[Yeredor 2009]

[Ono 2011]

N=2 case

\[ w_1^H U_1 w_1 = 1 \quad w_1^H U_2 w_2 = 0 \]

\[ w_2^H U_1 w_1 = 0 \quad w_2^H U_2 w_2 = 1 \]

\[ w_{fn} \leftarrow (W_f U_{fn})^{-1} e_n \]

\[ w_{fn} \leftarrow \frac{w_{fn}}{\sqrt{w_{fn}^H U_{fn} w_{fn}}} \]

\[ n = 1, \ldots, N \]
Outline

1. Existing BSS methods
   - Frequency-domain BSS
   - IVA, ICA, ILRMA

2. Proposed method
   - SSBSS: Similarity Search-based BSS
     - Differs in variance parameter updates

3. Experiments
   - Clean source databases: close and open
   - Convergence behavior
   - Computational time with a GPU
SSBSS Optimization

**Variance update**

- Similarity search
- Clean source database

\[ y_{ft} = W_f x_{ft} \]

**Frequency-wise separation matrix update**

\[ W_f \]

- Weighted covariance matrix

\[ U_{fn} = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\nu_{tn}} x_{ft} x_{ft}^H \]

**Solve HEAD: Hybrid Exact-Approximate Diagonalization**

- [Yeredor 2009]
- [Ono 2011]

\[ w_{fk}^H U_{fn} w_{fn} = \delta_{kn} \]

N=2 case

\[ w_1^H U_1 w_1 = 1 \quad w_1^H U_2 w_2 = 0 \]

\[ w_2^H U_1 w_1 = 0 \quad w_2^H U_2 w_2 = 1 \]

\[ w_{fn} \leftarrow \frac{w_{fn}}{\sqrt{w_{fn}^H U_{fn} w_{fn}}} \]

\[ n = 1, \ldots, N \]
Clean source database

**Database**

\[ \mathcal{S} = \{ \{ \mathbf{s}_{tl} \}^{T_l}_{t=1} \}^{L}_{l=1} \]

**Entry:** $F$-dimensional power spectra vector

\[ \mathbf{s}_{tl} = [ |s_{1tl}|^2, \ldots, |s_{Ftl}|^2]^T \quad s_{ftl} \in \mathbb{C} \]

- $f = F$
- $f = 1$
- $l = 1$
- $l = 2$
- $l = L$

$L$ source sound files
Objective function (same structure)

\[
J(\{W_f\}_{f=1}^F) = \sum_{t=1}^T \sum_{n=1}^N G(\tilde{y}_{tn}) - 2T \sum_{f=1}^F \log |\text{det} W_f|
\]

Contrast function \(G\)

\[
G(\tilde{y}_{tn}) = \sum_{f=1}^F \left( \frac{|y_{ftn}|^2}{v_{ftn}} + \log v_{ftn} \right)
\]

Variance \(v\) constrained

\(F\)-dimensional vector \(y_{tn} = [v_{1tn}, \ldots, v_{Ftn}]^T\)

\(\gamma S\) with \(\gamma\) arbitral scale

Clean source database
Variance update by similarity search

\[ f = F \]

\( n \)-th separated signal

Variance update

\[ \mathbf{v}_{tn} \leftarrow \gamma \mathbf{S}_* \]

adjusted scale

\[ \gamma = \frac{1}{F} \sum_{f=1}^{F} \frac{|y_{ftn}|^2}{|s_f|^2} \cdot \]

Query: power spectra vector

\[ \tilde{\mathbf{y}}_{tn} = [|y_{1tn}|^2, \ldots, |y_{Ftn}|^2]^T \]

Measure: Itakura-Saito divergence

\[ D_{IS}(\tilde{\mathbf{y}}_{tn}, \gamma \mathbf{S}) = \sum_{f=1}^{F} \left( \frac{|y_{ftn}|^2}{\gamma |s_f|^2} - \log \frac{|y_{ftn}|^2}{\gamma |s_f|^2} - 1 \right) \]

\[ f = F \]

clean source database

\[ f = 1 \]
SSBSS Optimization

Variance update

\[ y_{ft} = W_f x_{ft} \]

Frequency-wise separation matrix update

\[ U_{fn} = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\nu_{tn}} x_{ft} x_{ft}^H \]

Weighted covariance matrix

Solve HEAD: Hybrid Exact-Approximate Diagonalization

\[ w_{fk}^H U_{fn} W_{fn} = \delta_{kn} \]

[Yeredor 2009]

[Ono 2011]

N=2 case

\[ w_1^H U_1 w_1 = 1 \quad w_1^H U_2 w_2 = 0 \]

\[ w_2^H U_1 w_1 = 0 \quad w_2^H U_2 w_2 = 1 \]
Outline

1. Existing BSS methods
   - Frequency-domain BSS
   - IVA, ICA, ILRMA

2. Proposed method
   - SSBSS: Similarity Search-based BSS
   - Differs in variance parameter updates

3. Experiments
   - Clean source databases: close and open
   - Convergence behavior
   - Computational time with a GPU
Experimental conditions

- **Sources**: 2-second speeches

- **Mixtures**: 32 cases
  Various combinations of 2-second speech signals

- **Distance**: 120 cm

- **Loudspeakers**

- **Microphones**: 4 cm apart

- **Room size**: $4.45 \times 3.55 \times 2.5$ m

- **Height of microphones and loudspeakers**: 120 cm

- **Reverberation time**
  $RT_{60} = 200$ ms

- **Sources**: 2-second speeches

- **Mixtures**: 32 cases
  Various combinations of 2-second speech signals
Clean source databases

F = 1025

# database entries was around 30,000

- close
  - contained the sources used for mixtures
  - ideal situation for verifying the basic concept

- open
  - did not contain the source time frames used for mixtures
    - but contained the same speaker’s different utterances
  - In some settings, new entries were added aiming for better performance
Separation performance

The higher the better

- Each dotted line corresponds to a mixture case
- Solid grey line represents the average of 32 mixture cases

Did not perform well. Slightly improved as adding new entries.

Performed very well (ideal situation)
Variances & separated signals

Mixture

SSBSS (close)

ILRMA(1), SSBSS (open+c)

Frequency (kHz)

Time (sec)

0.5 1 1.5

Frequency (kHz)

Time (sec)

0.5 1 1.5

7.16 dB

9.79 dB

3.85 dB

6.45 dB
**Execution time**
- 20 seconds for 30 iterations and 2-second mixture

**Similarity search executed on a GPU**
- 158 queries (2 outputs $\times$ 79 time frames) for 30,000 entries with $F=1025$ dim. took around 230 ms.
Conclusion

- **Proposed SSBSS**
  - Searches clean database $\mathcal{S}$ for similar entries to $\tilde{y}$
  - Updates variance parameters $\mathbf{v}$ with the result

- **Experimental results**
  - Short observation of 2 seconds
  - High performance with ideal close database
  - Open database lowered the performance

- **Future work**
  - Constructing better databases for open cases
  - Accelerating the search to handle larger databases