31/Mar/2012

International Workshop on Statistical Machine Learning for Speech Processing
-Scalable Approach in the Era of Abundant Data-

Independent Component Analysis and Unsupervised Learning

國立交通大學
National Chiao Tung University

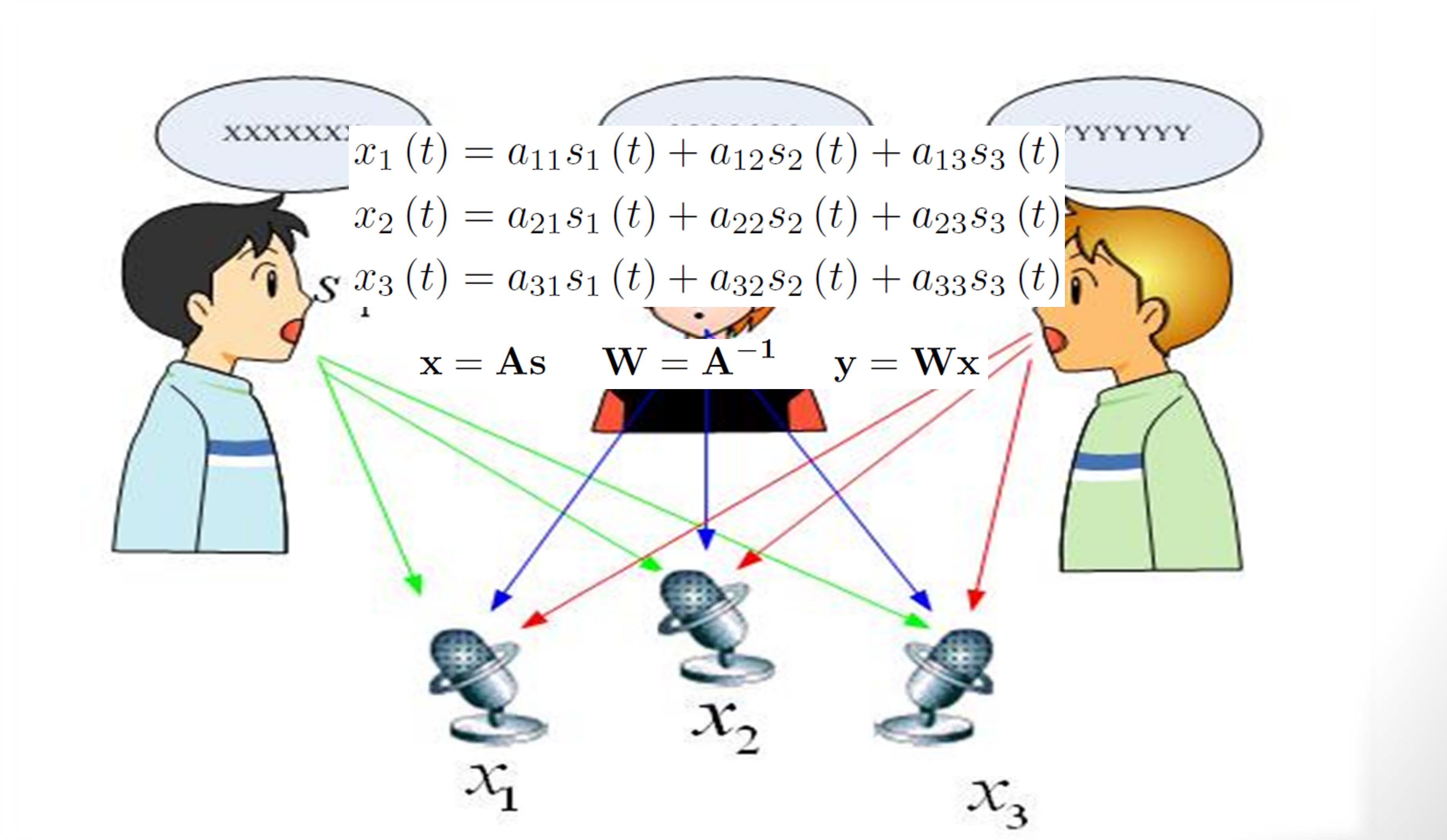
Jen-Tzung Chien

TABLE OF CONTENTS

Introduction

- Independent component analysis (ICA) is essential for blind source separation.
- ICA is applied to separate the *mixed signals* and find the *independent components*.
- The demixed components can be grouped into *clusters* where the *intra-cluster* elements are *dependent* and *inter-cluster* elements are *independent*.
- ICA provides *unsupervised learning* approach to acoustic modeling, signal separation and many others.

Blind Source Separation



$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) + a_{13}s_3(t)$$
$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) + a_{23}s_3(t)$$
$$x_3(t) = a_{31}s_1(t) + a_{32}s_2(t) + a_{33}s_3(t)$$

$$\mathbf{x} = \mathbf{A}\mathbf{s} \qquad \mathbf{W} = \mathbf{A}^{-1} \qquad \mathbf{y} = \mathbf{W}\mathbf{x}$$

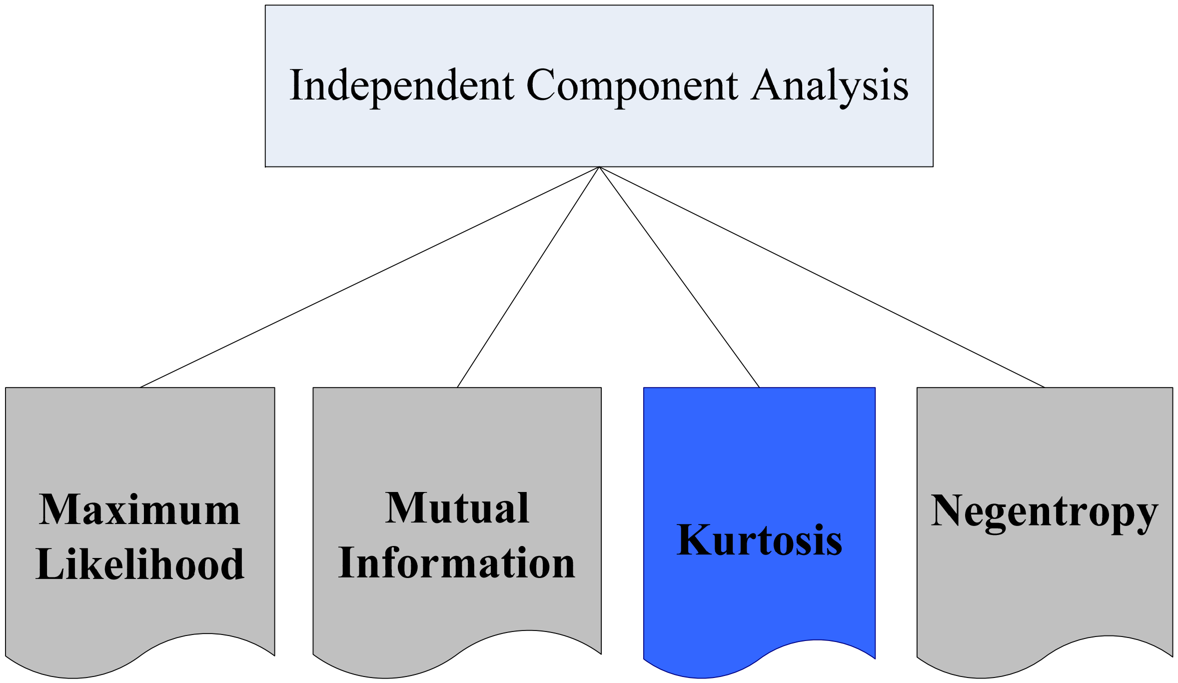
Independent Component Analysis

- Three assumptions
  - sources ⇨ statistically independent
  - independent component ⇨ nongaussian distribution
  - mixing matrix ⇨ square matrix

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

$$\underbrace{\begin{bmatrix} X_{11} & \cdots & X_{1t} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mt} \end{bmatrix}}_{m \times t} = \underbrace{\begin{bmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mm} \end{bmatrix}}_{m \times m} \underbrace{\begin{bmatrix} S_{11} & \cdots & S_{1t} \\ \vdots & \ddots & \vdots \\ S_{m1} & \cdots & S_{mt} \end{bmatrix}}_{m \times t}$$

ICA Objective Function



ICA Learning Rule

- ICA demixing matrix can be estimated by optimizing an *objective function* $D(\mathbf{X}, \mathbf{W})$ via *gradient descent* algorithm or *natural gradient* algorithm

$$\mathbf{W}^{(n+1)} = \mathbf{W}^{(n)} - \eta \frac{\partial D\left(\mathbf{X}, \mathbf{W}^{(n)}\right)}{\partial \mathbf{W}^{(n)}}$$

$$\mathbf{W}^{(n+1)} = \mathbf{W}^{(n)} - \eta \frac{\partial D\left(\mathbf{X}, \mathbf{W}^{(n)}\right)}{\partial \mathbf{W}^{(n)}} \mathbf{W}^{(n)\mathrm{T}} \mathbf{W}^{(n)}$$

TABLE OF CONTENTS

ICA for Speech Recognition

- Mismatch between training and test data always exists. *Adaptation* of HMM parameters is important.
- *Eigenvoice* (PCA) versus *Independent Voice* (ICA)
  - PCA performs a linear de-correlation process
  - ICA extracts the higher-order statistics

$E[e_1 e_2 \cdots e_M] = E[e_1]\,E[e_2] \cdots E[e_M]$	uncorrelation ⇨ PCA
$E[s_1^r s_2^r \cdots s_M^r] = E[s_1^r]\,E[s_2^r] \cdots E[s_M^r]$	higher-order correlations are zero ⇨ ICA

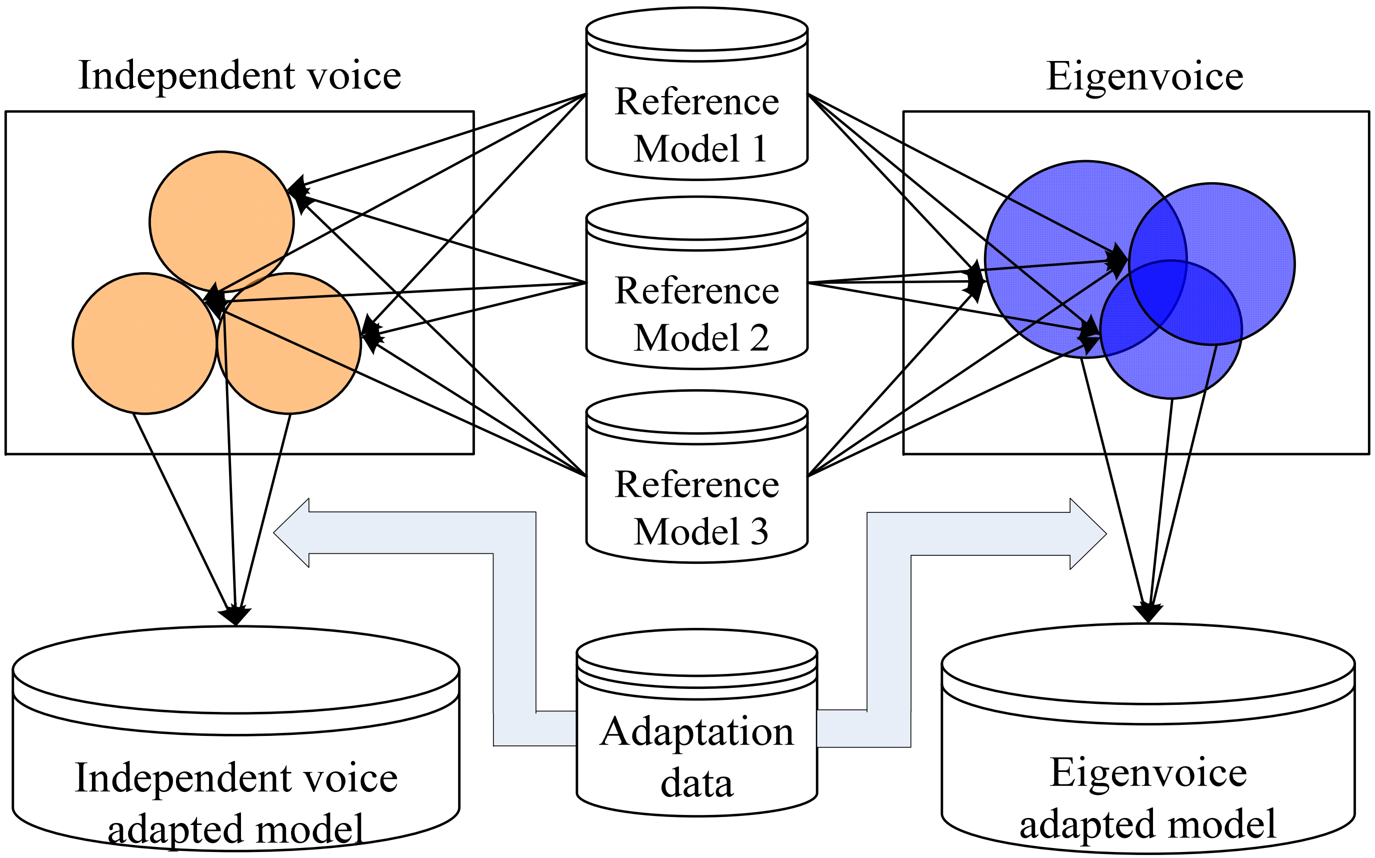
Sparseness & Information Redundancy

- The degree of *sparseness* in distribution of the transformed signals is proportional to the amount of *information* conveyed by the transformation.
- Sparseness measurement
  - fourth-order statistics (*kurtosis*) ⇨ nongaussianity

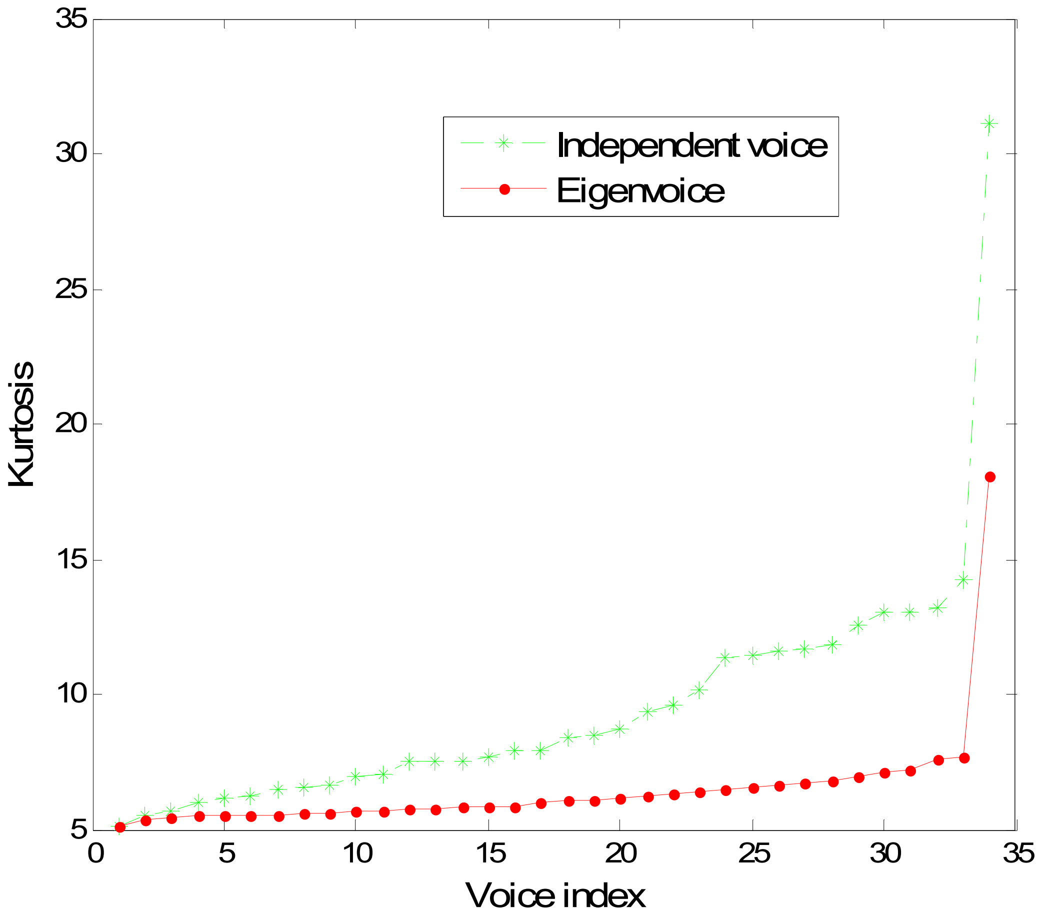
$$\mathrm{kurt}(s) = E[s^4] / E^2[s^2] - 3$$

- Information redundancy reduction using ICA is higher than that using PCA.

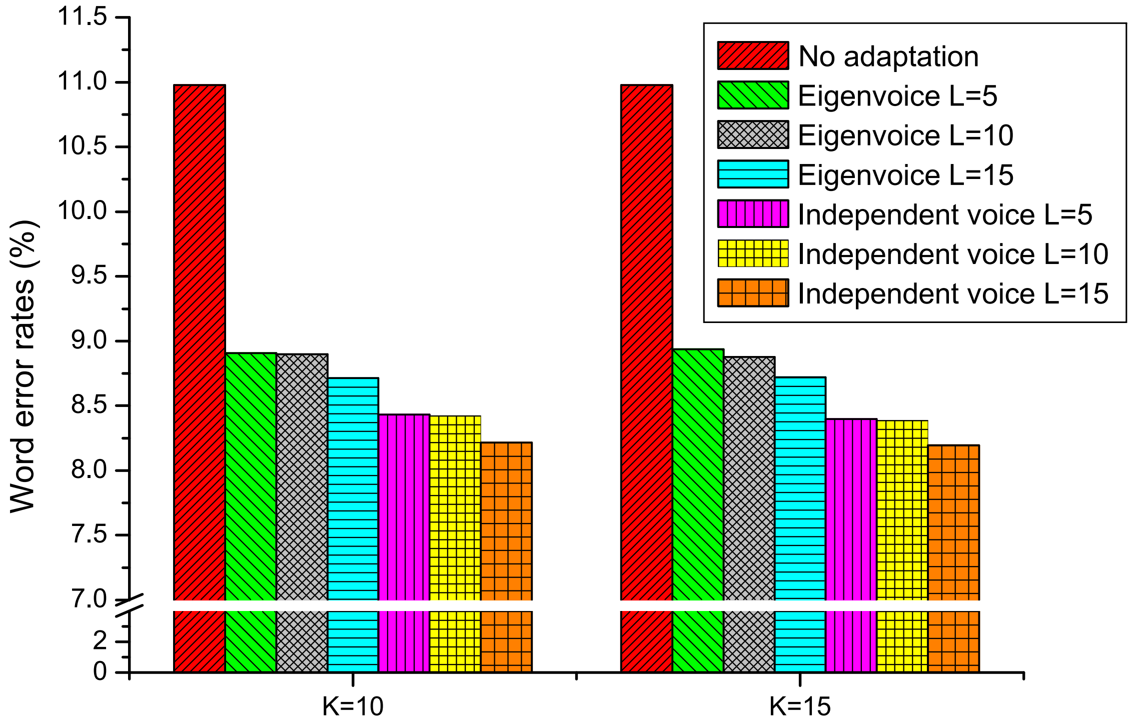
Eigenvoices versus Independent Voices



Evaluation of Kurtosis



Word Error Rates on Aurora2



K: number of components L: number of adaptation sentences

TABLE OF CONTENTS

Test of Independence

- Given the demixing signals $\mathbf{y} = \mathbf{W}\mathbf{x}$, the null & alternative hypotheses are defined as

$$H_0 : y^{(1)}, y^{(2)}, \ldots, y^{(d)} \text{ are mutually independent}$$
$$H_1 : y^{(1)}, y^{(2)}, \ldots, y^{(d)} \text{ are not mutually independent.}$$

- If $\mathbf{y}$ is *Gaussian* distributed, we are testing whether the correlation between $y^{(i)}$ and $y^{(j)}$ is equal to zero, i.e. $\Sigma_D = \mathrm{diag}\{\sigma_{ii}^2\}$ or

$$H_0 : \sigma_{ij}^2 = E\left\{\left(y^{(i)} - \mu^{(i)}\right)\left(y^{(j)} - \mu^{(j)}\right)\right\} = 0, \quad \text{for all } i \neq j.$$

Likelihood Ratio

$$\lambda_{\mathrm{LR}} = \frac{p(Y|H_0)}{p(Y|H_1)} = \frac{\max_{\boldsymbol{\mu},\Sigma_D} p(Y|\boldsymbol{\mu},\Sigma_D)}{\max_{\boldsymbol{\mu},\Sigma} p(Y|\boldsymbol{\mu},\Sigma)}$$

- LR serves as the *test statistics* which measures the confidence for H_0 against H_1 .
- LR is a measure of *independence* for $Y = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}$ and can act as an objective function for finding ICA demixing matrix.
- However, it is *not allowed* to assume Gaussianity for ICA problem.

Nonparametric Approach

- Let each sample be transformed by $\mathbf{y}_t = W\mathbf{x}_t$.
- Instead of assuming Gaussianity, we apply the *kernel density estimation*

$$p\left(y^{(i)}\right) = \frac{1}{Th}\sum_{t=1}^{T}\varphi\left(\frac{y^{(i)} - y_t^{(i)}}{h}\right), \quad i = 1, \ldots, d$$

using Gaussian kernel

$$\varphi(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$$

- Kernel centroid is given by $y_t^{(i)} = \mathbf{w}_i\mathbf{x}_t = \sum_{j=1}^{d} w_{ij}x_t^{(j)}$

Nonparametric Likelihood Ratio

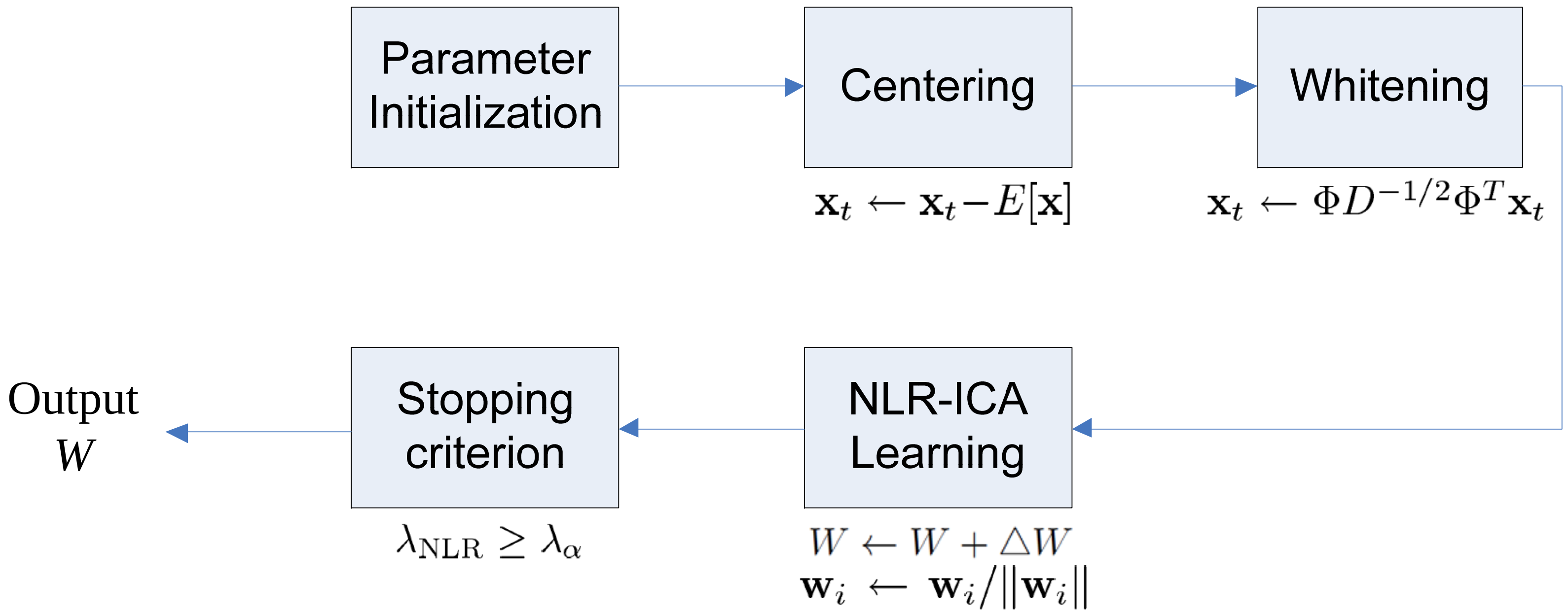
- NLR objective function

$$\lambda_{\mathrm{NLR}} = \frac{p(Y|H_0)}{p(Y|H_1)} = \frac{\prod_{t=1}^{T}\prod_{i=1}^{d} p\left(y_k^{(i)}\right)}{\prod_{t=1}^{T} p\left(\mathbf{y}_k\right)}$$

$$= \frac{\prod_{t=1}^{T}\prod_{i=1}^{d}\left[\frac{1}{Th}\sum_{k=1}^{T}\varphi\left(\frac{y_t^{(i)}-y_k^{(i)}}{h}\right)\right]}{\prod_{t=1}^{T}\left[\frac{1}{Th^d}\sum_{k=1}^{T}\psi\left(\frac{\mathbf{y}_t-\mathbf{y}_k}{h}\right)\right]}$$

with multivariate Gaussian kernel $\psi(\mathbf{v}) = \frac{1}{\sqrt{(2\pi)^d}}e^{-\frac{1}{2}\mathbf{v}^T\mathbf{v}}$

ICA Learning Procedure

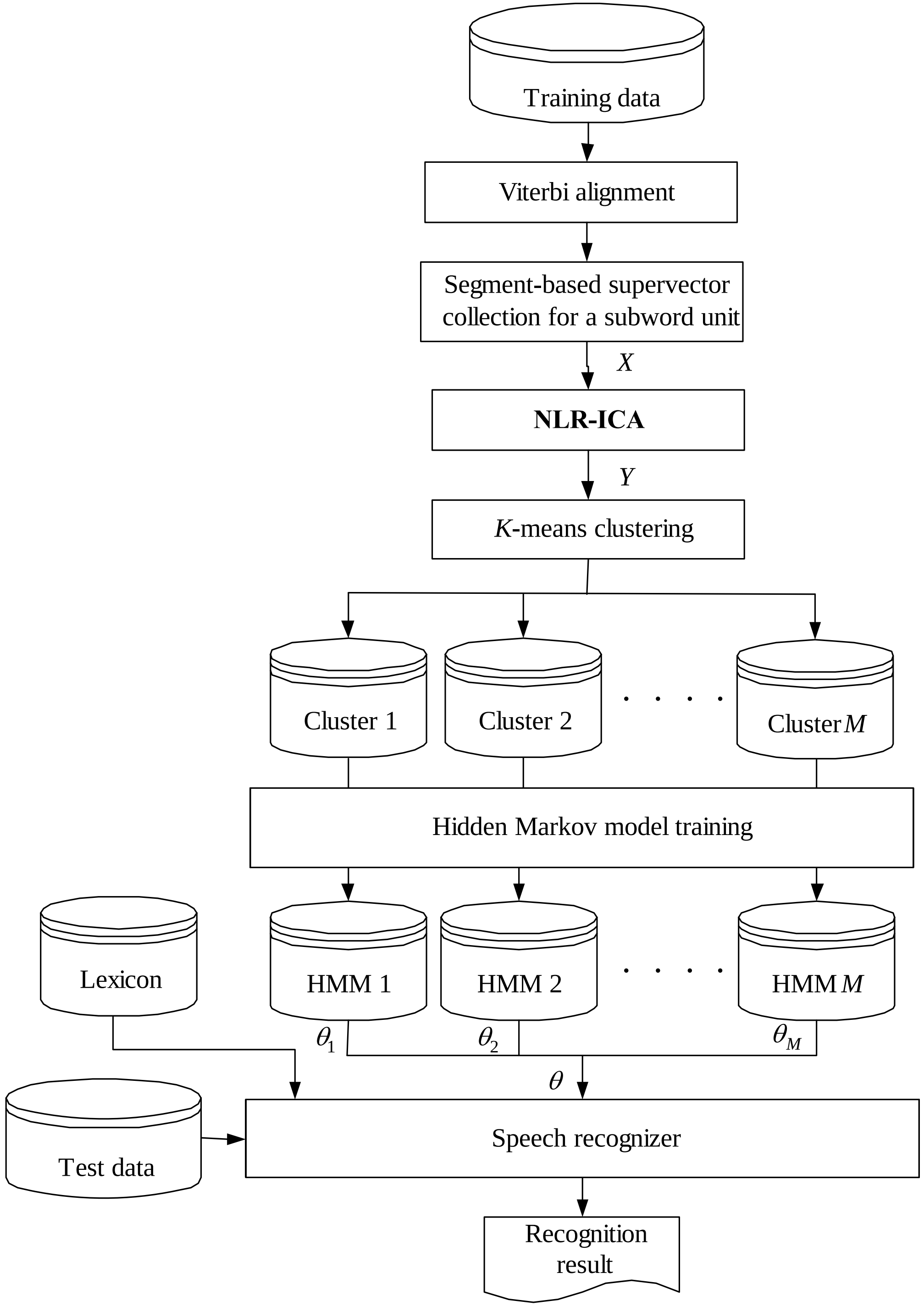


- Log likelihood ratio for null and alternative hypotheses

$$\log \lambda_{\mathrm{NLR}} = L_0(W) - L_1(W)$$

- Maximizing $\log \lambda_{\mathrm{NLR}}$ with respect to W_{NLR}, we obtain

$$\Delta W = -\eta \left(\nabla_W L_0(W) - \nabla_W L_1(W) \right)$$



Training data
Viterbi alignment
Segment-based supervector collection for a subword unit
X
NLR-ICA
Y
K-means clustering
Cluster 1
Cluster 2
. . . .
Cluster M
Hidden Markov model training
Lexicon
HMM 1
HMM 2
. . . .
HMM M
θ1
θ2
θM
θ
Test data
Speech recognizer
Recognition result

Segment-Based Supervector

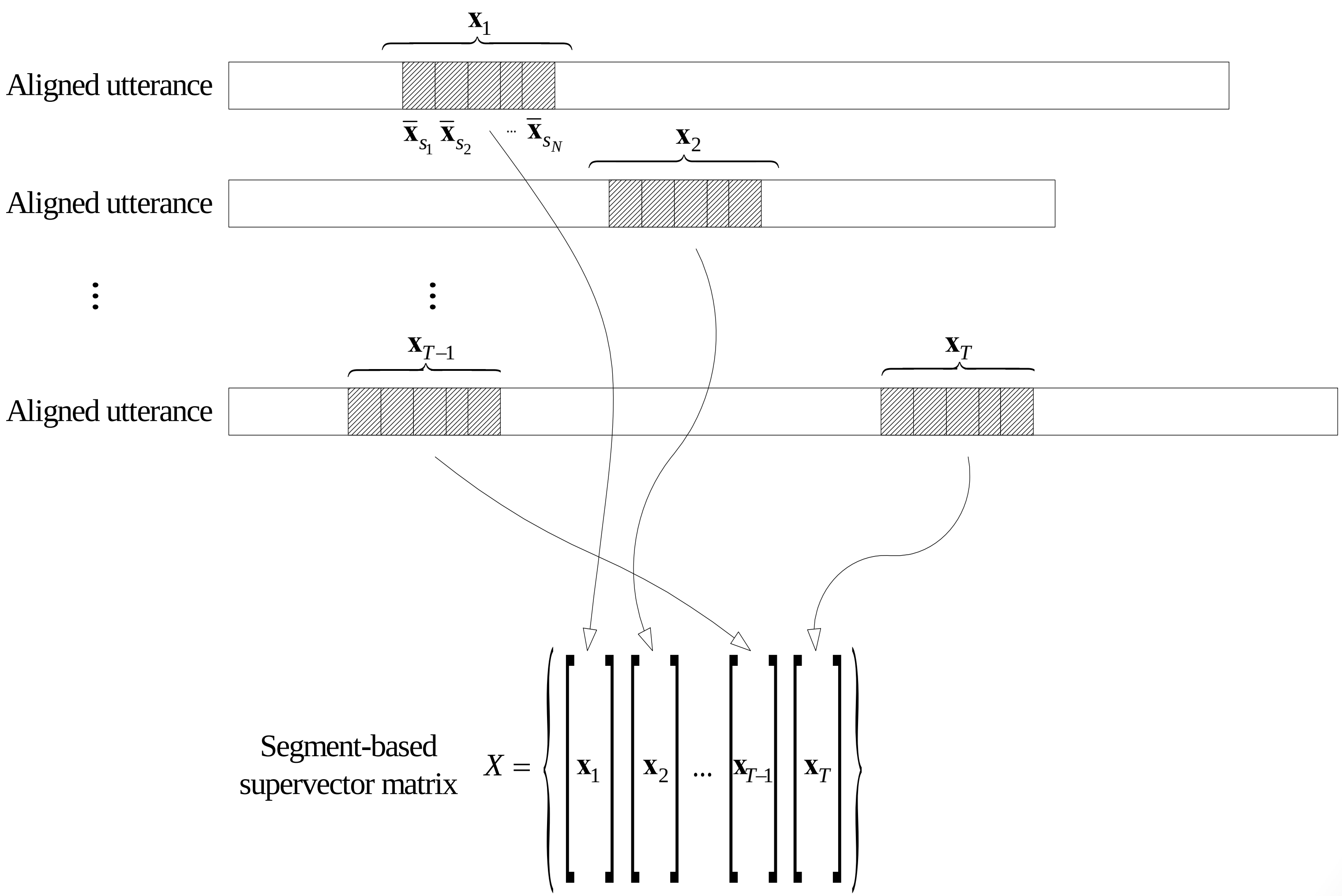
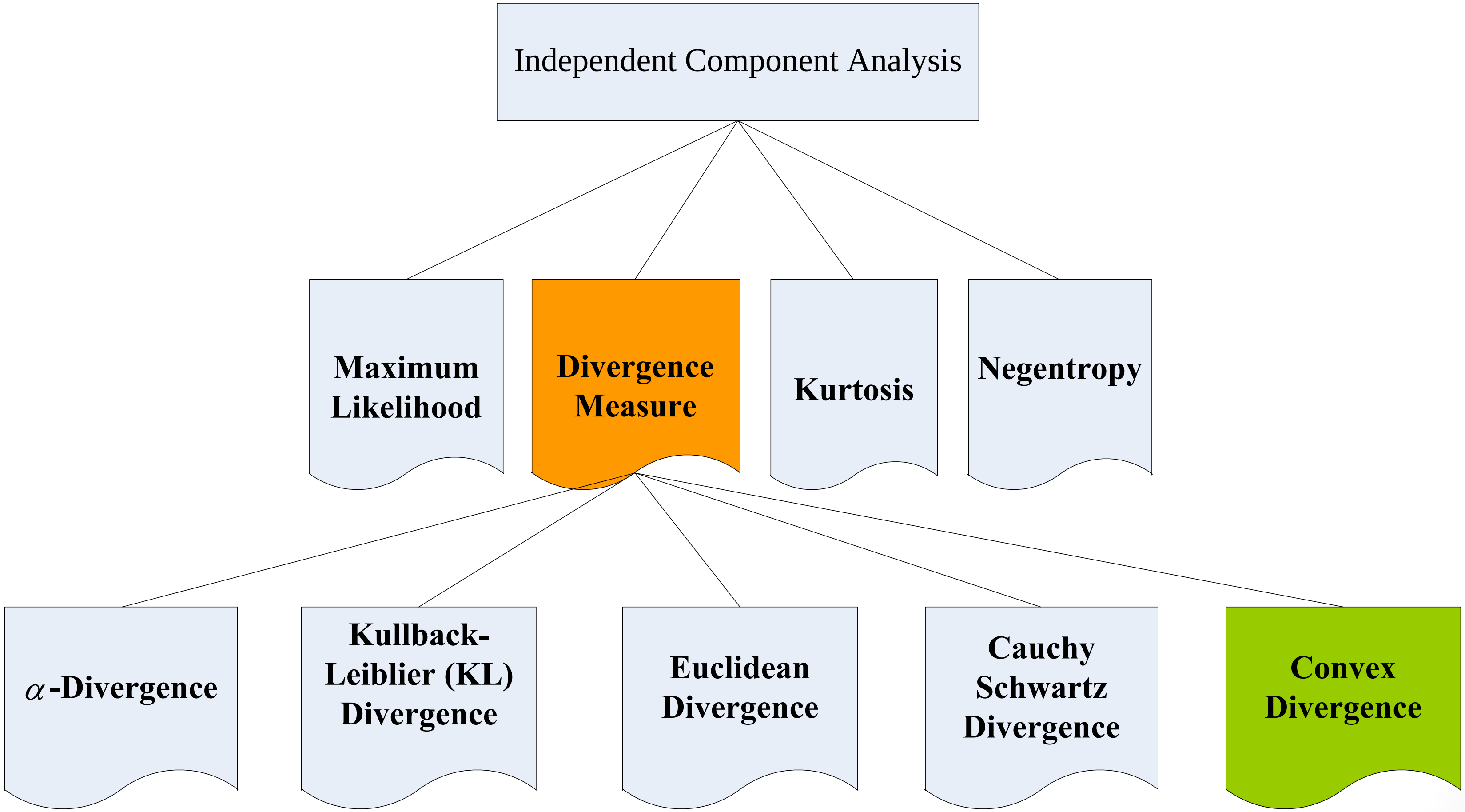


TABLE OF CONTENTS

ICA Objective Function



Mutual Information & KL Divergence

- Mutual information between two variables y_1 and y_2 is defined by using the *Shannon entropy* $H[\cdot]$.
- It can be formulated as the KL divergence or *relative entropy* between the joint distribution $p(y_1, y_2)$ and the product of marginal distribution $p(y_1)p(y_2)$

$$
\begin{aligned}
D_{\mathrm{KL}}(y_1, y_2) &= H[p(y_1)] + H[p(y_2)] - H[p(y_1, y_2)] \\
&= \iint p(y_1, y_2) \log \frac{p(y_1, y_2)}{p(y_1)p(y_2)} dy_1 dy_2
\end{aligned}
$$

where $D_{\mathrm{KL}}(y_1, y_2) \geq 0$.

Divergence Measures

- *Euclidean divergence* $D_{\mathrm{E}}(y_1, y_2)$

$$\iint (p(y_1, y_2) - p(y_1)p(y_2))^2 dy_1 dy_2$$

- *Cauchy-Schwartz divergence* $D_{\mathrm{CS}}(y_1, y_2)$

$$\log \left\{ \frac{\iint p(y_1, y_2)^2 dy_1 dy_2 \cdot \iint p(y_1)^2 p(y_2)^2 dy_1 dy_2}{\left[\iint p(y_1, y_2) p(y_1) p(y_2) dy_1 dy_2\right]^2} \right\}$$

- *α -divergence* $D_\alpha(y_1, y_2, \alpha)$

$$\frac{4}{1-\alpha^2} \iint \left[\frac{1-\alpha}{2} p(y_1, y_2) + \frac{1+\alpha}{2} p(y_1) p(y_2) \right.$$
$$\left. - p(y_1, y_2)^{(1-\alpha)/2} (p(y_1) p(y_2))^{(1+\alpha)/2} \right] dy_1 dy_2$$

Divergence Measures

- *f-divergence* $D_f(y_1, y_2)$

$$\int\int p(y_1)p(y_2)f\left(\frac{p(y_1, y_2)}{p(y_1)p(y_2)}\right)dy_1 dy_2$$

- *Jensen-Shannon divergence* $D_{\mathrm{JS}}(y_1, y_2)$

$$\begin{aligned}&H[\lambda p(y_1, y_2) + (1-\lambda)p(y_1)p(y_2)]\\&-\lambda H[p(y_1, y_2)] - (1-\lambda)H[p(y_1)p(y_2)]\end{aligned}$$

where $0 \leq \lambda \leq 1$. Entropy is a concave function.

Convex Function

- A convex function should meet the *Jensen's inequality*

$$
\begin{aligned}
f(\lambda p(y_1, y_2) + (1-\lambda)p(y_1)p(y_2)) \\
\leq \lambda f(p(y_1, y_2)) + (1-\lambda) f(p(y_1)p(y_2))
\end{aligned}
$$

- A *general convex function* is defined by

$$
f(t) = \frac{4}{1-\alpha^2}\left[\frac{1-\alpha}{2} + \frac{1+\alpha}{2}t - t^{(1+\alpha)/2}\right]
$$

$$
f(t) \geq 0 \text{ for } t \geq 0, f(1) = 0 \text{ and } f'(1) = 0.
$$

Convex Divergence

- By assuming equal weight $\lambda = 1/2$, we have

$$\frac{2}{1-\alpha^2}\iint \left\{ 2\left[\frac{p(y_1,y_2)+p(y_1)p(y_2)}{2}\right]^{(1+\alpha)/2} - \left[p(y_1,y_2)^{(1+\alpha)/2} + (p(y_1)p(y_2))^{(1+\alpha)/2}\right] \right\} dy_1 dy_2.$$

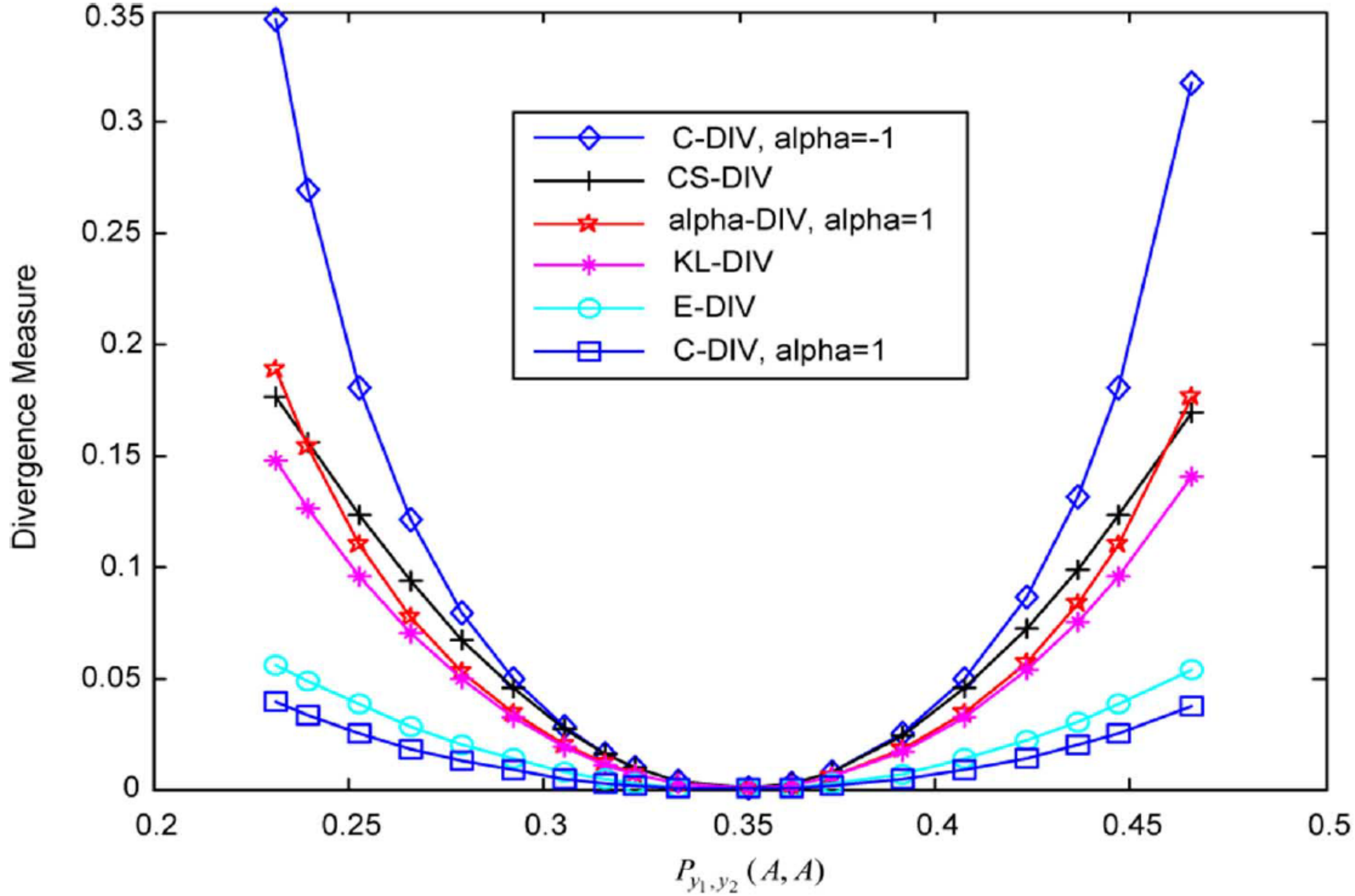
- When $\alpha = -1$, C-DIV is derived as a case with convex function $-\log(\cdot)$

$$D_{\mathrm{C}}(y_1, y_2, -1) = \iint \left\{ \log\left[\frac{p(y_1,y_2)+p(y_1)p(y_2)}{2}\right] - \frac{1}{2}\log(p(y_1,y_2)) - \frac{1}{2}\log(p(y_1)p(y_2)) \right\} dy_1 dy_2$$

Different Divergence Measures

	Symmetric Divergence	Convexity Parameter	Combination Weight	Special Realization from other DIV
KL-DIV	No	no	No	α-DIV ($\alpha = -1$)
α-DIV	No	α	No	f-DIV (using (6))
f-DIV	No	α	No	no
JS-DIV	$\lambda = 0.5$	no	λ (convex)	C-DIV ($\lambda = 0.5$, $\alpha = 1$)
Convex-log DIV	$\lambda = 0.5$	no	λ (convex)	C-DIV ($\lambda = 0.5$, $\alpha = -1$)
C-DIV	$\lambda = 0.5$	α	λ (convex)	no

Different Divergence Measures



Convex Divergence ICA

- C-ICA learning algorithm

$$\mathbf{W}^{(n+1)} = \mathbf{W}^{(n)} - \eta \frac{\partial D\left(\mathbf{X}, \mathbf{W}^{(n)}\right)}{\partial \mathbf{W}^{(n)}}$$

$$D_{\mathrm{C}}(\mathbf{X}, \mathbf{W}, \alpha) = \frac{2}{1-\alpha^2} \sum_{t=1}^{T} \left\{ 2 \left[\frac{1}{2} \left(p(\mathbf{W}\mathbf{x}_t) + \prod_{m=1}^{M} p(\mathbf{w}_m \mathbf{x}_t) \right) \right]^{(1+\alpha)/2} - \left[p(\mathbf{W}\mathbf{x}_t)^{(1+\alpha)/2} + \left(\prod_{m=1}^{M} p(\mathbf{w}_m \mathbf{x}_t) \right)^{(1+\alpha)/2} \right] \right\}$$

- *Nonparametric* C-ICA is established by using Parzen window density function.

Simulated Experiments

- A parametric demixing matrix

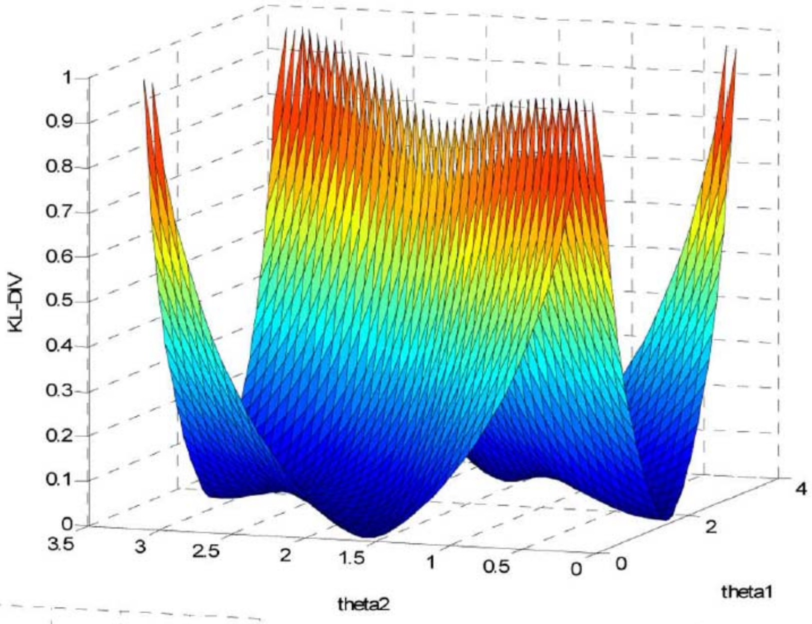
$$W = \begin{bmatrix} \cos\theta_1 & \sin\theta_1 \\ \cos\theta_2 & \sin\theta_2 \end{bmatrix}$$

- Two sources: *super-Gaussian* and *sub-Gaussian* distribution

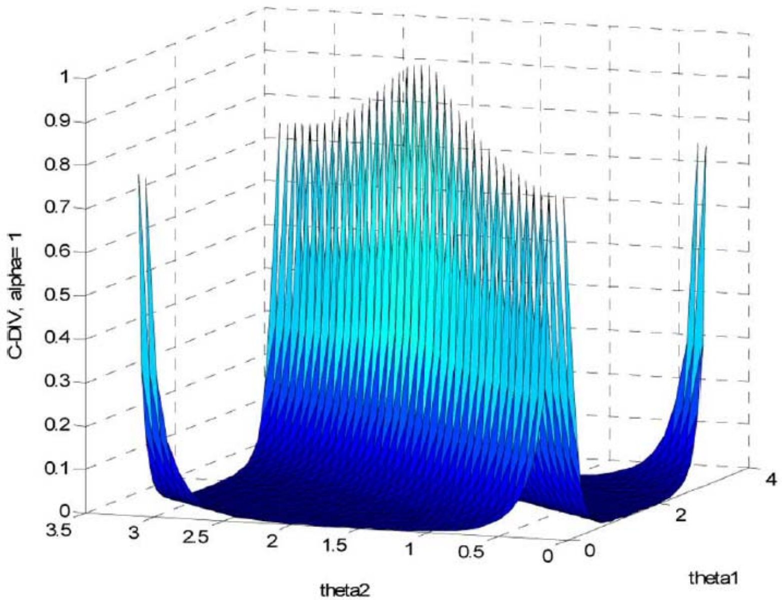
$$p(s_1) = \begin{cases} \dfrac{1}{2\tau_1}, & s_1 \in [-\tau_1, \tau_1] \\ 0, & \text{otherwise} \end{cases} \qquad p(s_2) = \frac{1}{2\tau_2} \exp\left(-\frac{|s_2|}{\tau_2} \right)$$

- Kurtosis
  - Source 1: -1.13, source 2: 2.23

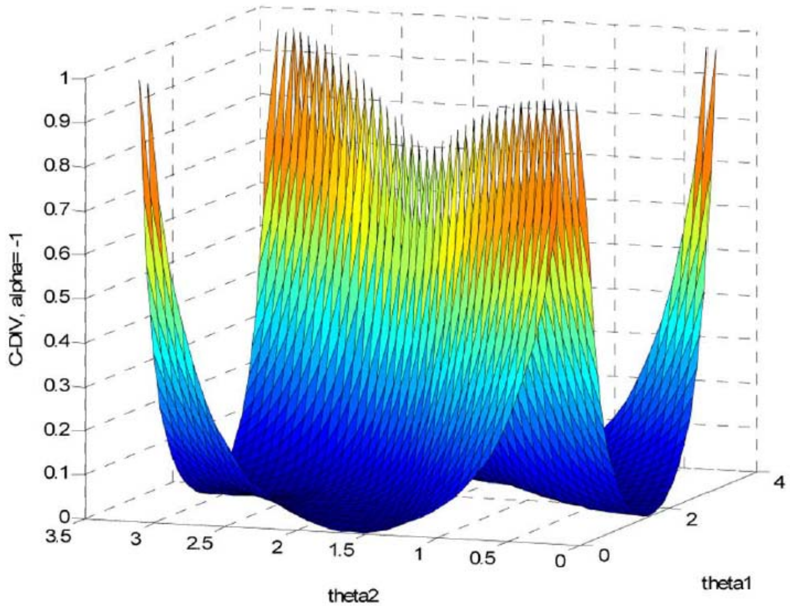
KL-DIV



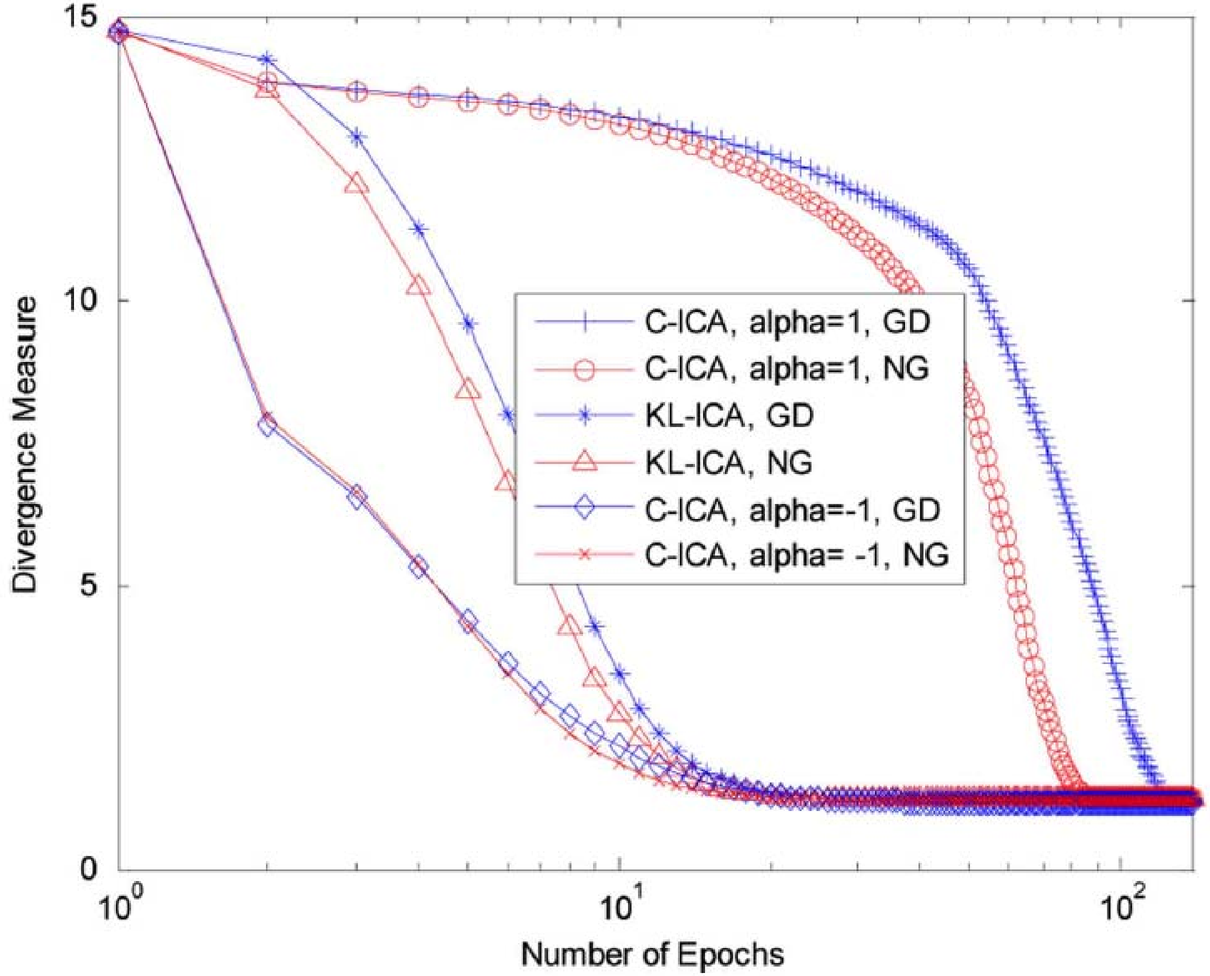
C-DIV alpha=1



C-DIV, alpha= -1



Learning Curves



Experiments on Blind Source Separation

- One music signal and two speech signals from two male speakers were sampled from ICA'99 BSS Test Sets at http://sound.media.mit.edu/ica-bench/

- Mixing matrix

$$A = \begin{bmatrix} 0.8 & 0.2 & 0.3 \\ 0.3 & -0.8 & 0.2 \\ -0.3 & 0.7 & 0.3 \end{bmatrix}$$

- Evaluation metric
  − signal-to-interference ratio (SIR)

$$\mathrm{SIR(dB)} = 10 \log_{10} \left(\sum_{t=1}^{T} \|\mathbf{s}_t\|^2 \Big/ \sum_{t=1}^{T} \|\mathbf{y}_t - \mathbf{s}_t\|^2 \right)$$

Comparison of Different Methods

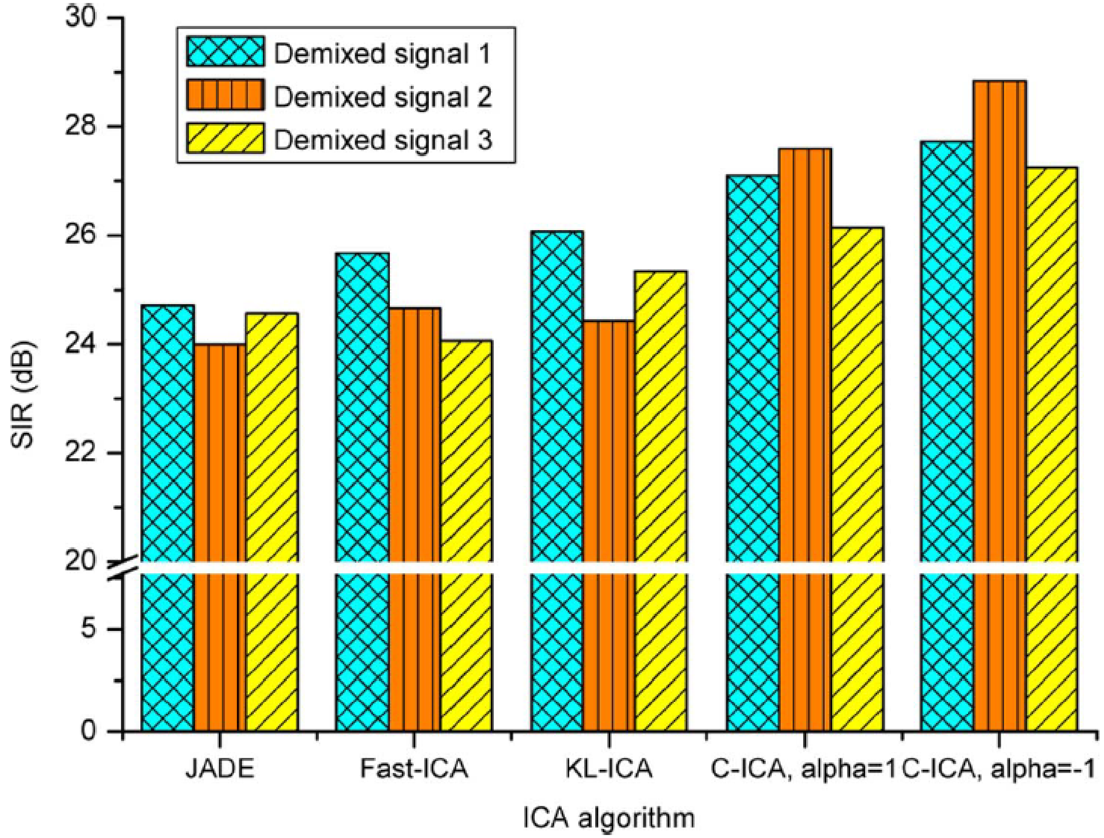


TABLE OF CONTENTS

1. ***Independent Component Analysis***
2. ***Case Study I: Speech Recognition***
   - Independent voices
   - Nonparametric likelihood ratio ICA
3. ***Case Study II: Blind Source Separation***
   - Convex divergence ICA
   - Nonstationary Bayesian ICA
   - Online Gaussian process ICA
4. ***Summary***

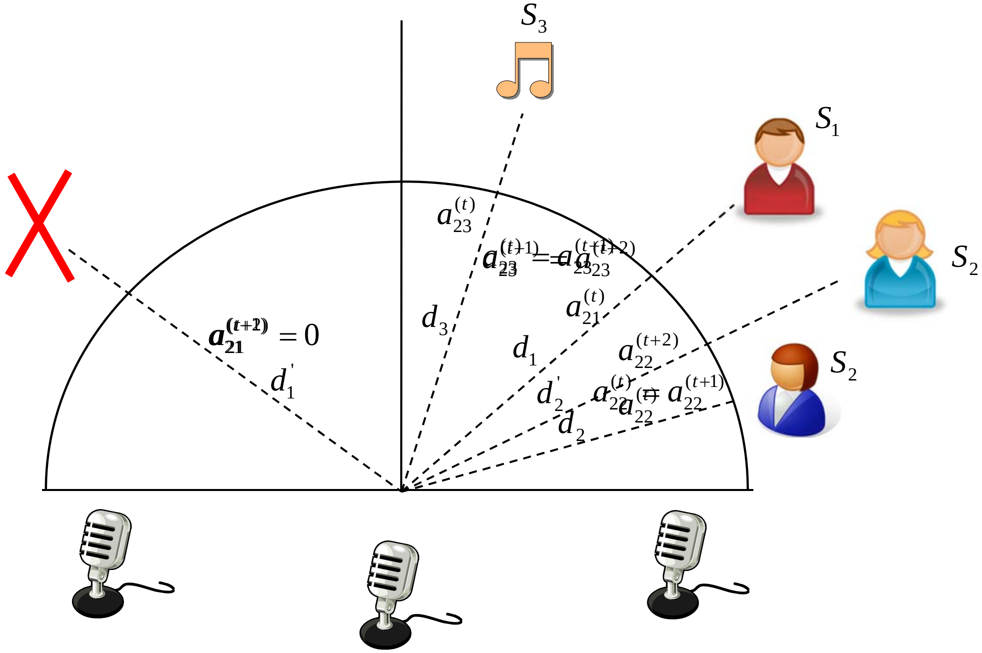
Why Nonstationary Source Separation?

- Real-world *blind source separation*
  - number of sources is unknown
  - BSS is a dynamic time-varying system
  - mixing process is nonstationary

- Why *nonstationary*?
  - Bayesian method using ARD can determine the changing number of sources
  - recursive Bayesian for online tracking of nonstationary conditions
  - Gaussian process provides a nonparametric solution to represent temporal structure of time-varying mixing system.

Nonstationary Mixing Systems

- Time-varying mixing matrix
- Source signals may abruptly appear or disappear



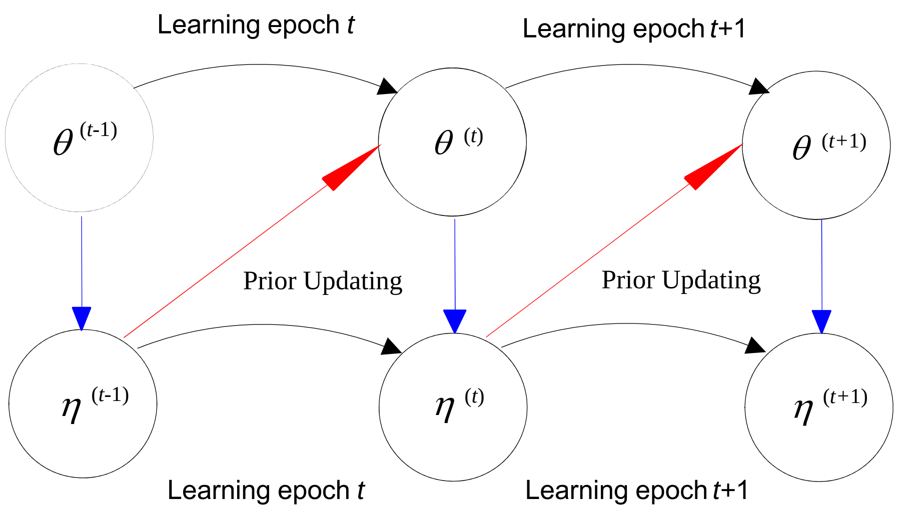
Nonstationary Bayesian (NB) Learning

- *Maximum a posteriori* estimation of NB-ICA parameters and compensation parameters

$$\theta^{(t)} = \arg\max_{\theta} \; p(X_t|\theta^{(t)}, \eta^{(t-1)})p(\theta)$$

$$\eta^{(t)} = \arg\max_{\eta} \; p(X_t|\theta^{(t)}, \eta)p(\eta|\phi^{t-1})$$

updating



Model Construction

- *Noisy ICA* model $\mathbf{x}_t = A\mathbf{s}_t + \boldsymbol{\varepsilon}_t$
- Likelihood function of an observation $\mathbf{x}_t$

$$p(\mathbf{x_t}|\mathbf{A}^{(t)}, \mathbf{s}^{(t)}, \beta^{(t)}) = N(\mathbf{x}_t|\mathbf{A}^{(t)}\mathbf{s}^{(t)}, {\beta^{(t)}}^{-1} I_N)$$

- Distribution of model parameters
  - source $p(\mathbf{s}^{(t)}|\boldsymbol{\pi}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\gamma}^{(t)}) = \prod_{m=1}^{M}\left[\sum_{k=1}^{K} \pi_k^{(t)} N(s_m^{(t)}|\mu_k^{(t)}, {\gamma_k^{(t)}}^{-1})\right]$
  - mixing matrix $p(A^{(t)}|\boldsymbol{\alpha}^{(t)}) = \prod_{m=1}^{M}\left[\prod_{n=1}^{N} N(a_{nm}^{(t)}|0, {\alpha_m^{(t)}}^{-1})\right]$
  - noise $p(\varepsilon_t|\beta^{(t)}) = N(\varepsilon_t|0, {\beta^{(t)}}^{-1} I_N)$

Prior & Marginal Distributions

- Prior distributions
  - precision of noise $p(\beta^{(t)}|u_\beta, w_\beta) = \mathrm{Gam}(\beta^{(t)}|u_\beta, w_\beta)$
  - precision of mixing matrix $p(\boldsymbol{\alpha}^{(t)}|u_\alpha, w_\alpha) = \prod_{m=1}^{M} \mathrm{Gam}(\alpha_m^{(t)}|u_\alpha, w_\alpha)$

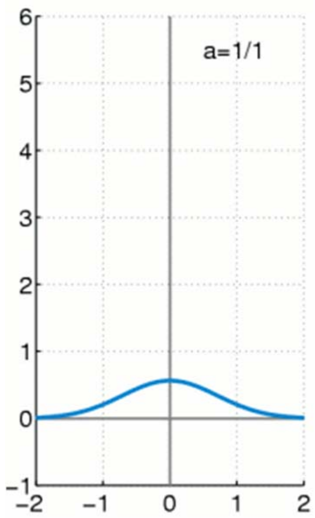
- *Marginal likelihood* of NB-ICA model

$$p(X) = \prod_{t=1}^{T} \int p(\mathbf{x}_t|A^{(t)}, \mathbf{s}^{(t)}, \boldsymbol{\alpha}^{(t)}, \beta^{(t)}) p(A^{(t)}|\boldsymbol{\alpha}^{(\boldsymbol{t})}) p(\boldsymbol{\alpha}^{(\boldsymbol{t})}|u_\alpha^{(t)}, w_\alpha^{(t)})$$
$$\times p(\mathbf{s}^{(t)}|\boldsymbol{\pi}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\gamma}^{(t)}) p(\boldsymbol{\beta}^{(t)}|\boldsymbol{u}_\beta^{(t)}, \mathbf{w}_\beta^{(t)}) dA^{(t)} d\mathbf{s}^{(t)} d\boldsymbol{\alpha}^{(t)} d\boldsymbol{\beta}^{(t)}$$

Automatic Relevance Determination

- Detection of source signals

$$\alpha_m^{(t)} = \begin{cases} \infty & , \quad a_m^{(t)} = \{a_{nm}^{(t)}\} \to 0 \\ < \infty & , \quad a_m^{(t)} = \{a_{nm}^{(t)}\} \neq 0 \end{cases}$$



  – number of sources can be determined

Compensation for Nonstationary ICA

$$G_{\eta^{(t)}}(\alpha^{(t)}) = \eta^{(t)}\alpha^{(t)}$$

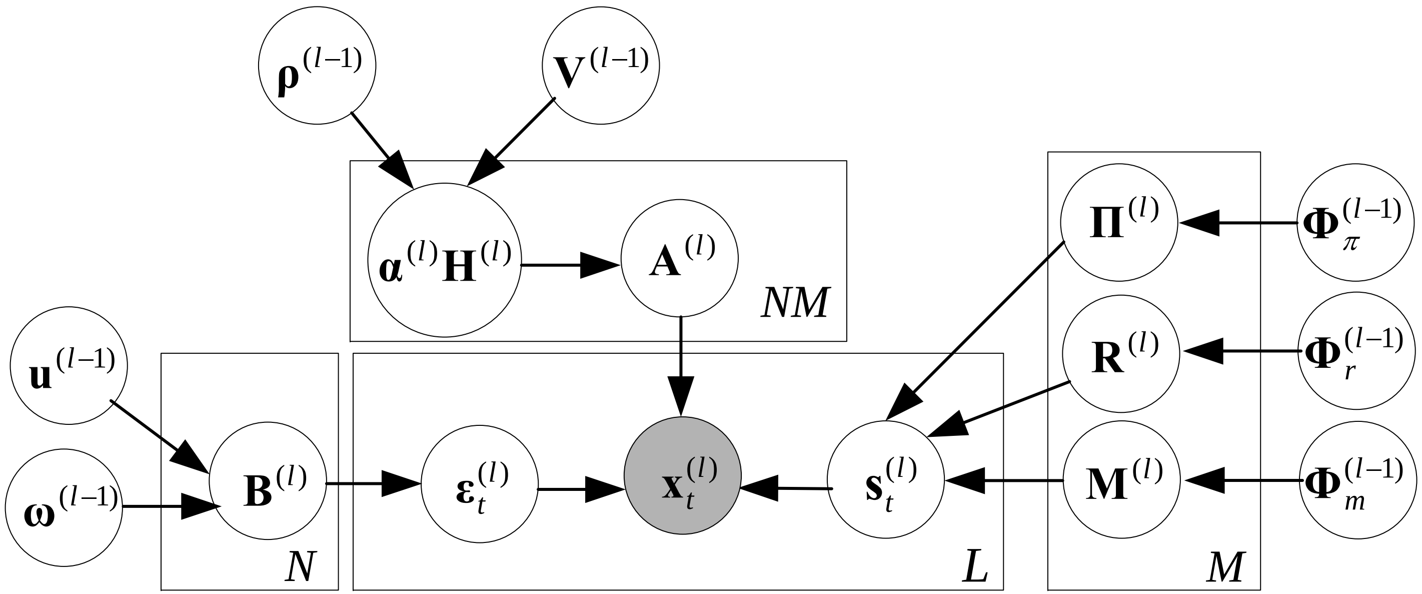
- Prior density of compensation parameter
  - conjugate prior (Wishart distribution)

$$p(\eta^{(t)}|\varphi^{t-1} = \{\nu^{t-1}, \Lambda^{t-1}\})c\left(M, \nu^{t-1}\right)\left|\frac{1}{2}\nu^{t-1}\Lambda^{t-1}\right|^{\left(\nu^{t-1}-1\right)/2}$$

$$\times\left|\eta^{(t)}\right|^{\left(\nu^{t-1}-M-2\right)/2}\exp\left[-\frac{1}{2}\mathrm{tr}\left(\nu^{t-1}\Lambda^{t-1}\eta^{(t)}\right)\right]$$

$$c\left(M, \nu^{t-1}\right) = \left(\pi^{M(M-1)/4}\prod_{m=1}^{M}\Gamma\left(\left(\nu^{t-1}-m\right)/2\right)\right)^{-1}$$

Graphical Model for NB-ICA



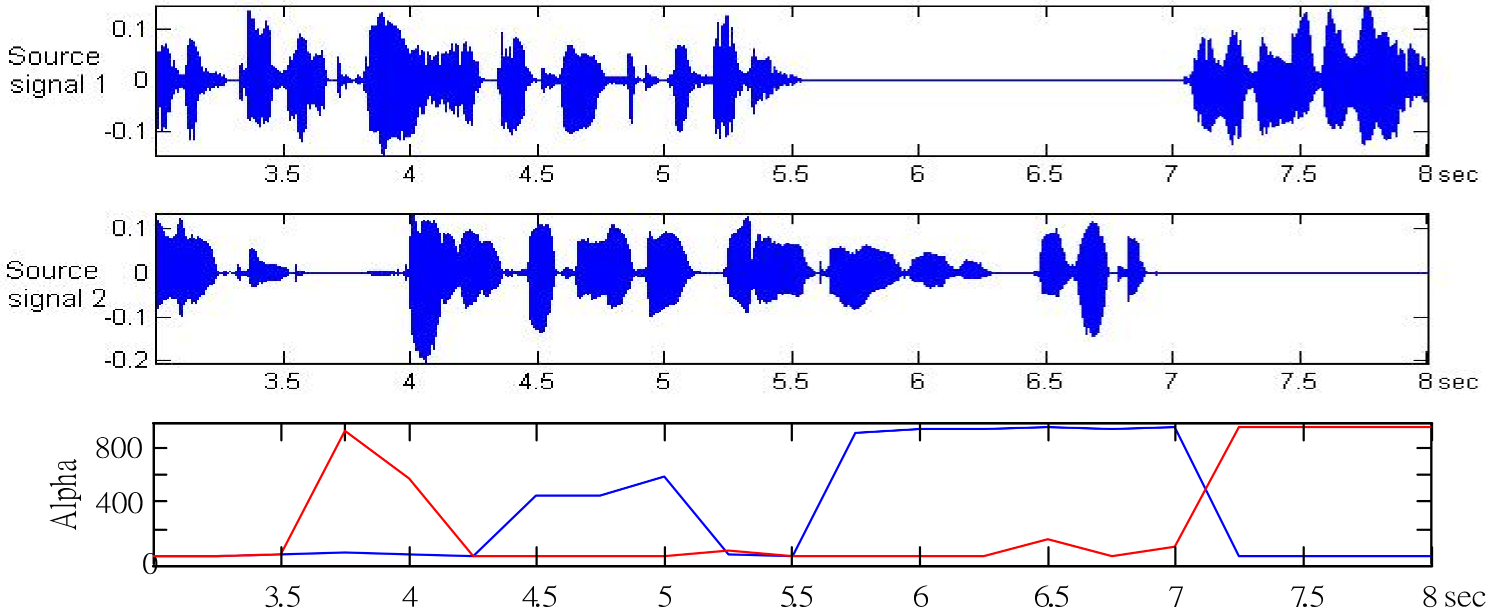
Experiments

- Nonstationary Blind Source Separation
  - ICA'99 http://sound.media.mit.edu/ica-bench/
- Scenarios
  - state of source signals: active or inactive
  - source signals or sensors are moving: nonstationary mixing matrix

$$A^{(t)} = \begin{bmatrix} \cos(2\pi f_1 t) & \sin(2\pi f_2 t) \\ -\sin(2\pi f_1 t) & \cos(2\pi f_2 t) \end{bmatrix}$$

$$f_1 = 1/5 \text{ Hz} \quad f_2 = 1/2.5 \text{ Hz}$$

Source Signals and ARD Curves



Blue: first source signal
Red: second source signal

TABLE OF CONTENTS

Online Gaussian Process (OLGP)

- Basic ideas
  - incrementally detect the status of source signals and estimate the corresponding distributions from online observation data $\chi^t = \left\{ X^{(1)}, X^{(2)}, \cdots, X^{(t)} \right\}$.
  - *temporal structure* of time-varying mixing coefficients $A^{(t)}$ are characterized by *Gaussian process*.
  - Gaussian process is a nonparametric model which defines the prior distribution over functions for Bayesian inference.

Model Construction

- *Noisy ICA* model $\mathbf{x}^{(t)} = A^{(t)}\mathbf{s}^{(t)} + \boldsymbol{\varepsilon}^{(t)}$
- Likelihood function

$$p(\mathbf{x}^{(t,i)} \mid \mathbf{A}^{(t,i)}, \mathbf{s}^{(t,i)}, \beta^{(t,i)}) = N(\mathbf{x}^{(t,i)} \mid \mathbf{A}^{(t,i)}\mathbf{s}^{(t,i)}, {\beta^{(t,i)}}^{-1} I_N)$$

- Distribution of model parameters
  - source

$$p(\mathbf{s}^{(t,i)} \mid \boldsymbol{\pi}^{(t,i)}, \boldsymbol{\mu}^{(t,i)}, \boldsymbol{\gamma}^{(t,i)}) = \prod_{m=1}^{M} \left[\sum_{k=1}^{K} \pi_{m,k}^{(t,i)} N(s_m^{(t,i)} \mid \mu_{m,k}^{(t,i)}, {\gamma_{m,k}^{(t,i)}}^{-1}) \right]$$

  - noise

$$p(\varepsilon^{(t,i)} \mid \beta) = N(\varepsilon^{(t,i)} \mid 0, {\beta_m^{(t,i)}}^{-1} I_N)$$

  - $p(\beta^{(t,i)}) = \mathrm{Gam}(u_\beta^{(t,i)}, w_\beta^{(t,i)})$

Gaussian Process

- *Mixing matrix*
  - $A^{(t)}$ is generated by the latent function $f_{nm}(\cdot)$

$$a_{nm}^{(t,i)} = f_{nm}\left(\mathbf{a}_{nm}^{(t,i-1,i-p)}\right)$$

$$\mathbf{a}_{nm}^{(t,i-1,i-p)} = \left[a_{nm}^{(t,i-1p)} a_{nm}^{(t,i-2)} \cdots \; a_{nm}^{(t,i-p)}\right]^T$$

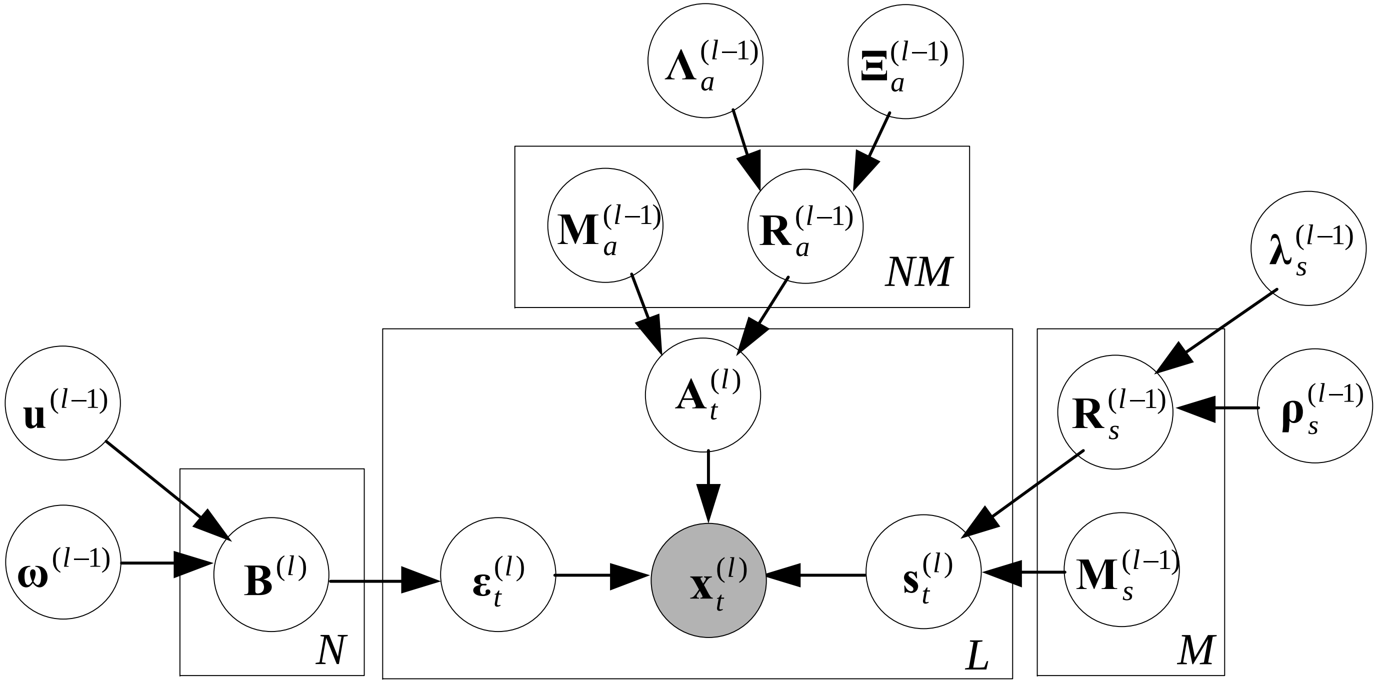
  - GP is adopted to describe the distribution of $f_{nm}(\cdot)$

$$f_{nm}(\mathbf{a}_{nm}^{(t,i-1,i-p)}) \sim N(0, \kappa(\mathbf{a}_{nm}^{(t,i-1,i-p)}, \mathbf{a}_{nm}^{(t,\tau-1,\tau-p)}))$$

$$\kappa(\mathbf{a}_{nm}^{(t,i-1,i-p)}, \mathbf{a}_{nm}^{(t,\tau-1,\tau-p)}) = \rho_{nm}^{(t,i)} \exp\left\{-\frac{\lambda_{nm}^{(t,i)}}{2}\left\|\mathbf{a}_{nm}^{(t,i-1,i-p)} - \mathbf{a}_{nm}^{(t,\tau-1,\tau-p)}\right\|^2\right\}$$

  - $\left\{\lambda_{nm}^{(t,i)}, \rho_{nm}^{(t,i)}\right\}$ are *hyperparameters* of kernel function

Graphical Model for OLGP-ICA

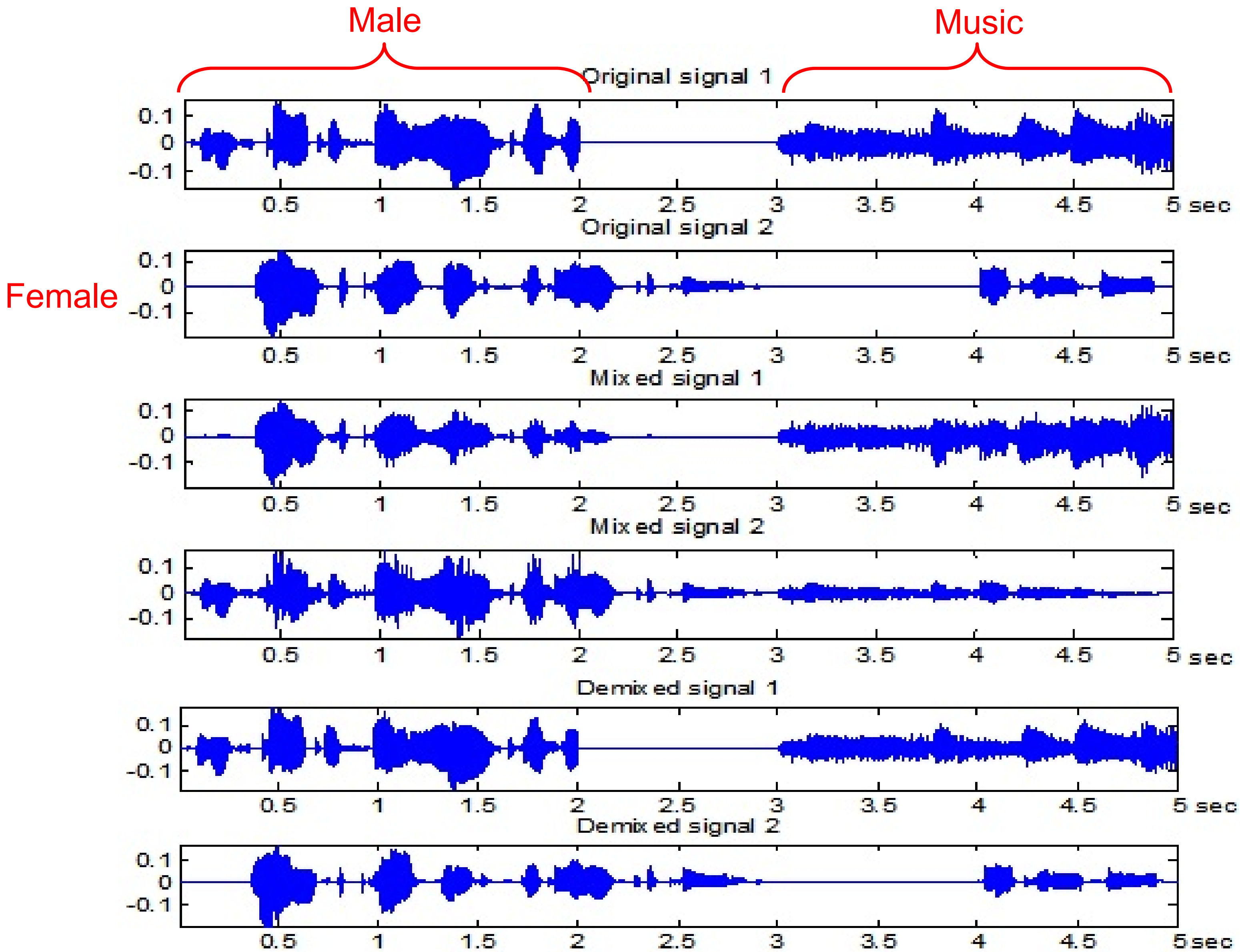


Experimental Setup

- Nonstationary source separation using source signals from
  - http://www.kecl.ntt.co.jp/icl/signal/
- Nonstationary scenarios
  - status of source signals: active or inactive
  - source signals or sensors are moving: nonstationary mixing matrix

$$A^{(t)} = \begin{bmatrix} \cos(2\pi f_1 t) & \sin(2\pi f_2 t) \\ -\sin(2\pi f_1 t) & \cos(2\pi f_2 t) \end{bmatrix}$$

$$f_1 = 1/20 \text{ Hz} \qquad f_2 = 1/10 \text{ Hz}$$


Male
Music
Original signal 1
Original signal 2
Female
Mixed signal 1
Mixed signal 2
Demixed signal 1
Demixed signal 2
0.1
0
-0.1
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5 sec

Comparison of Different Methods

- Signal-to-interference ratios (SIRs) (dB)

	VB-ICA	BICA-HMM	Switching-ICA	Online VB-ICA	OLGP-ICA
Demixed signal 1	7.97	9.04	12.06	11.26	17.24
Demixed signal 2	-3.23	-1.5	-4.82	4.47	9.96

Summary

- We presented speaker adaptation method based on *independent voices* by fulfilling ICA perspective.
- A *nonparametric likelihood ratio* ICA was proposed according to hypothesis test theory.
- A *convex divergence* was developed as an optimization metric for ICA algorithm.
- A *nonstationary Bayesian* ICA was proposed to deal with nonstationary mixing system.
- An *online Gaussian process* ICA was presented for nonstationary and temporally correlated source separation.
- ICA methods could be extended to solve *nonnegative matrix factorization* and single-channel separation.

References

- J.-T. Chien and B.-C. Chen, “A new independent component analysis for speech recognition and separation”, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1245-1254, 2006.
- J.-T. Chien, H.-L. Hsieh and S. Furui, “A new mutual information measure for independent component analysis”, in *Proc. ICASSP*, pp. 1817-1820, 2008.
- H.-L. Hsieh, J.-T. Chien, K. Shinoda and S. Furui, “Independent component analysis for noisy speech recognition”, in *Proc. ICASSP*, pp. 4369-4372, 2009.
- H.-L. Hsieh and J.-T. Chien, “Online Bayesian learning for dynamic source separation”, in *Proc. ICASSP*, pp. 1950-1953, 2010.
- H.-L. Hsieh and J.-T. Chien, “Online Gaussian process for nonstationary speech separation”, in *Proc. INTERSPEECH*, pp. 394-397, 2010.
- H.-L. Hsieh and J.-T. Chien, “Nonstationary and temporally-correlated source separation using Gaussian process”, in *Proc. ICASSP*, pp. 2120-2123, 2011.
- J.-T. Chien and H.-L. Hsieh, “Convex divergence ICA for blind source separation”, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 290-301, 2012.

Thanks to

H.-L. Hsieh

K. Shinoda

S. Furui

jtchien@nctu.edu.tw