Blind separation of infinitely many sparse sources

Hirokazu Kameoka[1,2], Misa Sato[1], Takuma Ono[1], Nobutaka Ono[3], Shigeki Sagayama[1]

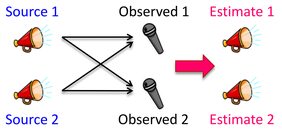
[1] The University of Tokyo, [2] NTT Communication Science Laboratories, [3] National Institute of Informatics

Abstract: This paper presents a novel BSS approach that simultaneously performs an estimation of the number of sources, source separation based on the sparseness of speech, and permutation alignment, based on Bayesian nonparametric approach.

1. Introduction

Blind source separation (BSS)

- Technique for separating sources only from microphone inputs
- Potential applications include hands-free teleconference system and automatic meeting transcription system



Motivation and objective

- It is often difficult to pre-specify the exact number of all possible sources present in real environments.

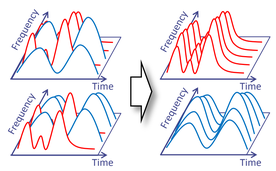
e.g.) In meeting situations, # of speakers can change during the meeting or loud, unexpected noise such as door slamming can occur.

→ When # of sources is unknown, we shall always consider underdetermined case (# of microphones < # of sources)

- One successful approach for underdetermined BSS involves utilizing "*sparseness" of speech [1]
  * T-F components of speech are near zero across most of T-F bins
- To exploit the sparseness of speech, mixing model must be represented in T-F domain

→ Permutation alignment problem needs to be solved

⇨ We propose BSS approach that simultaneously performs
(1) estimation of # of sources,
(2) source separation based on sparseness of speech, and
(3) permutation alignment.



2. Mixing Model

Instantaneous mixture in frequency domain:

(Narrowband approximation of convolutive mixture in time domain)

$$\boldsymbol{y}_{\omega,t} = \boldsymbol{A}_\omega \boldsymbol{s}_{\omega,t} = \sum_{k=1}^{K} \boldsymbol{a}_{k,\omega} s_{k,\omega,t} + \boldsymbol{n}_{\omega,t}$$

$\boldsymbol{y}_{\omega,t} := (y_{1,\omega,t}, \ldots, y_{M,\omega,t})^{\mathsf{T}}$

$y_{m,\omega,t}$: T-F component observed at microphone m

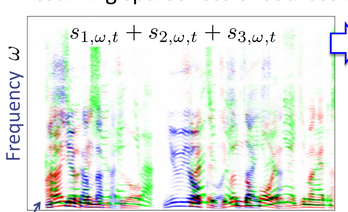
$s_{k,\omega,t}$: T-F component of source k

$\boldsymbol{a}_{k,\omega}$: Steering vector of source k

$\boldsymbol{n}_{\omega,t}$: T-F component of noise

K (# of sources) $>$ M (# of microphones)

Assuming sparseness of sources in T-F domain



Superposition of 3 speech spectrograms

Assumption: only one source is active in each T-F bin

$z_{\omega,t} \in \{1, \ldots, K\}$:
Index of source active at T-F bin (ω, t)

$$s_{k,\omega,t} = \begin{cases} \hat{s}_{\omega,t} & (k = z_{\omega,t}) \\ 0 & (k \neq z_{\omega,t}) \end{cases}$$

$$\boldsymbol{y}_{\omega,t} = \boldsymbol{a}_{z_{\omega,t},\omega} \hat{s}_{\omega,t} + \boldsymbol{n}_{\omega,t}$$

3. Generative Model

Generative process of observed signal

- Gaussian noise assumption: $\boldsymbol{n}_{\omega,t} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \Sigma_\omega^{(n)})$

⇨ $\boldsymbol{y}_{\omega,t} | \boldsymbol{a}_{1:K,\omega}, z_{\omega,t}, \hat{s}_{\omega,t} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{z_{\omega,t},\omega} \hat{s}_{\omega,t}, \Sigma_\omega^{(n)})$ … (1)

Mixture of infinitely many sparse sources

- Finite case: $\boldsymbol{y}_{\omega,t} | \boldsymbol{a}_{1:K,\omega}, z_{\omega,t}, \hat{s}_{\omega,t} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{z_{\omega,t},\omega} \hat{s}_{\omega,t}, \Sigma_\omega^{(n)})$

$K \to \infty$

$z_{\omega,t} | \boldsymbol{\pi} \sim \mathrm{Discrete}(z_{\omega,t}; \boldsymbol{\pi})$

$\boldsymbol{\pi} \sim \mathrm{Dirichlet}(\boldsymbol{\pi}; \alpha_0/K, \ldots, \alpha_0/K)$

- Infinite case: $\boldsymbol{y}_{\omega,t} | \boldsymbol{a}_{1:\infty,\omega}, z_{\omega,t}, \hat{s}_{\omega,t} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{z_{\omega,t},\omega} \hat{s}_{\omega,t}, \Sigma_\omega^{(n)})$

$z_{\omega,t} | \boldsymbol{\pi} \sim \mathrm{Discrete}(z_{\omega,t}; \boldsymbol{\pi})$

$\boldsymbol{\pi} \sim \mathrm{GEM}(\boldsymbol{\pi}; \alpha_0)$ … (2)

Process by which active source index is generated (π_k indicates how likely source index k is to be chosen)

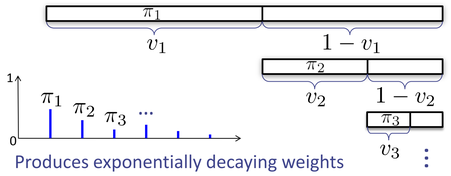
for $k = 1, \ldots, \infty$

$v_k \sim \mathrm{Beta}(v_k; 1, \alpha_0)$

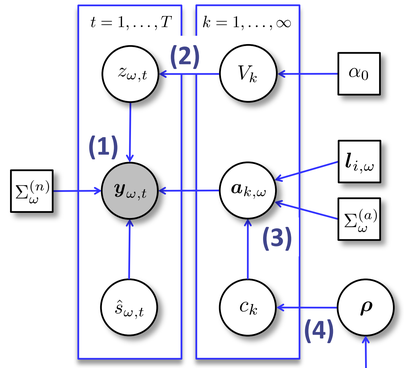
$\pi_k = v_k \prod_{k'=1}^{k-1} (1 - v_k)$

→ Stick-breaking process [3]

a.k.a. constructive definition of Dirichlet process [4]



Produces exponentially decaying weights



Mixture of Direction-of-Arrivals (DOAs)

- If index k indicates identical source across ω, $\boldsymbol{a}_{k,\omega}$ must have certain structure based on the property of acoustic wave propagation.

⇨ Incorporating appropriate constraint into $\boldsymbol{a}_{k,\omega}$ may help solve permutation alignment problem through parameter inference.

Assumption: Each source propagates as a plane wave and emanates from one of I directions.

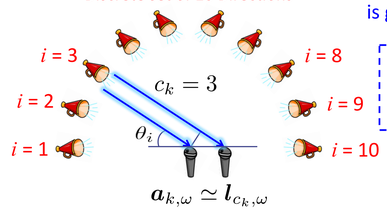
$c_k \in \{1, \ldots, I\}$: Direction index of source k

$\boldsymbol{l}_{i,\omega}$: Complex array response associated with direction i (**constant value**)

$\boldsymbol{a}_{k,\omega} | c_k \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{k,\omega}; \boldsymbol{l}_{c_k,\omega}, \Sigma_\omega^{(a)})$ … (3)

Process by which complex array response is generated given direction c_k

Discrete set of 10 Directions



$c_k | \boldsymbol{\rho} \sim \mathrm{Discrete}(c_k; \boldsymbol{\rho})$ … (4)

$\boldsymbol{\rho} \sim \mathrm{Dirichlet}(\boldsymbol{\rho}; \beta_0/I, \ldots, \beta_0/I)$

Process by which direction index is generated (ρ_i indicates how likely direction i is to be chosen)

Probability distributions

$\mathcal{N}_{\mathbb{C}}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) \propto e^{-(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{H}} \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$

$\mathrm{Beta}(y; z_1, z_2) \propto y^{z_1}(1-y)^{z_2}$

$\mathrm{Discrete}(x; \boldsymbol{y}) = y_x \quad (\sum_n y_n = 1)$

$\mathrm{Dirichlet}(\boldsymbol{y}; z_1, \ldots, z_N) \propto \prod_{n=1}^{N} y_n^{z_n - 1}$

4. Posterior Inference

Variational inference

- Goal is to compute $p(A, S, Z, V, C, \rho | Y)$ (Θ)
- From (1)-(4) :

$p(Y, A, S, Z, V, C, \rho)$

$= \underbrace{p(Y|A,S,Z)}_{(1)} \underbrace{p(Z|V)}_{(2)} \underbrace{p(V)}_{(3)} \underbrace{p(A|C)}_{(3)} \underbrace{p(C|\rho)}_{(4)} \underbrace{p(\rho)}_{(4)}$

To obtain exact posterior, we must derive marginal distribution $p(Y)$, but this involves intractable integrals.

$Y = \{\boldsymbol{y}_{\omega,t}\}_{1 \le \omega \le \Omega, 1 \le t \le T}$

$A = \{\boldsymbol{a}_{k,\omega}\}_{1 \le k \le \infty, 1 \le \omega \le \Omega}$

$S = \{\hat{s}_{\omega,t}\}_{1 \le \omega \le \Omega, 1 \le t \le T}$

$Z = \{z_{\omega,t}\}_{1 \le \omega \le \Omega, 1 \le t \le T}$

$V = \{v_k\}_{1 \le k \le \infty}$

$C = \{c_k\}_{1 \le k \le \infty}$

$\rho = (\rho_1, \ldots, \rho_I)$

- Instead, we consider finding $\hat{q}(\Theta) = \underset{q \in \mathcal{Q}}{\mathrm{argmin}}\, \mathrm{KL}(q(\Theta) \| p(\Theta|Y))$

Tractable choices of $\mathcal{Q}$ allows us to obtain principled approximate solutions

Kullback-Leibler divergence between $q(\Theta)$ and $p(\Theta|Y)$

- Mean-field approximation $q(\Theta) = q(A)q(S)q(Z)q(V)q(C)q(\rho)$
- Truncated approximation $q(v_{K^*} = 1) = 1$ ⇨ $\pi_k = 0 \quad (k > K^*)$

(Thus, we can ignore $\boldsymbol{a}_{k,\omega}, c_k$ for $k > K^*$)

Truncating q does not mean that true posterior itself is truncated. As truncation level K^* becomes larger, approximation to true posterior become more accurate.

Coordinate ascent

$\hat{q}(A) = \prod_{k,\omega} \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{k,\omega}; \boldsymbol{m}_{k,\omega}, \Gamma_{k,\omega})$

$\hat{q}(S) = \prod_{\omega,t} \mathcal{N}_{\mathbb{C}}(\hat{s}_{\omega,t}; \mu_{\omega,t}, \sigma^2_{\omega,t})$

$\hat{q}(Z) = \prod_{\omega,t} \mathrm{Discrete}(z_{\omega,t}; \boldsymbol{\phi}_{\omega,t})$

$\hat{q}(V) = \prod_{k} \mathrm{Beta}(v_k; \gamma_{k,0}, \gamma_{k,1})$

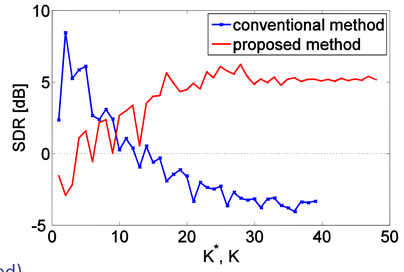
$\hat{q}(C) = \prod_{k} \mathrm{Discrete}(c_k; \boldsymbol{\psi}_k)$

$\hat{q}(\boldsymbol{\rho}) = \mathrm{Dirichlet}(\boldsymbol{\rho}; \zeta_1, \ldots, \zeta_I)$

5. Experiment

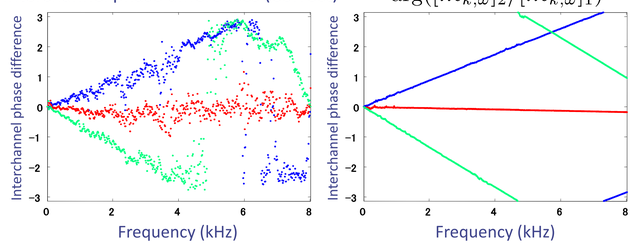
Experimental condition

- Speech data: excerpted from ATR Japanese speech database
- Observed signals: mixed three source signals (two female & one male speakers) into two mixtures using measured impulse response (reverberation time 0ms)
- T-F components were obtained by short-time Fourier transform (Frame length 64 ms, Hamming window, overlap length 16ms)
- Three sources were spaced 30 degrees apart
- Comparison: Sawada's method [5] (# of sources must be pre-specified)
- # of iterations = 100, # of DOA candidates I = 180
- Separated signals = $\mu_{\omega,t} \phi_{k,\omega,t}$



K^*: Truncation level (Proposed)
K : Assumed # of sources (Conventional)

Interchannel phase differences (Correct) $\arg([\boldsymbol{m}_{k,\omega}]_2 / [\boldsymbol{m}_{k,\omega}]_1)$



Results

- Average Signal-to-Distortion Ratios (SDRs) [6] obtained with Sawada's and proposed methods with various K and K^* settings (See upper-right)
- Proposed method improves with increasing K^* (implying that there is no need to pre-specify # of sources)

References

[1] O. Yılmaz, S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.
[2] Y. Izumi, N. Ono, S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in *Proc. WASPAA*, pp. 147–150, 2007.
[3] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.
[4] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.
[5] H. Sawada, S. Araki, S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. ASLP*, vol. 19, no. 3, pp. 516–527, 2010.
[6] E. Vincent, R. Gribonval, C. F´evotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, pp. 1462–1469, 2006.