A NEW MUTUAL INFORMATION MEASURE FOR INDEPENDENT COMPONENT ALALYSIS

Jen-Tzung Chien [a], *Hsin-Lung Hsieh* [a] *and Sadaoki Furui* [b]

[a] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan
[b] Department of Computer Science, Tokyo Institute of Technology, 2-12-1, O-okayama, Meguro-ku, Tokyo, Japan
{chien, hlhsieh}@chien.csie.ncku.edu.tw, furui@cs.titech.ac.jp

ABSTRACT

Independent component analysis (ICA) is a popular approach for blind source separation (BSS). In this study, we develop a new mutual information measure for BSS and unsupervised learning of acoustic models. The underlying concept of ICA unsupervised learning algorithm is to demix the observations vectors and identify the corresponding mixture sources. These independent sources represent the specific speaker, gender, accent, noise or environment, etc, embedded in acoustic models. The novelty of the proposed ICA is to derive a new metric of mutual information for measuring the dependence among mixture sources. We focus on building this metric based on the *Jensen's inequality*, which is illustrated to use smaller number of iterations in finding the demixing matrix compared to other types of mutual information. We present a parametric ICA using the generalized Gaussian distribution to characterize the non-Gaussianity of model parameters. Also, a nonparametric ICA is established by using the Parzen window based distribution. In the experiments on BSS and noisy speech recognition, we demonstrate the effectiveness of the proposed Jensen ICA compared to FastICA and other nonparametric ICA.

Index Terms— Independent component analysis, mutual information, Jensen's inequality, speech recognition

1. INTRODUCTION

Independent component analysis (ICA) [1][6] is a widely accepted mechanism in solving blind source separation (BSS) problem. In the BSS problem, a set of observations is given while the underlying source information is hidden. The mixing weights of the individual sources are unknown. The BSS problem is aimed to identify the source signals and/or the mixing weights so as to separate these information sources in signal domain, feature domain or model domain [5]. The basic assumptions in the ICA method have the statements that the source signals are mutually independent and non-Gaussian distributed. Using the ICA, an $M \times 1$ observation vector $\mathbf{x}$ is modeled from M statistically independent sources $\mathbf{s}$ by $\mathbf{x} = A\mathbf{s}$ where A is a square and invertible mixing matrix of size $M \times M$. The elements of $\mathbf{s} = [s_1, \cdots, s_M]^T$ are linearly mixed to the observations $\mathbf{x} = [x_1, \cdots, x_M]^T$ by the transformation matrix $\mathbf{A}$. We are engaged in an inverse problem and find the source signals by $\mathbf{s} = W\mathbf{x}$ where W is a demixing matrix. In this study, we highlight a new ICA procedure for separating audio signals as well as clustering hidden Markov models (HMMs). In general, if the data is properly transformed or demixed, the resulting components can be grouped into clusters where the elements are dependent in the same cluster and independent to the variables in different clusters [2]. We concern the *inter-cluster independence* and the *intra-cluster dependence* for building a new ICA clustering algorithm.

In the ICA procedure, it is critical to estimate the demixing matrix W by optimizing the metric of independence. The metrics of likelihood function, negentropy, kurtosis and mutual information have been successfully applied in developing ICA algorithms [6]. Minimization of the mutual information among sources is viewed as a meaningful information theoretic solution. Conventionally, the ICA method using minimum mutual information (MMI) was constructed by Shannon's mutual information where the difference between the marginal entropy and the joint entropy of different information sources was accumulated. One difficulty of using MMI was the estimation of marginal entropy. Comon approximated the output marginal probability density function by applying the truncated polynomial expansion [6]. Alternatively, the MMI method proposed by Xu et al. [12] prevented the polynomial expansion through approximating the Kullback-Leibler divergence using the Cauchy-Schwartz inequality. ICA estimation was performed by using the Parzen window based distribution. Also, Boscolo et al. [3] proposed an ICA algorithm where the mutual information between the reconstructed signals was minimized. Using nonparametric kernel density technique, this algorithm was carried out by estimating the unknown probability density functions of the source signals and finding the unknown mixing matrix. In acoustic feature analysis, the ICA method was used to project MFCC features and discover that the first MFCC feature was associated with the speaker's gender and the second feature was associated with the speaker's accent [8]. In this paper, we present a new ICA algorithm using the mutual information based objective function derived by the Jensen's inequality. The source densities are modeled by the generalized Gaussian function [4][9] for modeling the non-Gaussian structure. A wide class of distributions including uniform, Gaussian, Laplacian and other sub-Gaussian and super-Gaussian densities can be characterized. We develop the Jensen ICA (J-ICA) algorithm by using not only the parametric source density but also the non-parametric source density. This algorithm is examined by the experiments of BSS and speech recognition.

2. MUTUAL INFORMATION FOR ICA

Objective function is one of the most important issues in ICA implementation. Mutual information is known as a popular metric of measuring the dependence in the observed variables. There are several measures of mutual information existing in the literature.

2.1 Mutual Information Measures

Assume we have two continuous variables x_1, x_2 with marginal distributions $p(x_1)$, $p(x_2)$ and joint distribution $p(x_1, x_2)$, the Shannon's mutual information of x_1, x_2 is defined by [11]

$$I_S(x_1, x_2) = H(p(x_1)) + H(p(x_2)) - H(p(x_1, x_2))$$
$$= \iint p(x_1, x_2) \log \frac{p(x_1, x_2)}{p(x_1)p(x_2)} dx_1 dx_2 \,, \qquad (1)$$

where $H(\cdot)$ is a Shannon's entropy and $I_S(x_1,x_2)\geq 0$ with equality if and only if x_1, x_2 are independent. Also, $I_S(x_1,x_2)$ is known as a Kullback-Leibler divergence between $p(x_1,x_2)$ and $p(x_1)p(x_2)$. In [12], the quadratic mutual information based on the Euclidean distance (ED) and the Cauchy-Schwartz (CS) inequality was proposed as

$$I_{\mathrm{ED}}(x_1,x_2)=\iint (p(x_1,x_2)-p(x_1)p(x_2))^2\,dx_1dx_2\ , \qquad (2)$$

$$I_{\mathrm{CS}}(x_1,x_2)=\log\frac{\iint p(x_1,x_2)^2\,dx_1dx_2\iint p(x_1)^2p(x_2)^2\,dx_1dx_2}{(\iint p(x_1,x_2)p(x_1)p(x_2)dx_1dx_2)^2}\ . \qquad (3)$$

It is obvious that $I_{\mathrm{ED}}(x_1,x_2)\geq 0$ and $I_{\mathrm{CS}}(x_1,x_2)\geq 0$. In these two measures, the equalities hold if and only if x_1 and x_2 are independent. The resulting mutual information measures are meaningful for ICA implementation. In [10], a Jensen-Shannon (JS) divergence measure was presented from the viewpoint of decision theory. The Jensen's inequality and the Shannon entropy were combined. This measure was illustrated to provide the lower bound and the upper bound for Bayes probability of classification error. In this study, we are motivated to develop a new measure of mutual information based on the Jensen's inequality. The so-called Jensen ICA algorithm can be derived.

2.2 A New Measure Based on Jensen's Inequality

Let $f(\cdot)$ denote a convex function. Considering the joint distribution $p(x_1,x_2)$ and the product of marginal distributions $p(x_1)p(x_2)$ as inputs of convex function, the Jensen's inequality for two functions is written by

$$\begin{aligned} &f(\lambda_a p(x_1,x_2)+\lambda_b p(x_1)p(x_2)) \\ &\leq \lambda_a f(p(x_1,x_2))+\lambda_b f(p(x_1)p(x_2)), \end{aligned} \qquad (4)$$

where convex coefficients meet the constraints $\lambda=\{\lambda_a,\lambda_b\}\geq 0$ and $\lambda_a+\lambda_b=1$. As we know, the function $-\log(\cdot)$ is a popularly adopted convex function. We can derive the Jensen mutual information for measuring the dependence between variables x_1 and x_2 by

$$\begin{aligned} I_J(x_1,x_2,\lambda)=&\iint(\log(\lambda_a p(x_1,x_2)+\lambda_b p(x_1)p(x_2)) \\ &-\lambda_a\log(p(x_1,x_2))-\lambda_b\log(p(x_1)p(x_2)))dx_1dx_2\ . \end{aligned} \qquad (5)$$

It is straightforward to show that $I_J(x_1,x_2,\lambda)\geq 0$ and equality holds if and only if two variables x_1 and x_2 are independent, i.e. $p(x_1,x_2)=p(x_1)p(x_2)$. In developing ICA algorithm, it is reasonable to establish the Jensen mutual information by equally treating the contributions of $p(x_1,x_2)$ and $p(x_1)p(x_2)$, namely setting $\lambda_a=\lambda_b=0.5$. The mutual information measure can be generalized to $I_J(x_1,\cdots,x_M)$ for expressing the dependence measure for observation vector $\mathbf{x}=[x_1,\cdots,x_M]^T$.

To illustrate the relations among the mutual information measures of Shannon, Euclidean distance, Cauchy-Schwartz and Jensen, we use a simple case with two discrete random variables X_1, X_2. The joint probability of two events A, B is shown in Figure 1. If we consider the case of a marginal probability of X_1 with $P_{X_1}(A)=0.7$, $P_{X_1}(B)=0.3$ and the joint probabilities $P_{X_1,X_2}(A,A)$ and $P_{X_1,X_2}(B,A)$ with values in the ranges from 0 to 0.7 and 0 to 0.3, respectively, the mutual information measures can be calculated according to different values of $P_{X_1,X_2}(A,A)$ and $P_{X_1,X_2}(B,A)$. Figure 2 shows the mutual information measure versus the probability model $P_{X_1,X_2}(A,A)$ in the case of $P_{X_2}(A)=0.5$ and $P_{X_2}(B)=0.5$. We can see that different mutual information measures reach the same minimum point where the condition of independence between X_1 and X_2 happens. Also, among these four measures, the flattest curve and the steepest curve are attained by the Euclidean distance mutual information and the Jensen mutual information, respectively. This evaluation implies that the implementation of ICA based on the Jensen mutual information achieves the minimum value of mutual information efficiently. Comparably, a small number of iterations are needed in estimating Jensen ICA demixing matrix owing to the close relation between probability model and demixing matrix.

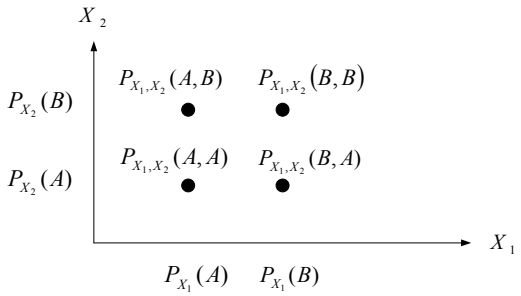

Figure 1 Probability distribution of two random variables

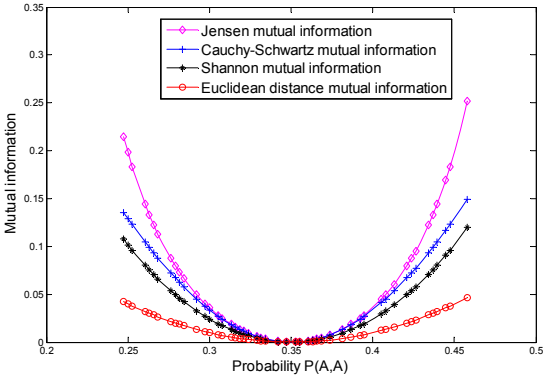

Figure 2 Comparison of four mutual information measures

3. JENSEN ICA ALGORITHM

In what follows, we address the Jensen ICA (J-ICA) algorithm based on the metric of Jensen mutual information. The basic assumption of ICA is that the sources should be non-Gaussian distributed. Here, we develop a parametric J-ICA and a nonparametric J-ICA by using the generalized Gaussian density and the Parzen window based density, respectively.

3.1 Parametric J-ICA Algorithm

First, the generalized Gaussian function [4] is applied to characterize the super-Gaussian or sub-Gaussian distribution for source signal. The multivariate distribution of $\mathbf{y}\in\Re^M$ and the univariate distribution of a component y_m are defined by

$$p(\mathbf{y}\,|\,\boldsymbol{\mu},\Sigma,\beta)=\frac{\omega(\beta)^M}{|\Sigma|^{1/2}}\exp[-c(\beta)\,|\,(\mathbf{y}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})\,|^{1/(1+\beta)}]\ , \qquad (6)$$

$$p(y_m \mid \mu_m, \sigma_m^2, \beta_m) = \frac{\omega(\beta_m)}{\sigma_m} \exp\left[-c(\beta_m) \left| \frac{y_m - \mu_m}{\sigma_m} \right|^{2/(1+\beta_m)} \right], \quad (7)$$

where $c(\beta) = \left[\frac{\Gamma\left[\frac{3}{2}(1+\beta)\right]}{\Gamma\left[\frac{1}{2}(1+\beta)\right]} \right]^{1/(1+\beta)}$, $\omega(\beta) = \frac{\Gamma\left[\frac{3}{2}(1+\beta)\right]^{1/2}}{(1+\beta)\Gamma\left[\frac{1}{2}(1+\beta)\right]^{3/2}}$ and $\Gamma(\cdot)$ is a Gamma function. These distributions are governed by means $(\mathbf{\mu}, \mu)$ and variances (Σ, σ^2). The parameter β is a measure of kurtosis. The case $\beta = 0$ represents the standard Gaussian distribution. In cases of $\beta = 1$ and $\beta = -1$, the distribution becomes Laplacian distribution and uniform distribution over the unit interval, respectively. As $\beta \to \infty$, the distribution becomes a delta function centered at zero. We can calculate β through maximum *a posteriori* estimation where Gamma prior for β was specified [4]. Since the preprocessing stages of mean removal and whitening process are performed in ICA procedure, we practically assume zero mean and unit variance in generalized Gaussian distribution.

Given an observation sequence $X = \{\mathbf{x}_t\} = \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$, we are estimating a demixed matrix $W = [\mathbf{w}_1^T, \cdots, \mathbf{w}_M^T]^T$ which minimizes the Jensen mutual information $I_J(x_{t1}, \cdots, x_{tM})$ accumulated from X. Here, the mth row of W is denoted by $\mathbf{w}_m$ and the corresponding demixed signal is given by $y_{tm} = \mathbf{w}_m \mathbf{x}_t$. The parametric J-ICA (PJ-ICA) objective function of demixed data $I_{\text{PJ-ICA}}(X, W)$ is yielded by

$$\sum_{t=1}^{T} \left\{ \begin{array}{l} \log\left[\frac{\omega(\beta)^M}{|\Sigma|^{1/2}} \exp[-c(\beta) |(W\mathbf{x}_t - \mathbf{\mu})^T \Sigma^{-1} (W\mathbf{x}_t - \mathbf{\mu})|^{1/(1+\beta)}] \right. \\ \left. \quad + \prod_{m=1}^{M} \frac{\omega(\beta_m)}{\sigma_m} \exp\left[-c(\beta_m) \left| \frac{\mathbf{w}_m \mathbf{x}_t - \mu_m}{\sigma_m} \right|^{2/(1+\beta_m)} \right] \right] - \log 2 \\ -\frac{1}{2} \left[\begin{array}{l} -c(\beta) |(W\mathbf{x}_t - \mathbf{\mu})^T \Sigma^{-1} (W\mathbf{x}_t - \mathbf{\mu})|^{1/(1+\beta)} + \log \frac{\omega(\beta)^M}{|\Sigma|^{1/2}} \\ + \sum_{m=1}^{M} \left(-c(\beta_m) \left| \frac{\mathbf{w}_m \mathbf{x}_t - \mu_m}{\sigma_m} \right|^{2/(1+\beta_m)} + \log \frac{\omega(\beta_m)}{\sigma_m} \right) \end{array} \right] \end{array} \right\} \quad (8)$$

By taking the gradient of $I_{\text{PJ-ICA}}(X, W)$ with respect to the demixing matrix W, we derive the ICA solution to demixing matrix by the iterative procedure

$$W^{(n+1)} = W^{(n)} - \eta \nabla_{W^{(n)}} I_{\text{PJ-ICA}}(X, W^{(n)}), \quad (9)$$

where n is an iteration index and η is the learning rate. The resulting algorithm is called the parametric J-ICA because the independent source signals are modeled by the parametric distribution.

3.2 Nonparametric J-ICA Algorithm

One popular alternative to avoiding the assumption of Gaussian distribution in ICA is to adopt the nonparametric approach. The nonparametric density using Parzen window estimation is attractive for data modeling because the distribution shape can be flexibly generated by a data-driven way. The nonparametric ICA has been successfully developed in [3][5]. This paper presents a new nonparametric ICA based on the Jensen mutual information. The demixed signals $\mathbf{y} = [y_1, \cdots, y_M]^T$ are characterized either by univariate distribution or multivariate distribution

$$p(\mathbf{y}) = \frac{1}{Th^M} \sum_{i=1}^{T} \psi\left(\frac{\mathbf{y} - \mathbf{y}_i}{h} \right) = \{p(y_m)\} = \left\{ \frac{1}{Th} \sum_{i=1}^{T} \varphi\left(\frac{y_m - y_{im}}{h} \right) \right\} \quad (10)$$

We adopt Gaussian kernels with the univariate form and the multivariate form expressed by $\varphi(u) = (2\pi)^{-1/2} e^{-u^2/2}$ and $\psi(\mathbf{u}) = (2\pi)^{-M/2} e^{-0.5 \cdot \mathbf{u}^T \mathbf{u}}$, respectively. Given the demixed component $y_{tm} = \mathbf{w}_m \mathbf{x}_t$ and the demixed signals $\mathbf{y}_t = W\mathbf{x}_t$, the non-parametric J-ICA (NJ-ICA) objective function $I_{\text{NJ-ICA}}(X, W)$ is generated by

$$\sum_{t=1}^{T} \left\{ \begin{array}{l} \log\left[\frac{1}{Th^M} \sum_{i=1}^{T} \psi\left(\frac{W(\mathbf{x}_t - \mathbf{x}_i)}{h} \right) + \prod_{m=1}^{M} \frac{1}{Th} \sum_{i=1}^{T} \varphi\left(\frac{\mathbf{w}_m(\mathbf{x}_t - \mathbf{x}_i)}{h} \right) \right] - \log 2 \\ -\frac{1}{2} \left[\begin{array}{l} \log\left(\frac{1}{Th^M} \sum_{i=1}^{T} \psi\left(\frac{W(\mathbf{x}_t - \mathbf{x}_i)}{h} \right) \right) \\ + \log\left(\prod_{m=1}^{M} \frac{1}{Th} \sum_{i=1}^{T} \varphi\left(\frac{\mathbf{w}_m(\mathbf{x}_t - \mathbf{x}_i)}{h} \right) \right) \end{array} \right] \end{array} \right\} \quad (11)$$

The nonparametric J-ICA algorithm is then derived by taking the gradient of $I_{\text{NJ-ICA}}(X, W)$ with respect to W and substituting this gradient into the learning rule in (9). In general, it is straightforward to find the gradients of objective function in (8) and (11) with respect to W. We don't show these equations.

4. EXPERIMENTS

In the experiments, we realized the proposed parametric and nonparametric J-ICA algorithms for blind source separation and speech recognition. In BSS, we used a speech signal and a music signal sampled from http://sound.media.mit.edu/ica-bench/. A 2×2 mixing matrix $A = [[0.31\ 0.63]^T [0.94\ 0.31]^T]$ was specified to mix the source signals as shown in Figure 3. For comparative study, we also carried out the FastICA algorithm [7] and the Boscolo's nonparametric ICA (denoted by BN-ICA) [3] which were based on the metrics of negentropy and MMI, respectively. In ICA implementation, the initial demixing matrix $W^{(1)}$ was randomly selected. The convergence condition was set to be $|W^{(n+1)}| - |W^{(n)}| \le 0.8$. We show the waveforms of the demixed signals in Figure 3 and report the signal-to-interference ratios (SIRs) of different ICAs in Table 1. The parametric and nonparametric J-ICA methods obtain good demixed signals. In the comparison of SIR, J-ICA algorithms achieve improvement compared to FastICA and BN-ICA. Two J-ICA methods perform comparably. In the preliminary evaluation, we did find that J-ICA used smaller number of iterations than ICA by other mutual information.

	FastICA	BN-ICA	PJ-ICA	NJ-ICA
Music	-1.372	-1.659	-0.734	-0.793
Speech	-1.039	-0.802	-0.092	0.020

Table 1 Comparison of SIR (dB) for different methods

We also evaluated ICA performance in clustering of HMMs for noisy speech recognition on AURORA2 database. The set A with four noises (subway, babble, car and exhibition hall), six

SNRs (-5, 0, 5, 10, 15 and 20 dB) and a clean condition was adopted. Feature vector consisted of twelve MFCCs and one log energy and their first derivatives. Conventionally, the multi-conditional training was performed by putting all training data for four noise types and five SNRs for HMM training. However, there were several independent sources existed in the model parameters. We attempted to explore the independent sources of noise conditions and build several clusters of HMM parameters. In the training phase, we first estimated the HMM parameters for individual noise types and SNRs. Twenty supervectors $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_{20}\}$ of HMM means for different dimensions, states and words were generated. We had 26 features, 16 states and 11 words in HMM modeling. Each observed vector $\mathbf{x}_t$ had the dimension of 26*16*11. We performed ICA transformation of HMM mean supervectors X and applied K-means algorithm to find independent HMM clusters. In the test phase, we used the first ten frames of each sentence and picked up the closest HMM cluster with the highest likelihood score. This HMM cluster was adopted to recognize the input sentence.

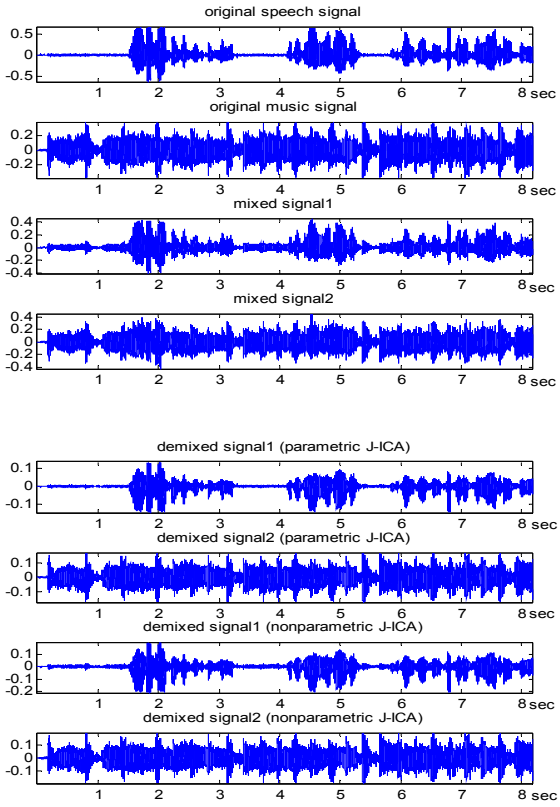

Figure 3 Waveforms of original signals, mixed signals and demixed signals using two J-ICA methods

In Table 2, we report the recognition accuracies averaged over test data in different noise conditions of set A. Here, the baseline system is the case of without ICA transformation. Also, we implement HMM clustering using the principal component analysis (PCA) method for comparison. One cluster means that no HMM clustering is applied. We investigate the cases when the number of independent sources are two, three and four, and only show the results of parametric J-ICA. In this preliminary evaluation, the improvement of HMM clustering is not significant. One important reason was the incorrectness of selecting HMM cluster. In general, the best performance 81.48% was achieved by using the parametric J-ICA in case of four HMM clusters.

Table 2 Recognition accuracies (%) of using different methods

	1 cluster	2 clusters	3 clusters	4 clusters
Baseline	80.4	80.6	81.3	81.3
PCA		81.3	80.8	79.7
FastICA		79.8	81.4	81.4
PJ-ICA		80.9	81.4	81.5

5. CONCLUSIONS

We have presented a new mutual information measure for developing an ICA algorithm for BSS and speech recognition. Some existing mutual information measures were surveyed. A Jensen mutual information measure was derived as an objective for convex optimization. This mutual information was illustrated to be the steepest among several measures in terms of the variations of probability models. By considering the parametric and the nonparametric distributions for independent sources, we exploited the parametric J-ICA as well as the nonparametric J-ICA for finding the demixing matrix. In the experiments of separating music and speech signals, J-ICA algorithms obtained quite good performance. In the application of noisy speech recognition, we got slight improvement compared to other methods. In the future, we are performing detailed evaluation of convergence speed and exploring ICA algorithm for other issues of speech recognition.

6. ACKNOWLEDGMENT

The authors thank the helpful discussion with Prof. Koichi Shinoda at Tokyo Institute of Technology.

7. REFERENCES

[1] H. Attias, "Independent factor analysis", *Neural Computation*, vol. 11, pp. 803-851, 1999.

[2] F. R. Bach and M. I. Jordan, "Finding clusters in independent component analysis", *Proc. of ICA*, 2003.

[3] R. Boscolo, H. Pan and V. P. Roychowdhury, "Independent component analysis based on nonparametric density estimation", *IEEE Transactions on Neural Networks,* vol. 15, no. 1, 2004.

[4] G. Box and G. Tiao, *Bayesian inference in statistical analysis*, John Wiley and Son, 1973.

[5] 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.

[6] P. Comon, "Independent component analysis, a new concept?", *Signal Processing*, vol. 36, pp. 287-314, 1994.

[7] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis", *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626-634, 1999.

[8] J.-H. Lee, H.-Y. Jung, T.-W. Lee and S. Y. Lee, "Speech feature extraction using independent component analysis", *Proc. of ICASSP*, vol. 3, pp. 1631-1634, 2000.

[9] T.-W. Lee and M. S. Lewicki, "The generalized Gaussian mixture model using ICA", *Proc. of ICA*, pp. 239-244, 2000.

[10] J. Lin, "Divergence measures based on the Shannon entropy", *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145-151, 1991.

[11] C. E. Shannon, "A mathematical theory of communication", *Bell System Technical Journal*, vol. 27, pp. 379-423, 1948.

[12] D. Xu, J. Principle, J. Fisher and H.-C. Wu, "A novel measure for independent component analysis (ICA)", *Proc. of ICASSP*, vol. 2, pp. 1161-1164, 1998.