PARAMETRIC-PEARSON-BASED INDEPENDENT COMPONENT ANALYSIS FOR FREQUENCY-DOMAIN BLIND SPEECH SEPARATION

Hiroko Kato, Yuichi Nagahara, Shoko Araki, Hiroshi Sawada and Shoji Makino

NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, 619-0237, Kyoto, Japan
phone: + (81) 774 93 5138, fax: + (81) 774 93 1945, email: katohi@cslab.kecl.ntt.co.jp
web: www.kecl.ntt.co.jp

* School of Political Science and Economics, Meiji University
1-1, Kanda-Surugadai, Chiyoda-ku, 101-8301, Tokyo, Japan
phone: + (81) 3 3296 2118, fax: + (81) 3 3296 2350, email: nagahara@kisc.meiji.ac.jp

ABSTRACT

Separation performance is improved in frequency-domain blind source separation (BSS) of speech with independent component analysis (ICA) by applying a parametric Pearson distribution system. ICA adaptation rules include a score function determined by approximated source distribution, and better approximation improves separation performance. Previously, conventional hyperbolic tangent (tanh) or generalized Gaussian distribution (GGD) was uniformly applied to the score function for all frequency bins, despite the fact that a wideband speech signal has different distributions at different frequencies. To obtain better score functions, we propose the integration of a parametric Pearson distribution system with ICA learning rules. The score function is estimated by using appropriate Pearson distribution parameters for each frequency bin. We consider three estimation methods with Pearson distribution parameters and conduct separation experiments with real speech signals convolved with actual room impulse responses. Consequently, the signal-to-interference ratio (SIR) of the proposed methods significantly improve over 3 dB compared to conventional methods.

1. INTRODUCTION

Frequency-domain blind source separation (BSS) based on independent component analysis (ICA) [1][2] has been proposed for the BSS of convolutive mixtures of speech signals [3][4][5][6][7]. ICA adaptation rules have a score function determined by approximated source distribution. In conventional applications of ICA to speech signals, super-Gaussian distribution is used as a distribution model, typically with the score function *tanh* [1][2]. Recently, a GGD-based modeling approach that is more adaptable than *tanh* has also been presented [8][9]. In these modeling approaches, the score function is uniformly applied to all frequencies.

However, when transforming the speech signals to the frame series (i.e., the time sequence in the frequency domain) by short-time Fourier transform (STFT), it is clear that the distribution of the frame series illustrates different patterns for each frequency. The shapes seem to have a non-Gaussian appearance, as shown by fat-tailed and skewed characteristics, that resembles various distribution shapes of the Pearson distribution system [10]. A previous study investigated the application of the Pearson distribution system to ICA (Pearson-ICA) [11]. This approach showed better separation performance than such conventional nonlinear functions as *tanh*. Furthermore, a nonparametric ICA approach to estimating the source distribution was proposed, and its separation performance was compared to those of several methods such as Pearson-ICA, Fast ICA, and Kernel-ICA [12]; however, [11] and [12] were performed in the time-domain and used artificial data. In order to achieve BSS of convolutive mixtures, which is the focus of this paper, time-domain processing is inefficient.

Therefore, this article proposes a Pearson distribution system approach to frequency-domain BSS. In this approach, the source distribution for each frequency bin is modeled by an adaptive parametric Pearson distribution, and the score function is formed by the parameters of the distribution to improve separation performance. So far, practical implementation of distribution parameters that depend on the moment of the signal has been unwieldy. To overcome such problems, [13], [14] and [15] have proposed the solution of introducing a discrimination method for Pearson distribution types and transforming formulae between the moment and distribution parameters. Accordingly, we also employ a new implementation in the proposed score function's estimation procedure.

This paper is organized as follows: Section 2 introduces the basic framework of BSS and conventional nonlinear functions. Section 3 outlines the parametric Pearson distribution system. Section 4 describes our proposed BSS methods. Section 5 shows experimental results, and conclusions are given in Sec. 6.

2. BLIND SOURCE SEPARATION OF SPEECH

In this paper, we consider the blind source separation (BSS) of speech signals observed in real environments, i.e., the BSS of convolutive mixtures of speech. In such environ-

ments, N source signals $s_i(n)$ are observed with their reverberant components and delays by M sensors. Therefore, observations are modeled as convolutive mixtures:

$$x_j(n) = \sum_{i=1}^{N}\sum_{p=1}^{P} h_{ji}(p)s_i(n-p+1) \quad (j=1,\cdots,M)\ , \qquad (1)$$

$h_{ji}(n)$: P – taps impulse response from source i to sensor j.

Our goal is to obtain separated signals $y_k(n)$ $(k=1,\cdots,N)$ using only the information provided by observations $x_j(n)$. In this paper, we handle the case of $N=M=2$ (Fig. 1); however, we can expand our method to any $N=M$ case without loss of generality.

This paper employs a frequency-domain approach, that is, a short-time Fourier transform (STFT) is performed to convert our problem into a linear instantaneous mixture at each frequency. In the frequency domain, mixtures (1) are modeled as

$$X_j(f,m) = \sum_{i=1}^{N} H_{ji}(f)S_i(f,m)\ , \qquad (2)$$

where f denotes a frequency and m is the frame index. With matrices, (2) can be written as

$$\mathbf{X}(f,m) = \mathbf{H}(f)\mathbf{S}(f,m),$$

$\mathbf{H}(f)$: $M \times N$ mixing matrix,

$\mathbf{S}(f,m) = [S_1(f,m),\cdots,S_N(f,m)]^T$: STFT of source signal,

$\mathbf{X}(f,m) = [X_1(f,m),\cdots,X_N(f,m)]^T$: STFT of observed signal.

In a blind scenario, $\mathbf{H}(f)$ and $\mathbf{S}(f,m)$ are unknown.

The separation process can be formulated in each frequency f:

$$\mathbf{Y}(f,m) = \mathbf{W}(f)\mathbf{X}(f,m)\ , \qquad (3)$$

$\mathbf{Y}(f,m) = [Y_1(f,m),\cdots,Y_N(f,m)]^T$: estimated source signal vector,

$\mathbf{W}(f)$: $N \times M$ separation matrix.

$\mathbf{W}(f)$ is determined so that $Y_1(f,m)$,$\cdots$, $Y_N(f,m)$ become mutually independent using ICA. After getting the separated signals (3), we convert the frequency-domain signal $Y_k(f,m)$ into a time-domain signal by using inverse STFT.

The separation matrix is independently estimated at each frequency. An algorithm based on a natural gradient [16] is widely used for this. The adaptation rule of the i-th iteration is

$$\mathbf{W}_{i+1}(f) = \mathbf{W}_i(f) + \eta[\mathbf{I} - \left\langle \Phi(\mathbf{Y}(f,m))\mathbf{Y}^H(f,m)\right\rangle]\bullet\mathbf{W}_i(f), \qquad (4)$$

$\langle Y(f,m)\rangle$: average with respect to m,

H: transpose conjugate, η: adaptation stepsize.

Here, $\Phi(Y)$ ($Y(f,m)$ is simplified by Y) indicate the score function. If source distributions p are known, score functions (5) are defined as [1][2]:

$$\Phi(Y) = -\frac{p'(|Y|)}{p(|Y|)}\exp(j\angle Y)\ , \qquad (5)$$

Y: complex number, $|\cdot|$: absolute value, $\angle Y$: argument, where $p'(x) = dp(x)/dx$. In blind separation, however, source distribution cannot be obtained *a priori*. Usually, the score function has to be approximated, and the conventional *tanh* is widely used at all frequencies for speech separation because speech signals have a super-Gaussian distribution [1][2]:

$$\Phi(Y) = \tanh(g|Y|)\exp(j\angle Y)\ . \qquad (6)$$

As mentioned in Sec. 1, conventional GGD has also been applied to BSS [8]. The GGD-based score function, uniformly applied to all frequencies [9], is represented by

$$\Phi(Y) = |Y|^{\beta-1}\, sign(|Y|)\exp(j\angle Y). \qquad (7)$$

β: shape parameter.

As is well known in speech signal processing/engineering, for $\beta = 1$, 2, and 0.5, the GGD becomes, respectively, a Laplacian distribution, whose speech closely follows it, a standard Gaussian distribution, and a Gamma distribution.

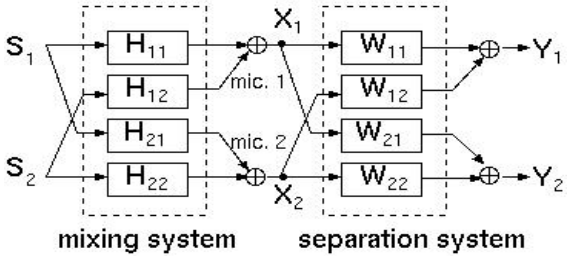

Figure 1. Frequency-domain speech BSS system ($N = M = 2$)

3. PARAMETRIC PEARSON SYSTEM APPROACH

The above conventional nonlinear function approach is uniformly applied to all frequencies; however, the actual speech signal for each frequency has a different distribution. Therefore, Fig. 2 shows the distribution shapes a, b, c, d, e, and f for bins 3, 5, 30, 125, 250, and 500, respectively. To adapt these different shapes, we apply the Pearson distribution system, which is widely used to model such various distributions as Gaussian, Student's *t*, gamma, and beta. Distribution parameters are detected by the sample moments as mean, variance, skewness (Skew), and kurtosis (Kurt).

Pearson [10] defined the differential equation related to probability density function $p(|Y|)$. If the random variable is complex value Y, the form is defined by

$$-\frac{p'(|Y|)}{p(|Y|)} = \frac{b_0 + b_1|Y|}{c_0 + c_1|Y| + c_2|Y|^2}\exp(j\angle Y)\ . \qquad (8)$$

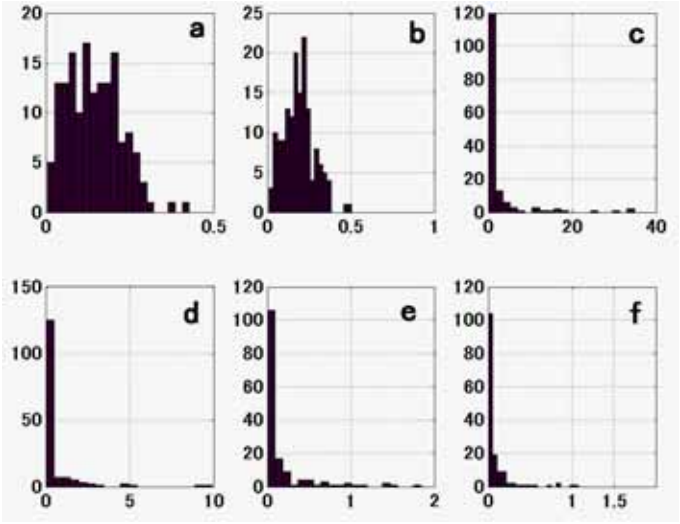

Figure 2. Histograms a, b, c, d, e and f of the frame series for several frequency bins f =3, 5, 30, 125, 250, and 500. STFT frame size: 512, Sampling rate: 8 kHz

Note that form (8) has the same shape as score function (5) of the ICA. That is, if the coefficients of (8) can be estimated by an appropriate method through the observed data at each frequency, we can obtain the score function to approximate the source distribution at each frequency. For estimation of several kinds of distribution families in a Pearson system, [13] introduced parameter κ using Skew and Kurt obtained by the data:

$$\kappa = \frac{(\text{Skew})^2(\text{Kurt}+3)^2}{4(2\times\text{Kurt}-3\times(\text{Skew})^2-6)(4\times\text{Kurt}-3\times(\text{Skew})^2)}. \quad (9)$$

For $0>\kappa$, $0<\kappa<1$, $1<\kappa$, the types are discriminated to I, IV, and VI, respectively. In Fig. 2, panels a and b represent Types IV and VI. Panels c to f represent the shapes of Type I. In our preliminary consideration of the STFT series of real speech data, we calculated κ values for each frequency bin (Fig. 3). Figure 3 shows the distribution of the frame series as classified by Types I, VI, and IV. Type I was widely detected in the mid- and high-frequency bins; however, the height and tail of the distribution were different for each bin, as shown in panels c, d, e, and f of Fig. 2. The distribution figures are each different J-shaped distributions.

Then the score functions of Types I, IV and VI were applied, as described below in terms of distribution parameters:

$$\begin{aligned}
\text{I} \quad &: \Phi(|Y|) = \frac{-(p+q+2)|Y|+(p+q-2)a+(p-1)b}{|Y|^2-(2a+b)|Y|+a(a+b)}\exp(j\angle Y),\\
\text{IV} &: \Phi(|Y|) = \frac{2b|Y|-2b\mu-2b\tau\delta}{|Y|^2-2\mu|Y|-\mu^2+\tau^2}\exp(j\angle Y),\\
\text{VI} &: \Phi(|Y|) = \frac{(c+1)|Y|-(c+1)a-(\beta-1)\alpha}{|Y|^2-(2a-\alpha)|Y|+a(a-\alpha)}\exp(j\angle Y).
\end{aligned} \quad (10)$$

Parameters $p, q, b, a, \mu, \tau, \delta, c, \beta,$ and α can be calculated by the moments of the frame series (see [13][14]). When applying the Pearson system to frequency-domain BSS, our proposed method sets forms (8) and (10) as the score functions.

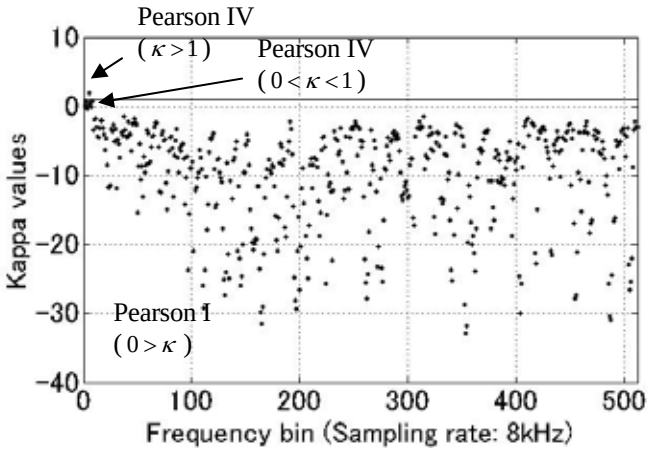

Figure 3. κ values for frequency bin. The horizontal axis indicates frequency bin within 512 STFT frame size and the vertical axis indicates κ value calculated by (9), which was applied to each series obtained by STFT.

4. PROPOSED METHODS

To estimate the score function mentioned above, we propose the following three methods:

Method 1: *Minimization of cross-correlation*

Here, we use the score function (8). To estimate the Pearson parameters $\{b_0(f), b_1(f), c_0(f), c_1(f), c_2(f)\}$, we select the parameters that minimize the sum of the absolute values of the off-diagonal components of $[\mathbf{I}-\langle\Phi(\mathbf{Y})\mathbf{Y}^H\rangle]$ in (4); that is,

$$\sum_i |\phi(Y_i(f,m)Y_j^*(f,m)| \quad (i\neq j). \quad (11)$$

These off-diagonal components represent the higher-order cross-correlation of the outputs. If output signals are well-separated, they become mutually independent, and the value of (11) becomes 0.On the other hand, when the separation is incomplete, the absolute value of off-diagonal components is far from zero. Therefore, we can use the off-diagonal components as measures of separation performance. According to this measure, we search for the Pearson system parameters $\{b_0(f), b_1(f), c_0(f), c_1(f), c_2(f)\}$ that minimize (11) in an arbitrary range with a grid search. Using these parameters, we determine the score function of ICA corresponding to (8) and estimate the separating matrix with (4). In this paper, Pearson parameters are determined by a grid search.

Method 2: *Estimation of appropriate Pearson distribution type*

This method directly determines the appropriate Pearson type for each bin in the learning process of the separation matrix. Score function (10) is used in this method. The specific calculation procedure for each frequency proceeds

specific calculation procedure for each frequency proceeds as follows:

1) Estimate separation matrix $\hat{\mathbf{W}}(f)$ using (6) and set initial value $\mathbf{W}_0(f) = \hat{\mathbf{W}}(f)$;
2) Calculate κ (see (9)) by the skewness and kurtosis of the absolute value of $\mathbf{Y}(f,m)$ obtained with (4);
3) Following κ, the appropriate Pearson distribution type is specified and the parameters of the score function defined in (10) are calculated by the moments of the STFT frame series according to [13] [14];
4) Renew $\mathbf{W}(f)$ by (4); and
5) Iterate procedures 2) to 4) until there is a convergence of (4).

Compared with Method 1, the computational burden is significantly reduced because it is not necessary to perform grid search.

Method 3: *Combining Methods 1 and 2*

In Fig. 3, the κ values for low-frequency bins are unstable. The distribution shape of the frame series at high frequency is illustrated by the various J-shape patterns shown as Type I. In a preliminary investigation, we found that the individual histograms, corresponding to the frequencies of the estimated parameters $b_0(f)$, $b_1(f)$, $c_0(f)$, $c_1(f)$, and $c_2(f)$, have similar tendencies for all speaker combinations at high frequency. On the other hand, at low frequency, the distribution types depend on each speaker. From this fact, we propose another method, Method 3, that combines Methods 1 and 2. Here, the score function (10) is applied at low frequencies, while score function (8) is applied at high frequencies. However, score function (8) is not estimated by a grid search; instead, the pre-estimated mean value for each parameter is used for (8), thus significantly reducing the time needed for calculation. Concretely, the procedure is

1) Calculate mean values $\{\overline{b}_0(f), \overline{b}_1(f), \overline{c}_0(f), \overline{c}_1(f), \overline{c}_2(f)\}$ of parameters estimated by applying Method 1 to arbitrary data combinations;
2) Define f_0 as the boundary point;
3) Apply Method 2 to low-frequency $f \le f_0$ according to the appropriate Pearson type for each frequency bin; and
4) To high-frequency $f > f_0$, input averaged parameters $\{\overline{b}_0(f), \overline{b}_1(f), \overline{c}_0(f), \overline{c}_1(f), \overline{c}_2(f)\}$ for each bin directly into (8).

For choosing the best f_0 in advance, we compared SIR values when using f_0 between 0 and 200 bins and selected the f_0 that provided the highest SIR.

5. EXPERIMENTAL RESULTS

5.1 Experimental conditions

We conducted separation experiments with real speech signals and measured room impulse responses. The speech data were convolved with impulse responses measured in an actual room (Fig. 4) whose reverberation time was 130 ms.

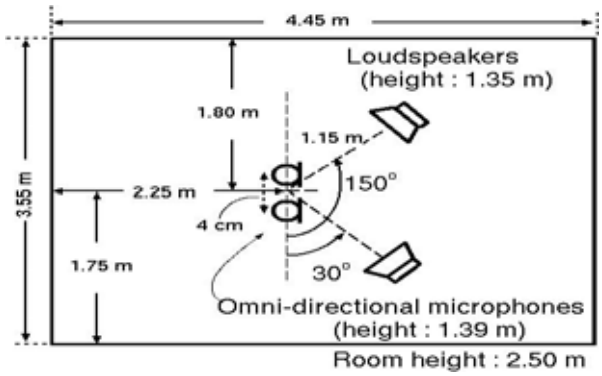

Figure 4. Room layout used for experiments

As original speech, we used Japanese sentences of 3 seconds spoken by male and female speakers. We made observation signals with (1) and investigated four combinations of speakers. The STFT frame size was 512, and the frame shift was 256 at a sampling rate of 8 kHz. To solve the permutation problem of frequency-domain ICA, we employed the direction of arrival and correlation approach [7], and to solve the scaling problem we used the minimum distortion principle [17]. For numerical analysis, we arranged four data sets: female and female (f & f), two types of female and male (f & m), and male and male (m & m). As an evaluation measure, we used the signal-to-interference ratio (SIR) as a separation performance measure:

$$\mathrm{SIR}_i = 10\log_{10}\frac{\sum_n y_{ii}^2(n)}{\sum_{k\neq i}(\sum_n y_{ik}(n))^2}, \qquad (12)$$

$$y_{ik}(n) = \sum_l u_{ik}(l)s_k(n-l), \quad u_{ik}(l) = \sum_{j=1}^{M}\sum_{t=0}^{L-1} w_{ij}(t)h_{jk}(l-t),$$

$w_{ij}(t)$: inverse discrete Fourier transform of ij component of separation matrix $\mathbf{W}(f)$,
L: STFT frame size.

5.2 Results

Results using Methods 1, 2, and 3, conventional *tanh*, and GGD-based modeling methods with the above four types of data sets are summarized in Table 1. The SIR for GGD was the value obtained for the best SIR for $\beta \in [0.5, 1.0]$. For conventional nonlinear functions, the GGD-based modeling method was slightly better than *tanh*. For the proposed Pearson approach, we obtained maximum improvement in separation performance of around 3 dB better than conventional *tanh* and around 2 dB better than conventional GGD.

Method 3, combining Methods 1 and 2, showed better separation performances than that by using only Method 2, even though the mean parameters used in Method 3 were first estimated by using only two data combinations. The mean parameters were applied to all data combinations, including the data that were not used to estimate the mean parameters. In fact, the SIR values sometimes became unstable in the calculation applying only Method 2. Therefore, Method 3's results suggest that using the mean parameter values pre-estimated by Method 1 at high frequencies did not lead to

instability and that at low frequencies the parameters estimated with data moments worked well.

Furthermore, we considered the computational time needed to perform these methods. We obtained the results using Matlab® profile report and summarize them in Table 2. The CPU clock speed was 594 MHz. Methods applying conventional nonlinear functions to the score function were faster than Methods 1 and 2. In particular, Method 1 took about two hours because the calculation algorithm included a grid search to obtain optimized parameters. By reducing this optimization procedure, Method 3 could work at a reasonable computation speed, thus improving performance.

Table 1: SIR (dB) values for conventional nonlinear functions and three proposed methods

	tanh	GGD	Meth. 1	Meth. 2	Meth. 3
m & m	17.71	17.58	17.00	17.54	18.12
f&m 1	14.99	15.89	15.99	16.38	16.42
f&m 2	15.92	17.22	18.61	17.08	18.38
f & f	17.27	17.76	20.12	18.62	19.43

Table 2: Computational time required to perform BSS

	Methods	Time [s]
conventional nonlinear function	*tanh*	10.38
	GGD	18.09
Proposed Pearson system approach	Meth. 1	6888.66
	Meth. 2	56.11
	Meth. 3	13.07

6. CONCLUSION

To estimate the frequency-domain separation matrix for ICA, we proposed a practical parametric Pearson distribution system. This system detects a score function for the source distribution at each frequency. We first confirmed the efficiency of three methods developed to apply the Pearson system to frequency-domain speech BSS under blind conditions. As the first method, unknown parameters were estimated to minimize the cross-correlation of the separation matrix. In the second method, we directly calculated the transform formulae based on κ discrimination. The third method was a combination of these two methods. The proposed approach significantly improved separation performance compared with conventional *tanh* and GGD-based modeling approaches. Furthermore, the combined method showed better performance than applying the individual methods, and its computational speed was reasonable.

REFERENCES

[1] S. Haykin, ed. Unsupervised Adaptive Filtering (Volume I: Blind Source Separation), John Wiley & Sons, 2000.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.

[3] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Frequency-domain blind source separation," in Speech Enhancement, eds. J. Benesty, S. Makino and J. Chen, Springer, 2005.

[4] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," Neurocomputing, vol. 22, pp. 21-34, 1998.

[5] L. Parra and C. Spence, "Convolutive blind separation of nonstationary sources," IEEE Trans. SAP. vol. 8, no. 3, pp. 320-327, May 2000.

[6] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency domain blind source separation," IEICE Trans. Fundamentals, vol. E86-A, no. 3, pp. 590-596, March 2003.

[7] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," IEEE Trans. SAP. vol. 12, no. 5, pp. 530-538, Sept. 2004.

[8] K. Kokkinakis and A.K. Nandi, "Multichannel speech separation using adaptive parametrization of source PDFs," eds. C.G. Puntonet and A. Prieto, ICA 2004, LNCS 3195, Springer-Verlag Berlin Heidelberg, pp. 486-493, 2004.

[9] R. Prasad, H. Saruwatari, and K. Shikano, "Blind separation of speech by fixed-point ICA with source adaptive negentropy approximation," IEICE Trans. Fundamentals, vol. E88-A, no. 7, July 2005.

[10] K. Pearson, "Memoir on skew variation in homogeneous material," Philos. Trans. Roy. Soc. A, vol. 186, pp. 343-414, 1895.

[11] J. Karvanen, J. Eriksson, and V. Koivunen, "Pearson system based method for blind separation," Proc. the Second International Workshop on ICA and BSS, pp. 585-590, 2000.

[12] R. Boscolo, H. Pan, and V. P. Roychowdhury, "Independent component analysis based on nonparametric density estimation," IEEE Trans. NN., vol. 15, no. 1, pp. 55-65, Jan. 2004.

[13] Y. Nagahara, "Non-Gaussian filter and smoother based on the Pearson distribution system," J. Time Ser. Anal, vol. 24, no. 6, pp. 721-738, 2003.

[14] Y. Nagahara, "The PDF and CF of Pearson type IV distributions and the ML estimation of the parameters," Stat. Prob. Letters, vol. 43, pp. 251-264, 1999.

[15] Y. Nagahara, "A method of simulating multivariate nonnormal distributions by the Pearson distribution system and estimation", Comp. Stat. Data. Anal, vol. 47, issue 1, pp. 1-29, 2004.

[16] S. Amari, "Natural gradient works efficiently in learning," Neural Computation, vol. 10, pp. 251-276, 1998.

[17] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," Proc. ICA2001, pp. 722-727, Dec. 2001.