A New Independent Component Analysis for Speech Recognition and Separation

Jen-Tzung Chien, Senior Member, IEEE, and Bo-Cheng Chen

Abstract—This paper presents a novel nonparametric likelihood ratio (NLR) objective function for independent component analysis (ICA). This function is derived through the statistical hypothesis test of independence of random observations. A likelihood ratio function is developed to measure the confidence toward independence. We accordingly estimate the demixing matrix by maximizing the likelihood ratio function and apply it to transform data into independent component space. Conventionally, the test of independence was established assuming data distributions being Gaussian, which is improper to realize ICA. To avoid assuming Gaussianity in hypothesis testing, we propose a nonparametric approach where the distributions of random variables are calculated using kernel density functions. A new ICA is then fulfilled through the NLR objective function. Interestingly, we apply the proposed NLR-ICA algorithm for unsupervised learning of unknown pronunciation variations. The clusters of speech hidden Markov models are estimated to characterize multiple pronunciations of subword units for robust speech recognition. Also, the NLR-ICA is applied to separate the linear mixture of pronunciations of subword units for robust speech recognition.

Index Terms—Acoustic modeling, blind source separation (BSS), independent component analysis (ICA), nonparametric likelihood ratio (NLR), pronunciation variation, speech recognition, unsupervised learning.

I. INTRODUCTION

INDEPENDENT component analysis (ICA) [11] has been attracting the researchers in societies of signal processing and neural networks for many years. It is because that the ICA principle is essential to deal with fundamental issues of blind source separation, blind deconvolution, feature extraction, unsupervised learning, data analysis, and compression. Many applications have been developed for text document clustering [21], facial feature representation [40], image enhancement, financial data analysis, and neurobiological signal processing [17]. For the applications on speech processing, ICA was employed for extraction of salient speech features [25], analysis of speaker variations [18], separation of multiple speakers [24], and cancellation of reverberation in speech signals [4]. ICA was also applied to establish basis functions for individual speakers for speaker recognition [20]. In general, ICA network is a higher-order and nonlinear extension of principal component analysis (PCA), which enables the network to separate statistically independent components [4], [29]. It has the capability of exploring latent factors in random signals with strong relations to factor analysis (FA) [2], [16], [19]. The main idea of ICA focuses on finding the independent sources from mixed data rather than reducing the feature dimension toward maximum expression direction via PCA and extracting the Gaussian common factors embedded in unknown data via FA.

Basically, ICA is powerful to capture the unknown structure of data by minimizing the statistical dependence of different components. It is meaningful to use ICA to solve blind source separation (BSS) problem. BSS aims to estimate a \( d \times d \) demixing matrix \( W \), which separates the mixed signal \( \mathbf{x} = [x^{(1)}, \ldots, x^{(d)}]^T \) and recovers its original independent components \( \mathbf{s} = [s^{(1)}, \ldots, s^{(d)}]^T \). The observed signal \( \mathbf{x} \) was mixed with an unknown matrix \( \mathbf{A} \), \( \mathbf{x} = \mathbf{As} \). The key idea of estimating ICA model is to maximize the non-Gaussianity for achieving the independence of sources [17]. Traditionally, the high-order statistics and information-theoretic criteria were exploited to measure the non-Gaussianity or independence. For example, the high-order statistics using absolute value of kurtosis was maximized to find independent components. But, kurtosis was sensitive to outlier data [17]. Also, the measurements using negentropy, likelihood function, and mutual information were popular to construct ICA model. The mutual information between the transformed sources was minimized to find the demixing matrix. Such optimization was shown to be equivalent to maximum likelihood and maximum negentropy principles under some assumptions [17]. In addition to the BSS problem, ICA was useful for unsupervised learning [30]. In [33], ICA was used to separate a multivariate distribution into a mixture of independent components. Because ICA was inherent in exploring data structure, it was also attractive to derive decision tree-based unsupervised classification using ICA [32]. In [26], ICA mixture models were presented for unsupervised classification of non-Gaussian classes. A local ICA model was formed to represent the nonlinear data distributions via merging linear ICA and k-means clustering [22].

Actually, due to the powerfulness of ICA on solving fundamental problems, we can apply ICA methods for speech feature extraction as well as acoustic modeling for speech recognition. In [18], the first and the second independent components of speech features were shown containing sexual and accent information, respectively, which were used to improve speech recognition performance. In acoustic modeling for large vocabulary continuous speech recognition (LVCSR), we characterize context-dependent subword unit using speech data from a pool of speakers with varying genders, accents, ages, emotions, etc. The top-down clustering based on decision tree state
tying was exploited to estimate hidden Markov model (HMM) parameters handling the contextual variations [3], [9]. The use of word- and syllable-level contextual modeling was presented for conversational speech recognition [36]. Also, the variations from inter- and intraspeakers in acoustic modeling were tackled through speaker clustering algorithms [23], [31] where a set of speaker-clustered models was estimated to represent the statistics from different speaker sources. More generally, we can compensate the pronunciation variations from changing contexts and speakers through the construction of multiple pronunciations in the dictionary [14] or unsupervised learning of unknown pronunciation sources [37]. Attractively, ICA provides a meaningful mechanism of finding unknown sources/mixtures for acoustic modeling. Each source represents a specific kind of pronunciation variation. Using ICA, we can demix the acoustic data for data clustering such that the parameters corresponding to individual sources can be calculated. As a result, the desirable LVCSR performance is achieved.

In this paper, we are motivated to develop a new ICA model for unsupervised learning of acoustic models. The clustered HMM parameters are established for speech recognition in presence of pronunciation variations. Importantly, we survey a series of ICA approaches and propose a statistical approach to construct ICA model using the hypothesis test principle. We are testing the hypothesis whether the transformed set of variates is independent or not. The hypothesis is verified when the test statistics in a form of likelihood ratio exceeds the specified significant level. Accordingly, the likelihood ratio of independence hypothesis to dependence hypothesis is derived as an objective (contrast) function to estimate demixing matrix. We maximize the likelihood ratio function, or the confidence for independence, to realize ICA. In multivariate statistical analysis [1], [15], the test of independence was presented under the assumption of Gaussian distribution, which is undesirable to resolve ICA problem. Instead of assuming Gaussian, we discover a nonparametric approach where the kernel density functions are adopted to approximate the data distributions. A new nonparametric likelihood ratio (NLR) criterion is introduced to build ICA model. In this study, we carry out the nonparametric likelihood ratio for clustering of speech signals for hidden Markov modeling. The proposed approach is also applied to blind separation of speech and music signals. We achieved desirable performance for both applications in terms of speech recognition rates and signal-to-interference ratios. In what follows, we survey and compare several objective functions to be optimized for finding demixing matrix. In Section III, the new NLR-based objective function is presented. We describe the derivation of NLR using ideas of hypothesis testing and nonparametric approaches. Next, Section IV mentions the applications of ICA for speech separation and recognition and the corresponding experimental setups. Several sets of evaluation on BSS of speech and music signals and unsupervised learning of HMMs are reported. Finally, the conclusions drawn from this study are given in Section V.

II. SURVEY OF OBJECTIVE FUNCTIONS FOR ICA

ICA problem is essentially comprised of the following three parts [12]:

1) formulation of an objective function measuring the degree of data independence;
2) determination of objective function \( \Gamma(X, W) \) from \( T \) observations \( X = \{x_1, \ldots, x_T\} \) using a demixing matrix \( W \);
3) optimization of objective function.

There is no doubt that the design of objective function \( \Gamma(X, W) \) mostly expressing the degree of independence is critical for ICA problem. The optimal matrix \( \hat{W} \) is then estimated through maximizing \( \Gamma(X, W) \). The gradient descent algorithm can be applied for iterative learning of \( \hat{W} \).

\[
W^{(n+1)} = W^{(n)} + \Delta W^{(n)} = W^{(n)} - \eta \nabla_{W^{(n)}} \Gamma(X, W^{(n)}) \tag{1}
\]

where \( \eta \) is iteration index and \( \eta \) is learning rate. The gradient term \( \nabla_{W^{(n)}} \Gamma(X, W^{(n)}) \) with respect to \( W^{(n)} \) is calculated to determine the adjustment \( \Delta W^{(n)} \). Using the derived \( \hat{W} \), linear ICA model is constructed by demixing observed signal \( x_t \) by \( y_t = \hat{W} x_t \).

In data analysis procedure, we usually perform whitening transform as a preprocessing step. After transformation, the data distribution has been uncorrelated. If data is Gaussian distributed, it is not possible to derive \( \hat{W} \) for ICA. Also, according to the central limit theorem, the Gaussianity of random variables is increased by the linear transformation. The assumption of Gaussian distribution provides no additional information for finding \( \hat{W} \). Hence, the implementation of ICA toward maximizing the measure of independence becomes equivalent to that maximizing the non-Gaussianity [17]. The classical measure of non-Gaussianity is the fourth-order cumulant called kurtosis. In [5], the cumulative density function instead of probability density function was adopted to measure the non-Gaussianity where the generalized Gaussian family [20] was considered to represent data distribution. In [12], [28], the characteristic function was employed to measure the independence of random variables.

Further, the maximum-likelihood (ML) criterion is popular to serve as objective function for ICA [7], [17], [38]. The idea of ML is to estimate the mostly likely \( W_{ML} \) by maximizing the likelihood function from data \( X \). Because the unknown independent source \( s \) is related to mixed data \( x \) with \( x = As = W^{-1}s \), the transformed data \( y \) becomes independent, i.e., \( y = Wx = WW^{-1}s = s \). The ML objective function can be determined by accumulating log likelihoods from \( X = \{x_1, \ldots, x_T\} \)

\[
\Gamma_{ML}(X, W) = \sum_{t=1}^{T} \log p(x_t) = \sum_{t=1}^{T} \log \left( |\det W| p(s_t) \right) = T \log |\det W| + \sum_{t=1}^{T} d \log p(w_t x_t) \tag{2}
\]

where \( w_t \) is the \( i \)th row of matrix \( W \). Another objective function is introduced from a neural network viewpoint using the maximum-entropy (ME) criterion [4], [16], [39]. The underlying concept is to find \( W_{ME} \) by maximizing the output entropy of a neural network with nonlinear output \( z = g(y) \). It can be shown that \( W_{ME} \) is equivalent to \( W_{ML} \), when the nonlinearity is done
by a cumulative distribution function corresponding to the density \( g_i(y^{(i)}) = p(y^{(i)}) \)

\[
\Gamma_{\text{ME}}(X, W) = H(z) = -E[\log p(g(Wx))] \\
= E \sum_{i=1}^{d} \log p(y^{(i)}) \\
+ \log |\det W| - E[\log p(x)] \\
\propto T \log |\det W| + \sum_{i=1}^{d} \sum_{t=1}^{T} \log p(w_{it}) \\
= \Gamma_{\text{ML}}(X, W). \tag{3}
\]

In (3), there are two Jacobian terms due to the transformations \( z = g(y) \) and \( y = Wx \). This is also called infomax principle because ME criterion maximizes the information conveyed to system output.

More intuitively, the information-maximization approach can be fulfilled through the minimum mutual information (MMI) criterion \([6], [16], [17], [39]\). According to the concept of differential entropy, MMI-ICA algorithm derives the demixing matrix \( W_{\text{MMI}} \) by measuring the mutual information among the components of \( y \)

\[
I(y^{(1)}, \ldots, y^{(d)}) = \sum_{i=1}^{d} H(y^{(i)}) - H(y) \\
= \int p(y) \frac{\log \hat{p}(y)}{\hat{p}(y)} dy \\
= D(p(y) || \hat{p}(y)) \tag{4}
\]

which can be interpreted as a Kullback-Leibler divergence between the joint probability density \( p(y) \) and its independent version \( \hat{p}(y) = \prod_{i=1}^{d} p(y^{(i)}) \). This mutual information is equivalent to the negentropy if \( \{y^{(i)}, \ i = 1, \ldots, d\} \) are uncorrelated \([16], [17]\). Correspondingly, we express MMI objective function as the negative of mutual information to be maximized

\[
\Gamma_{\text{MMI}}(X, W) = -I(y^{(1)}, \ldots, y^{(d)}) \\
= H(x) + \log |\det W| + \sum_{i=1}^{d} E[\log p(w_{ix})] \\
= \Gamma_{\text{ML}}(X, W) \tag{5}
\]

where \( H(x) \) is a constant for estimating \( W \). We can see that MMI and ML criteria are the same. The MMI-ICA algorithm is derived using the gradient

\[
\nabla_W \Gamma_{\text{MMI}}(X, W) = \nabla_W \left\{ \log |\det W| + \sum_{i=1}^{d} E[\log p(w_{ix})] \right\} \\
= (W^{-1})^T - E[\phi(y)\phi(y)^T]. \tag{6}
\]

Here, \( \phi(y) = [\phi(y^{(1)}), \ldots, \phi(y^{(d)})]^T \) and \( \phi(y^{(i)}) = -p(y^{(i)})/p(y^{(i)}) \). In \([39]\), a positive-definite operator \( W^TW \) was introduced in gradient descent algorithm so as to obtain a computationally efficient MMI-ICA learning rule

\[
\Delta W = -\eta (I - E[\phi(y)\phi(y)^T]) W. \tag{7}
\]

In \([4]\), the nonlinear function \( \phi(y) = 2\tan^{-1}(y) \) was used when the density \( p(y^{(i)}) \) was chosen to be the derivative of a logistic function.

### III. Nonparametric Likelihood Ratio

Subsequently, we demonstrate why the statistical hypothesis testing is meaningful to construct ICA model. The conventional test of independence of random variables is described. We explain why the nonparametric approach is required to fulfill the independence test for ICA problem. A new ICA objective function based on NLR is presented.

#### A. Test of Independence

In multivariate statistical analysis \([1], [15]\), the test of independence of normally distributed \( d \)-dimensional vector \( y = [y^{(1)}, \ldots, y^{(d)}]^T \) has been established. The basic idea is to verify the null hypothesis \( H_0 \) that the components \( y^{(1)}, y^{(2)}, \ldots, y^{(d)} \) are mutually independent against the alternative hypothesis \( H_1 \) that components are dependent. Namely

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

Assuming that \( y \) is Gaussian distributed with unknown mean vector \( \mu = [\mu^{(1)}, \ldots, \mu^{(d)}]^T \) and covariance matrix \( \Sigma \), the null hypothesis is equivalent to see the covariance between two components \( y^{(i)} \) and \( y^{(j)} \)

\[
H_0 : \sigma_{ij}^2 = E\left\{ (y^{(i)}-\mu^{(i)})(y^{(j)}-\mu^{(j)}) \right\} = 0, \quad \text{for all } i \neq j. \tag{8}
\]

When any two distinct components \( \{ (y^{(i)}, y^{(j)}), i \neq j \} \) are uncorrelated, \( y^{(1)}, y^{(d)} \) are mutually independent. This property holds only for Gaussian distributions. Under \( H_0 \), the covariance matrix \( \Sigma \) becomes \( \Sigma_D = \text{diag} \{ \sigma_i^2 \} \) with diagonal elements \( \{ \sigma_i^2, i = 1, \ldots, d \} \). Given the training samples \( Y = \{ y_1, y_2, \ldots, y_T \} \), the optimal solution to this hypothesis testing is to determine the ratio of likelihoods of null hypothesis \( p(Y|H_0) \) to alternative hypothesis \( p(Y|H_1) \) as

\[
\lambda_{\text{LR}} = \frac{p(Y|H_0)/p(Y|H_1)}{\max_{\mu, \Sigma_D} p(Y|\mu, \Sigma_D)} = \frac{\max_{\mu, \Sigma_D} p(Y|\mu, \Sigma_D)}{\max_{\mu, \Sigma} p(Y|\mu, \Sigma)}. \tag{9}
\]

The likelihood ratio is determined with the maximal likelihoods over the parameter spaces \( \{ \mu, \Sigma_D \} \) and \( \{ \mu, \Sigma \} \) for hypotheses \( H_0 \) and \( H_1 \), respectively. This random variable \( \lambda_{\text{LR}} \) serves as the test statistics. With a level of significance \( \alpha \), we can seek the decision boundary \( \lambda_\alpha \) for distribution of \( \lambda_{\text{LR}} \). The null hypothesis \( H_0 \) is verified when \( \lambda_{\text{LR}} \) located in the acceptance region \( \lambda_{\text{LR}} \geq \lambda_\alpha \). Basically, the ratio \( \lambda_{\text{LR}} \) measures the confidence for null hypothesis, or equivalently the independence for \( y^{(1)}, \ldots, y^{(d)} \). The larger the likelihood ratio is, the higher confidence the components are mutually independent.

In this study, we are interested in resolving ICA via test of independence for the transformed data \( y = Wx \). We highlight on developing the likelihood ratio (LR) objective function for ICA. Our goal is to estimate the demixing matrix \( W \) according to the evidence originated from the likelihood ratio test \([1], [15]\).
The demixing matrix $W_{LR}$ is estimated with the largest likelihood ratio, or the highest confidence toward the independence of $y^{(1)}, \ldots, y^{(d)}$. This matrix is feasible to optimally separate the observed signal $\mathbf{x}$ into $\mathbf{y}$ which is nearest to original independent signal $\mathbf{s}$ in likelihood ratio manner. The objective function is defined by $\Gamma_{LR}(X, W) = \lambda_{LR}$. However, the key to estimating $W$ for ICA is non-Gaussianity. The assumption of Gaussian distribution is forbidden for ICA problem. We cannot simply use the parametric likelihood ratio in (9) to find demixing matrix because $\lambda_{LR}$ is valid only for Gaussian distributions.

### B. Nonparametric Density Estimator

In general, incorrect assumptions on distribution of unknown signals can result in poor estimation performance. To avoid assuming Gaussian distribution in ICA, the generalized Gaussian distribution [5], [20] and the generalized Lambda distribution [13] were adopted to determine the ICA objective function. As suggested in [6], Boscolo et al. presented the nonparametric density estimation for the source signals. They used the mutual information as the objective function where a kernel density estimation technique was applied. Similarly, the ICA based on quantizing density estimators was presented [27]. In [38], a mixture of Gaussian models was referred as a kernel smoothing density estimate. The maximum likelihood objective function was adopted to fulfill ICA.

Differently, we are presenting a new NLR objective function for ICA. There is no assumption of parametric distributions in NLR criterion. To fulfill ICA, we investigate the distribution of the transformed signals $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}$ from the observed samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ under $\mathbf{y}_t = W \mathbf{x}_t$. The nonparametric density of component $y^{(i)}$ is provided by

$$p(y^{(i)}) = \frac{1}{T h} \sum_{t=1}^{T} \varphi\left(\frac{y^{(i)} - y^{(i)}_t}{h}\right), \quad i = 1, \ldots, d \tag{10}$$

where $h$ is the kernel bandwidth and $\varphi$ is the Gaussian kernel

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}. \tag{11}$$

The kernel centroid $y^{(i)}_t$ is the $i$th component of sample $\mathbf{y}_t$

$$y^{(i)}_t = w_i \mathbf{x}_t = \sum_{j=1}^{d} w_{ij} x^{(j)}_t. \tag{12}$$

Matrix $W$ is also expressed by $W = [w_{ij}]_{d \times d} = [\mathbf{w}_1^T \cdot \mathbf{w}_d^T]^T$.

### C. Derivation of NLR Objective Function

Interestingly, we employ the nonparametric density function in maximum likelihood ratio criterion and develop the NLR objective function for ICA. Under the null hypothesis, or equivalently assuming independence between $y^{(1)}, y^{(2)}, \ldots, y^{(d)}$, the joint distribution of $\mathbf{y}$ can be factored into the product of distributions of individual components

$p(\mathbf{y}|H_0) = \tilde{p}(\mathbf{y}) = \prod_{i=1}^{d} p(y^{(i)}).$  

The nonparametric likelihood ratio function accumulated from the transformed samples $Y = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}$ is established by

$$\lambda_{NLR} = \frac{p(Y|H_0)}{p(Y|H_1)} = \frac{T}{\prod_{i=1}^{d} p(y^{(i)})}{\prod_{i=1}^{d} p(y^{(i)})} = \frac{1}{T h^d} \sum_{k=1}^{T} \sum_{t=1}^{T} \varphi\left(\frac{y^{(i)}_t - y^{(i)}_k}{h}\right). \tag{13}$$

Here, $\varphi$ is the Gaussian kernel for vectors $\mathbf{v} \in \mathbb{R}^d$ and given by

$$\psi(\mathbf{v}) = \frac{1}{\sqrt{(2\pi)^d}} e^{-\frac{1}{2} \mathbf{v}^T \mathbf{v}}. \tag{14}$$

Notably, we do not specify any parametric distributions for component $y^{(i)}$ and vector $\mathbf{y}$ in calculating $\lambda_{NLR}$. The NLR objective function is then defined by the logarithm of nonparametric likelihood ratio $\lambda_{NLR}$ which is a difference of log likelihoods between null hypothesis $L_0(W)$ and alternative hypothesis $L_1(W)$

$$\Gamma_{NLR}(X, W) = \log \lambda_{NLR} = L_0(W) - L_1(W). \tag{15}$$

We maximize $\Gamma_{NLR}(X, W)$ to find $W_{NLR}$. Again, when applying gradient descent algorithm, two gradient terms $\nabla_W L_0(W)$ and $\nabla_W L_1(W)$ should be determined. Specifically, we derive the components of $\nabla_W L_0(W)$ as

$$\frac{\partial L_0(W)}{\partial w_{ij}} = \frac{T}{h^d} \sum_{k=1}^{T} \sum_{t=1}^{T} \frac{\partial \psi}{\partial w_{ij}} \left(\frac{w_i (x_t - x_k)}{h}\right) = \frac{T}{h^d} \sum_{k=1}^{T} \sum_{t=1}^{T} \psi\left(\frac{w_i (x_t - x_k)}{h}\right) \frac{w_i}{h}. \tag{16}$$

Also, the gradient due to alternative hypothesis has the form

$$\nabla_W L_1(W) = \frac{T}{h^d} \sum_{k=1}^{T} \sum_{t=1}^{T} \psi\left(\frac{w_i (x_t - x_k)}{h}\right) \frac{w_i}{h}. \tag{17}$$

The adjustment for iterative learning of $W_{NLR}$ is obtained by

$$\Delta W = -\eta (\nabla_W L_0(W) - \nabla_W L_1(W)). \tag{18}$$

The derived $W_{NLR}$ is able to optimally demix $\mathbf{x}$ into $\mathbf{y}$ in maximum confidence manner.
Parameter Initialization

Initialize $W$ & Set parameters $\eta$, $h$ and $\lambda_a$

Centering

$$x_i \leftarrow x_i - E[x]$$

Whitening

1. $E[xx^T] = \Phi D \Phi^T$
2. $x_i \leftarrow \Phi D^{-1/2} \Phi^T x_i$

Repeat

1. Compute $\nabla_w L_0(W)$ in (16) and $\nabla_w L_1(W)$ in (17)
2. $W \leftarrow W - \eta (\nabla_w L_0(W) - \nabla_w L_1(W))$
3. $w_i \leftarrow w_i / ||w_i||$, $1 \leq i \leq d$
4. Compute $\lambda_{NLR} = \prod_{i=1}^{T} \prod_{i=1}^{d} p(w_i|x_i) / \prod_{i=1}^{T} p(W|x_i)$ in (13)

Until $\lambda_{NLR} \geq \lambda_a$ or sufficient iterations are run (stopping criterion)

Output $W$

Fig. 1. ICA procedure using NLR objective function.

In this paper, our contributions are focused on developing a new ICA model based on the nonparametric hypothesis testing approach. We also present how ICA can be applied to unsupervised learning of HMMs for speech recognition. There were two works related to proposed ICA [6], [28]. In [28], the test of independence using characteristic functions was presented for measuring the independence of sources for compensating crosstalk interference. In [6], a nonparametric ICA was developed to build the MMI objective function while the simulation data was generated for system evaluation. Different from [6], [28], we completely derive a demixing matrix by maximizing a new NLR objective function, which is interpreted from the hypothesis testing principle. This NLR objective function is unlike ML, ME, and MMI objective functions originated from information maximization viewpoint. To see the relation between mutual information and log-likelihood ratio, it is interesting that the negative of mutual information in (4) can be expressed as $E[\log(p(y)/p(y))]$ which is an expectation of log likelihood ratio between null and alternative hypotheses. In Fig. 1, the ICA procedure based on NLR objective function is shown. We follow the data preprocessing instructions as suggested in [17]. The centering process $x_i \leftarrow x_i - E[x]$ and whitening transformation $x_i \leftarrow \Phi D^{-1/2} \Phi^T x_i$ are performed before applying NLR-ICA algorithm. In whitening process, the eigenvalue matrix $D$ and eigenvector matrix $\Phi$ of covariance matrix $E[xx^T]$ are computed. The learning process is terminated when $\lambda_{NLR} \geq \lambda_a$ or sufficient iterations are run. Typically, $\lambda_{NLR}$ has an upper bound 1 happening in case of independence $p(y) = \tilde{p}(y) = \prod_{i=1}^{d} y_i^{(i)}$. The normalization step $w_i \leftarrow w_i / ||w_i||$ is done to assure the preservation of data variance due to ICA transformation.

IV. APPLICATIONS AND EXPERIMENTS

In this study, we develop new ICA algorithm for speech applications. The proposed NLR-ICA was merged with $k$-means clustering algorithm for estimation of multiple HMMs. The variability of speech production for a subword unit could be compensated for speech recognition. Also, we carried out NLR-ICA algorithm for blind separation of speech and audio signals as mentioned below.

A. Application on Blind Separation of Speech and Audio Signals

When examining NLR-ICA for the BSS problem, we used the benchmark evaluation data sampled from ICA’99 BSS Test Sets [35]. The speech signal from a male speaker1 and the music signal from the Beethoven Symphony no. 52 are shown in Fig. 2. A 2 × 2 mixing matrix $A$ was specified to mix the source signals as given in Fig. 3. No delayed and convolved sources were considered. For comparison, we also implemented ICA algorithm based on MMI objective function as derived in (7). During the implementation of NLR-ICA, we searched the best kernel bandwidth $h$ from several values between 0.05 and 0.25. The nonparametric likelihood ratio $\lambda_{NLR}$ served as the searching criterion. Using parameter $h$, a 2 × 2 demixing matrix $W_{NLR}$ was calculated for blind separation. The performance of NLR-ICA was not sensitive to the choice of bandwidth as similar to [6]. Also, we used the identity matrix as the initial value in all gradient descent algorithms. Empirically, the learning rate $\eta = 0.01$ and 30 iterations were specified in our experiments.

1http://sound.media.mit.edu/ica-bench/sources/mike.wav
2http://sound.media.mit.edu/ica-bench/sources/beet.wav
As illustrated in Figs. 4 and 5, the demixed signals using MMI-ICA and NLR-ICA algorithms are obtained, respectively. It is obvious to see significant improvement by using NLR-ICA compared to MMI-ICA. Readers may download the waveforms for listening at http://chien.csie.ncku.edu.tw/nlr-ica. To measure the performance of separating speech and music signals, we also calculate the signal-to-interference ratio (SIR) in decibels for different cases. Given original source signals \( \{s_1, \ldots, s_T\} \) and demixed signals \( \{y_1, \ldots, y_T\} \), the SIR is computed by

\[
\text{SIR (dB)} = 10 \log_{10} \left( \frac{\sum_{t=1}^{T} ||s_t||^2}{\sum_{t=1}^{T} ||y_t - s_t||^2} \right)
\]  

where \( s_t, y_t \in \mathbb{R}^2 \) in this study. As listed in Table I, we compare the SIRs for the cases of mixed signals (without ICA) and separated signals with MMI-ICA and NLR-ICA. The SIR of mixed signals is measured as low as \(-1.76 \) dB. Using MMI-ICA, SIR is increased to \(5.80 \) dB. Further, using proposed NLR-ICA algorithm, the SIR is substantially improved to \(17.93 \) dB.

### B. Application on HMM Clustering for Speech Recognition

More importantly, we are interested in applying NLR-ICA for unsupervised learning of multiple HMM parameters. In [23], [31], the speaker clustering was developed to select the HMM clusters closest to test speaker for speaker adaptation. Here, the HMM clusters are trained for speaker-independent speech recognition without adaptation. There are two motivations behind multiple HMM approach [34]. First, putting all variability from inhomogeneous pronunciation sources leads to unnecessarily complex models. Second, the inhomogeneity in data sources may be known \textit{a priori} for sophisticated acoustic modeling. Accordingly, the pronunciation variations can be compensated through clustering of speech HMMs [37]. It is popular to use \(k\)-means clustering algorithm to generate multiple HMMs from aligned speech data belonging to a subword unit. The HMM clusters contain the variations of gender, accent, emotion, etc. Having multiple HMMs \( \theta = \{\theta_1, \ldots, \theta_M\} \) covering different variations, we assign the nearest-neighbor HMM cluster \( \theta_{m} \) to fit the environmental condition of observation sequence \( X \)

\[
\hat{m} = \arg \max_m p(X | \theta_m).
\]  

The better the speech data match the HMM parameters \( \theta_{m} \), the higher the recognition accuracy can be achieved.
In the real world, the observations $X$ of a subword unit are noisy and mixed with unknown variations caused by HMM alignment errors, contextual and speaker variability, etc. We need to recover their independent sources $S$ prior to $k$-means clustering so that the desirable clusters of HMMs can be estimated to characterize unknown variations. ICA provides a powerful mechanism to analyze and separate the underlying variation sources. We try to project the sample mean vectors corresponding to the same subword/HMM unit are collected. Then, we calculate the mean vectors $\bar{x}_n$ for the states $\{s_n, 1 \leq n \leq N\}$ belonging to a subword unit. This operation is done segment by segment. Accordingly, we generate the segment-based supervectors consisting of sample mean vectors of HMM states $X = [x^T_{s_1}, x^T_{s_2}, \ldots, x^T_{s_N}]$. Fig. 7 illustrates the idea of generating segment-based supervectors $X = [x_1, x_2, \ldots, x_T]$ from $T$ segments appearing the target subword. Given the supervector matrix $X$, we perform NLR-ICA to find $W_{\text{NLR}}$ and project $X$ to independent component subspace by $Y = W_{\text{NLR}}X$. Intuitively, $Y$ is better than $X$ for $k$-means clustering. After clustering, we identify cluster labels for all subword segments. The segments corresponding to the same cluster are gathered to calculate continuous-density HMM parameters $\theta_m = \{\omega_{sk}, \mu_{sk}, \Sigma_{sk}\}$ with mixture weights $\omega_{sk}$, mean vectors $\mu_{sk}$, and covariance matrices $\Sigma_{sk}$ for states $s$ and mixture components $k$.

C. Speech Database and Experimental Setup

To evaluate the HMM clustering using NLR-ICA, we sampled 8080 sentences uttered by 50 males and 50 females from the benchmark Mandarin speech corpus TCC300 [10]. This reading-style speech database was recorded in ordinary office environments via close-talking microphones. We used 7080 utterances from 40 males and 40 females for HMM training. The remaining 1000 utterances from 10 males and 10 females served as test data. We conducted the experiments on continuous Mandarin speech recognition. Total number of syllables in test data was 15 187. We report the syllable error rates (%) for evaluation. Mandarin is a syllabic and tonal language. Without tone information, there are 408 highly confused Mandarin syllables. Each syllable is composed of an initial (consonant) subsyllable and a final (vowel) subsyllable. The subsyllable serves as the subword/HMM unit. In the experiments, there were 99 right-context-dependent HMMs for initials and 37 context-independent HMMs for finals. The initials and finals were modeled by three and five states, respectively. Without HMM clustering, 485 HMM states (7823 Gaussians) including three nonspeech states were estimated in baseline subsyllable modeling. Each HMM state had 16 Gaussian mixtures on average. Each speech frame was characterized by 12 Mel-frequency cepstral coefficients (MFCCs), 12 delta MFCCs, one log energy, and one delta log energy. The details of HMM modeling for Mandarin speech were mentioned in [8] and [9]. During HMM training, the HMM clusters were estimated for
clear that the transformed data are distinguishable compared to original data. The $k$-means clustering algorithm is applied to determine the HMM clusters for the transformed data.

### D. Evaluation of Speech Recognition Performance

Finally, we evaluated the model size, computational load and recognition accuracy for different methods in speech recognition experiments. The effects of HMM clustering and ICA unsupervised learning were investigated. The baseline system was the case without HMM clustering. In HMM clustering, we compared the performance of $k$-means clustering using original data $X$ and ICA transformed data $Y$. In ICA transformation, we carried out ICA procedure of Fig. 1 based on the MMI objective function in (7) and the NLR objective function in (18). The demixing matrix derived by MMI objective function was used to transform $X$ to $Y$ with minimal mutual information between components $y^{(1)}, \ldots, y^{(d)}$. The NLR demixing matrix was able to maximize the confidence toward independence. The nonparametric density function was used to realize a more realistic ICA approach compared to forcing a parametric density function. For comparative study, we also implemented the nonparametric ICA approach using mutual information criterion proposed by Boscolo et al. [6]. The resulting NMI-ICA was based on the nonparametric mutual information (NMI) objective function. Further, we carried out a simple way to deal with pronunciation variations by increasing the number of Gaussian mixtures in baseline system. To conduct fair comparison, we empirically control the splitting of Gaussian densities and tune the baseline model size comparable to that using HMM clustering. As listed in Table II, we compare the number of HMM Gaussians, training times and syllable error rates (SERs) for the following cases: baseline without HMM clustering, HMM clustering without ICA, HMM clustering with MMI-ICA, NMI-ICA, and proposed NLR-ICA. The baseline results with/without increasing Gaussian mixtures are listed in the first two rows. We find that the number of HMM Gaussians is increased more than double due to HMM clustering. The Gaussian numbers are comparable for different HMM clustering methods. As seen in the table, the baseline SERs are reduced from 38.9% to 37.4% because we use more parameters to cover different kinds of variations from unknown test data. However, both results are worse than 36.5% using standard HMM clustering. The contribution of HMM clustering is obvious. Notably, when ICA method is applied to unsupervised learning of multiple HMMs, we can see that SERs are significantly reduced to 33.6% and 32.5% by adopting MMI and NMI objection functions, respectively. This shows the effectiveness of ICA preprocessing step for $k$-means clustering algorithm. Also, Boscolo’s nonparametric ICA using mutual information criterion did outperform parametric mutual information approach MMI-ICA [11], [39]. This reveals that the nonparametric approach is better than parametric approach to derive ICA model. Furthermore, when the nonparametric likelihood ratio derived from hypothesis test principle is applied to fulfill ICA, the best recognition performance with SER 31.4% is achieved. The likelihood ratio criterion is better than mutual information criterion for ICA data clustering. This promising result demonstrates the effectiveness of NLR-ICA.

![Scattering diagrams](image)

Fig. 8. Scattering diagrams of the first two dimensions of supervectors of Mandarin subsyllable “le” (讃) in (a) original space and (b) NLR-ICA space. “s” and “o” represent male and female samples, respectively.

Each subsyllable. The number of HMM cluster $M$ was set to be two or four by considering the number of segments allocated to target subsyllable. A threshold was set to be 80 to determine $M$. Also, the dimension of segment-based supervector $d$ was the product of feature dimension and state number corresponding to a HMM initial ($d = 26 \times 3$) or HMM final ($d = 26 \times 5$). We estimated $d \times d$ demixing matrix $W_{\text{NLR}}$ for ICA transformation of $x_t \in \mathbb{R}^d$. Clustering of supervectors was affected by the selection of speech features and the goodness of time alignment. To conduct the computational analysis, we implemented different algorithms on a personal computer with RAM 512 MB and CPU Pentium IV 1.6 GHz. The training times in hours were measured for comparison. In Fig. 8, we display the scattering diagrams for the first two dimensions of supervectors of Mandarin subsyllable “le” (讃) from male and female samples marked by “s” and “o,” respectively. The cases of original data $X$ and NLR-ICA transformed data $Y$ are illustrated.
algorithm for modeling pronunciation variations. When evaluating the computational cost, we find that the training time of performing HMM clustering without ICA is comparable to that of directly increasing Gaussian mixtures. Nevertheless, the HMM clustering using NLR-ICA algorithm spends additional 34% computation compared to that without ICA algorithm. How to alleviate the computational load is a concern. Also, in current implementation, HMM clustering is applied for estimation of multiple HMMs corresponding to a HMM unit. No state tying is performed. To reduce the HMM model size and simultaneously compensate the pronunciation variations, it is a good challenge to merge the process of state tying and balance the salience and sharing of the estimated HMM parameters.

V. CONCLUSION

We have surveyed a series of information-theoretic objective functions including ML, ME, and MMI for estimating ICA demixing matrix. These objective functions originated from different criteria were shown to be the same under some specifications. Attractively, we presented maximized to estimate the demixing matrix and detect the independent sources. In the experiments on blind separation of speech and music signals, we found that NLR-ICA approach obtained better separation performance compared to conventional MMI-ICA. More importantly, we discovered the applications of NLR-ICA for robust speech recognition. The procedure of applying ICA in unsupervised learning of HMMs was demonstrated. Because NLR-ICA method transformed training samples to a well-clustered subspace, we were able to estimate multiple and homogenous HMM parameters for speech recognition. The experiments on Mandarin speech recognition showed that the HMM clustering did improve the performance. The models became complex and the size was enlarged accordingly. Due to the contribution of ICA processing, we estimated multiple HMM parameters from compact data clusters. The syllable error rates were significantly reduced. We also found that the nonparametric ICA combined with likelihood ratio criterion was better than parametric and nonparametric approaches combined with mutual information criterion. In the future, we will investigate more applications of ICA on other issues of speech recognition. The issues of controlling model size and reducing computational load will be studied.

ACKNOWLEDGMENT

The authors would like to thank Dr. L. Deng and anonymous reviewers providing valuable comments, which considerably improved the quality of this paper.

REFERENCES


Jen-Zueng Chien received the Ph.D. degree in electrical engineering from the National Tsing Hua University (NTHU), Hsinchu, Taiwan, R.O.C., in 1997.

He was an Instructor in the Department of Electrical Engineering, NTHU, in 1995. In 1997, he joined the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan. He was a Visiting Researcher at the Speech Technology Laboratory, Panasonic Technologies Inc., Santa Barbara, CA, in the summer of 1998, and the Institute of Information Science, Academia Sinica, Taiwan, in the summers of 2003 and 2005. His research interests include general pattern recognition, speech recognition, speaker adaptation, language modeling, information retrieval, face recognition, biometrics, multimedia signal processing, and multimodal human–computer interaction. He serves on the board of the International Journal of Speech Technology and is a Guest Editor of the special issue on Chinese spoken language technology.

He received the Outstanding Young Investigator Award (Ta-You Wu Memorial Award) from the National Science Council, Taiwan, in 2003 and the Research Award for Junior Research Investigators from Academia Sinica in 2004. Dr. Chien is member of the IEEE Signal Processing Society, the International Speech Communication Association, the Chinese Image Processing and Pattern Recognition, and the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). He serves on the board of ACLCLP and chairs the Special Interest Group on Spoken Language Processing.

Bo-Cheng Chen received the M.S. degree in computer science and information engineering from National Cheng Kung University, Tainan, Taiwan, R.O.C., in 2004. His Master’s thesis involved research on independent component analysis and speech recognition.