Journal of Machine Learning Research 4 (2003) 1365-1392 Submitted 10/02; Published 12/03

A Maximum Likelihood Approach to Single-channel Source Separation

Gil-Jin Jang JANGBAL@BAWI.ORG
Spoken Language Laboratory
Division of Computer Science, KAIST
Daejon 305-701, South Korea
Phone: +82-42-869-5556, Fax: +82-42-869-3510

Te-Won Lee TEWON@UCSD.EDU
Institute for Neural Computation
University of California, San Diego
La Jolla, CA 92093, USA
Phone: +1-858-534-9662, Fax: +1-858-534-2014

Editors: Te-Won Lee, Jean-François Cardoso, Erkki Oja and Shun-ichi Amari

Abstract

This paper presents a new technique for achieving blind signal separation when given only a single channel recording. The main concept is based on exploiting *a priori* sets of time-domain basis functions learned by independent component analysis (ICA) to the separation of mixed source signals observed in a single channel. The inherent time structure of sound sources is reflected in the ICA basis functions, which encode the sources in a statistically efficient manner. We derive a learning algorithm using a maximum likelihood approach given the observed single channel data and sets of basis functions. For each time point we infer the source parameters and their contribution factors. This inference is possible due to prior knowledge of the basis functions and the associated coefficient densities. A flexible model for density estimation allows accurate modeling of the observation and our experimental results exhibit a high level of separation performance for simulated mixtures as well as real environment recordings employing mixtures of two different sources.

Keywords: Computational auditory scene analysis (CASA), blind signal separation (BSS), independent component analysis (ICA), generalized Gaussian distribution, sparse coding.

1. Introduction

In natural conversation a speech signal is typically perceived against a background of other sounds carrying different characteristics. The human auditory system processes the acoustic mixture reaching the ears to enable constituent sounds to be heard and recognized as distinct entities, even if these sounds overlap in both spectral and temporal regions with the target speech. This remarkable human speech perception is flexible and robust to various sound sources of different characteristics, therefore spoken communication is possible in many situations even though competing sound sources are present (Bregman, 1990). Researchers in signal processing and many other related fields have strived for the realization of this human ability in machines; however, except in limited certain applications, thus far they have failed to produce the desired outcomes.

In order to formulate the problem, we assume that the observed signal y^t is the summation of P independent source signals

$$y^t = \lambda_1 x_1^t + \lambda_2 x_2^t + \ldots + \lambda_P x_P^t\,, \tag{1}$$

where x_i^t is the t^{th} observation of the i^{th} source, and λ_i is the gain of each source, which is fixed over time. Note that superscripts indicate sample indices of time-varying signals and subscripts identify sources.[1] The gain constants are affected by several factors, such as powers, locations, directions and many other characteristics of the source generators as well as sensitivities of the sensors. It is convenient to assume all the sources to have zero mean and unit variance. The goal is to recover all x_i^t given only a single sensor input y^t. The problem is too ill-conditioned to be mathematically tractable since the number of unknowns is $PT+P$ given only T observations.

Various sophisticated methods have been proposed over the past few years in research areas such as computational auditory scene analysis (CASA; Bregman, 1994, Brown and Cooke, 1994) and independent component analysis (ICA; Comon, 1994, Bell and Sejnowski, 1995, Cardoso and Laheld, 1996). CASA separation techniques are mostly based on splitting mixtures observed as a single stream into different auditory streams by building an active scene analysis system for acoustic events that occur simultaneously in the same spectro-temporal regions. The acoustic events are distinguished according to rules inspired intuitively or empirically from the known characteristics of the sources. Example proposals of CASA are auditory sound segregation models based on harmonic structures of the sounds (Okuno et al., 1999, Wang and Brown, 1999), automatic tone modeling (Kashino and Tanaka, 1993), and psycho-acoustic grouping rules (Ellis, 1994). Recently Roweis (2001) presented a refiltering technique that estimates λ_i in Equation 1 as time-varying masking filters that localize sound streams in a spectro-temporal region. In his work, sound sources are supposedly disjoint in the spectrogram and a “mask” divides the mixed streams completely. These approaches are, however, only applicable to certain limited environments due to the intuitive prior knowledge of the sources such as harmonic modulations or temporal coherency of the acoustic objects.

The use of multiple microphones, such as stereo microphones, binaural microphones, or microphone arrays, may improve separation accuracy. ICA is a data driven method that makes good use of multiple microphone inputs and relaxes the strong characteristic frequency structure assumptions. The ICA algorithms estimate the inverse-translation-operator that maps observed mixtures to the original sources. However, ICA algorithms perform best when the number of observed signals is greater than or equal to the number of sources (Comon, 1994). Although some recent overcomplete representations may relax this assumption (Lewicki and Sejnowski, 2000, Bofill and Zibulevsky, 2001), separating sources from a single channel observation remains problematic.

ICA has been shown to be highly effective in other aspects such as encoding image patches (Bell and Sejnowski, 1997), natural sounds (Bell and Sejnowski, 1996, Abdallah and Plumbley, 2001), and speech signals (Lee and Jang, 2001). The notion of effectiveness adopted here is based on the principle of redundancy reduction (Field, 1994), which states that a useful representation is to transform the input in such a manner that reduces the redundancy due to complex statistical dependencies among elements of the input stream. If the coefficients are statistically independent, that is, $p(x_i, x_j) = p(x_i)p(x_j)$, then the coefficients have a minimum of common information and are

[1] This notation may be confused with n^{th} power. However, this compact notation allows the source separation algorithm given in Section 3 to be to presented in a more orderly fashion by expressing the long formula in one line. Note also that superscripts denoting n^{th} power only appear in Section 2.3.

thus least redundant. In constrast, correlation-based transformations such as principal component analysis (PCA) are based on dimensionality reduction. They search for the axis that has minimum correlations, which does not always match the least redundant transformation. Given segments of predefined length out of a time-ordered sequence of a sound source, ICA infers time-domain basis filters and, at the same time, the output coefficients of the basis filters estimate the least redundant representation. A number of notable research findings suggest that the probability density function (pdf) of the input data is approximated either implicitly or explicitly during the ICA adaptation processes (Pearlmutter and Parra, 1996, MacKay, 1996). *"Infomax"*, a well-known implicit approximation technique proposed by Bell and Sejnowski (1995), models the pdf at the output of the ICA filter by a nonlinear squashing function, and adapts the parameters to maximize the likelihood of the given data.

Our work is motivated by the pdf approximation property involved in the basis filters adapted by ICA learning rules. The intuitive rationale behind the approach is to exploit the ICA basis filters to the separation of mixed source signals observed in a single channel. The basis filters of the source signals are learned a priori from a training data set and these basis filters are used to separate the unknown test sound sources. The algorithm recovers the original auditory streams in a number of gradient-ascent adaptation steps maximizing the log likelihood of the separated signals, computed by the basis functions and the pdfs of their coefficients—the output of the ICA basis filters. We make use of not only the ICA basis filters as strong prior information for the source characteristics, but also their associated coefficient pdfs as an object function of the learning algorithm. The theoretical basis of the approach is **"sparse coding"** (Olshausen and Field, 1996), once termed **"sparse decomposition"** (Zibulevsky and Pearlmutter, 2001). Sparsity in this case means that only a small number of coefficients in the representation differ significantly from zero. Empirical observations show that the coefficient histogram is extremely sparse, and the use of generalized Gaussian distributions (Lewicki, 2002) yields a good approximation.

The remainder of this paper is organized as follows. Section 2 introduces two kinds of generative models for the mixture and the sound sources. Section 3 describes the proposed signal separation algorithm. Section 4 presents the experimental results for synthesized mixtures, and compares them with Wiener filtering. Finally Section 5 summarizes our method in comparison to other methods, and Section 6 draws conclusions.

2. Adapting Basis Functions and Model Parameters

The algorithm first involves the learning of the time-domain basis functions of the sound sources that we are interested in separating. This corresponds to the prior information necessary to successfully separate the signals. We assume two different types of generative models in the observed single channel mixture as well as in the original sources. The first one is depicted in Figure 1-**A**. As described in Equation 1, at every $t \in [1, T]$, the observed instance is assumed to be a weighted sum of different sources. In our approach only the case of $P = 2$ is considered. This corresponds to the situation defined in Section 1: two different signals are mixed and observed in a single sensor.

2.1 A Model for Signal Representation

For the individual source signals, we adopt a decomposition-based approach as another generative model. This approach has been formerly employed in analyzing natural sounds (Bell and Sejnowski, 1996, Abdallah and Plumbley, 2001), speech signals (Lee and Jang, 2001), and colored

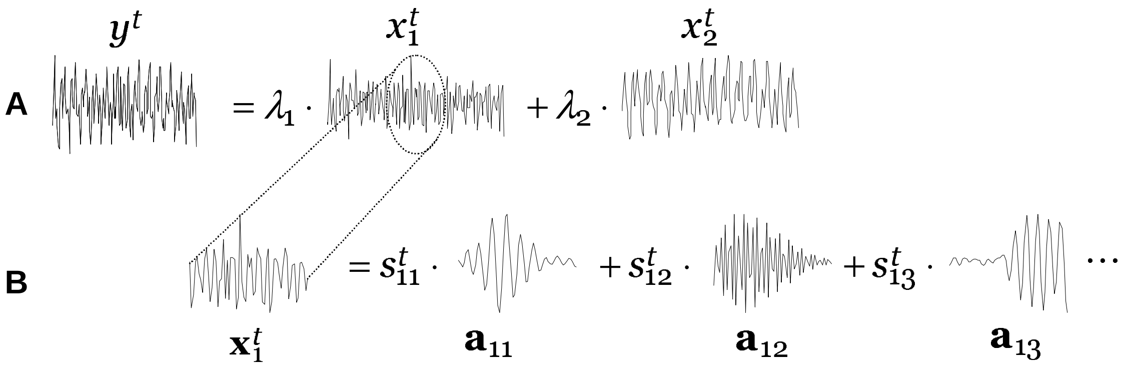


Figure 1: Generative models for the observed mixture and the original source signals. (**A**) The observed single channel input of T samples long is assumed to be generated by a weighted sum of two source signals of the same length: $y^t = \lambda_1 x_1^t + \lambda_2 x_2^t$. (**B**) Decomposition of the individual source signals. The method is to chop x_i^t into blocks of uniform length N starting at t, represented as vectors $\mathbf{x}_i^t = [x_i^t \; x_i^{t+1} \; \ldots \; x_i^{t+N-1}]'$, which is in turn assumed to be generated by weighted linear superpositions of basis functions: $\mathbf{x}_i^t = \sum_k s_{ik}^t \mathbf{a}_{ik}$.

noise (Zibulevsky and Pearlmutter, 2001). A fixed-length segment drawn from a time-varying signal is expressed as a linear superposition of a number of elementary patterns, called basis functions, with scalar multiples (Figure 1-**B**). Continuous samples of length N with $N \ll T$ are chopped out of a source, from t to $t + N - 1$, and the subsequent segment is denoted as an N-dimensional column vector in a boldface letter, $\mathbf{x}_i^t = [x_i^t \; x_i^{t+1} \; \ldots \; x_i^{t+N-1}]'$, attaching the lead-off sample index for the superscript and representing the transpose operator with $'$. The constructed column vector is then expressed as a linear combination of the basis functions such that

$$\mathbf{x}_i^t = \sum_{k=1}^{M} \mathbf{a}_{ik} s_{ik}^t = \mathbf{A}_i \mathbf{s}_i^t \,, \tag{2}$$

where M is the number of basis functions, $\mathbf{a}_{ik}$ is the k^{th} basis function of i^{th} source denoted by an N-dimensional column vector, s_{ik}^t is its coefficient (weight) and $\mathbf{s}_i^t = [s_{i1}^t \; s_{i2}^t \ldots s_{iM}^t]'$. The right-hand side is the matrix-vector notation. The second subscript k followed by the source index i in s_{ik}^t represents the component number of the coefficient vector $\mathbf{s}_i^t$. We assume that $M = N$ and $\mathbf{A}$ has full rank so that the transforms between $\mathbf{x}_i^t$ and $\mathbf{s}_i^t$ are reversible in both directions. The inverse of the basis matrix, $\mathbf{W}_i = \mathbf{A}_i^{-1}$, refers to the ICA basis filters that generate the coefficient vector: $\mathbf{s}_i^t = \mathbf{W}_i \mathbf{x}_i^t$. The purpose of this decomposition is to model the multivariate distribution of $\mathbf{x}_i^t$ in a statistically efficient manner. The ICA learning algorithm searches for a linear transformation $\mathbf{W}_i$ that makes the components as statistically independent as possible. Amari and Cardoso (1997) showed that the solution is achieved when all the individual component pdfs, $p(s_{ik}^t)$, are maximized, provided the linear transformation is invertible:

$$\begin{aligned} \mathbf{W}_i^* &= \arg\max_{\mathbf{W}_i} \prod_t p(\mathbf{x}_i^t | \mathbf{W}_i) \\ &= \arg\max_{\mathbf{W}_i} \prod_t \left\{ \prod_k p(s_{ik}^t) \right\} \cdot |\det(\mathbf{W}_i)| \,, \end{aligned}$$

where $\det(\cdot)$ is the matrix determinant operator, and the term $|\det(\mathbf{W}_i)|$ gives the change in volume produced by the linear transformation (Pham and Garrat, 1997), constraining the solution $\mathbf{W}_i^*$ to be a nonsingular matrix. Independence between the components and over time samples factorizes the joint probabilities of the coefficients into the product of marginal component pdf. Thus the important issue is the degree to which the model distribution is matched to the true underlying distribution $p(s_{ik}^t)$. We do not impose a prior distribution on the source coefficients. Instead, we are interested in inferring the distribution that results in maximally independent coefficients for the sources. Therefore we use a generalized Gaussian prior (Lewicki, 2002) that provides an accurate estimate for symmetric non-Gaussian distributions in modeling the underlying distribution of the source coefficients. The generalized Gaussian prior, also known as exponential power distribution, whose simplest form is $p(s) \propto \exp(-|s|^q)$, can describe Gaussian, platykurtic, and leptokurtic distributions by varying the exponent q. The optimal value of q for given data can be determined from the *maximum a posteriori* value and provides a good fit for the symmetric distributions. In the following sections we present an ICA learning algorithm using a generalized Gaussian function as a flexible prior that estimates the distributions of the sources.

2.2 Learning Basis Functions

In the generative sound model the key parameters are the basis filters. The ICA transformation is performed by the basis filters, the rows of $\mathbf{W}$. They change the coordinates of the original data so that the output coefficients are statistically independent. Initially, we do not know the structure of the basis filters, and therefore we adapt the filters using a generalized formulation of the ICA cost function. First we briefly describe the ICA learning rule.

The goal of ICA is to adapt the filters by optimizing $\mathbf{s}$ so that the individual components s_k are statistically independent, and this adaptation process minimizes the mutual information between s_k. A learning algorithm can be derived using the information maximization principle (Bell and Sejnowski, 1995) or the maximum likelihood estimation (MLE) method (Pearlmutter and Parra, 1996), which can be shown to be equivalent to estimating the density functions (Cardoso, 1997). In our approach, we use the infomax learning rule with natural gradient extension and update the basis functions by the following learning rule (Lee et al., 2000b):

$$\Delta\mathbf{W} \propto \left[\mathbf{I} - \varphi(\mathbf{s})\mathbf{s}'\right]\mathbf{W}, \tag{3}$$

where $\mathbf{I}$ is the identity matrix, $\varphi(\mathbf{s}) = \partial \log p(\mathbf{s})/\partial \mathbf{s}$ and $\mathbf{s}'$ denotes the matrix transpose of $\mathbf{s}$. We assume that $\mathbf{W}$ is square; that is, the number of sources is equal to the number of sensors. The coefficient vector $\mathbf{s}$ can be replaced with any of $\mathbf{s}_i^t$ in Equation 2. To learn the basis filter for the i^{th} source, only $\{\mathbf{s}_i^t | t \in [1,T]\}$ are used. We omit the subscripts and the superscripts in this section for compact notations. $\Delta\mathbf{W}$ is the change of the basis functions that is added to $\mathbf{W}$ and will converge to zero once the adaptation process is complete. Calculating $\varphi(\mathbf{s})$ requires a multivariate density model for $p(\mathbf{s})$, which factorizes to component pdf: $p(\mathbf{s}) = \prod_k^N p(s_k)$. The parametric density estimate $p(s_k)$ plays an essential role in the success of the learning rule. Pham and Garrat (1997) stated that local convergence is assured if $p(s_k)$ is an estimate of the true source density. Note that the global shape of $p(s_k)$ was fixed in previous work (Olshausen and Field, 1996, Hyvärinen, 1999, Lee et al., 2000a).

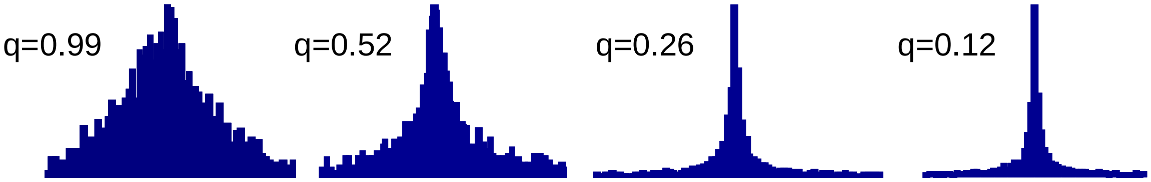


Figure 2: Examples of the actual coefficient distributions and the estimated values of the exponent parameters of the exponential power distributions. The distributions generally have more sharpened summits and longer tails than a Gaussian distribution, and would be classified as super-Gaussian. Generalized Gaussian density functions provide good matches by varying exponents as shown in the equation. From left to right, the exponent decreases, and the distributions become more super-Gaussian.

2.3 Generalized Gaussian Distributions

The success of the ICA learning algorithm for our purpose depends highly on how closely the ICA density model captures the true source coefficient density. The better the density estimation, the better the basis features in turn are responsible for describing the statistical structure. The generalized Gaussian distribution models a family of density functions that is peaked and symmetric at the mean, with a varying degree of normality in the following general form (Lewicki, 2002, Box and Tiao, 1973):

$$p_{\mathbf{g}}(s|\theta) = \frac{\omega(q)}{\sigma} \exp\left[-c(q)\left|\frac{s-\mu}{\sigma}\right|^q\right], \quad \theta = \{\mu, \sigma, q\}$$

where $\mu = E[s]$, $\sigma = \sqrt{E[(s-\mu)^2]}$, $c(q) = \left[\frac{\Gamma[3/q]}{\Gamma[1/q]}\right]^{q/2}$, and $\omega(q) = \frac{\Gamma[3/q]^{1/2}}{(2/q)\Gamma[1/q]^{3/2}}$. The exponent q regulates the deviation from normality. The Gaussian, Laplacian, and strong Laplacian—speech signal—distributions are modeled by putting $q = 2$, $q = 1$, and $q < 1$ respectively. The q parameter is optimized by finding the maximum *a posteriori* value from the data. See work by Box and Tiao (1973) and Lee and Lewicki (2000) for detailed algorithms for q estimation. Each scalar component of the score function in Equation 3 can be computed by using the parametric univariate pdf $p_{\mathbf{g}}(s|\theta)$ for the source coefficient s with suitable generalized Gaussian parameters:

$$\varphi(s) = \frac{\partial \log p_{\mathbf{g}}(s)}{\partial s} = -\frac{cq}{\sigma^q}|s-\mu|^{q-1}\text{sign}(s-\mu)\,. \qquad (4)$$

Gradient ascent adaptation is applied in order to attain the maximal log likelihood. The detailed derivations of the learning algorithm can be found in the original papers (Box and Tiao, 1973, Lee and Lewicki, 2000).

In Figure 2, the coefficient histogram of real data reveals that the distribution has a highly sharpened point at the peak around zero and has heavy and long tails; there is only a small percentage of informative quantities (non-zero coefficients) in the tails and most of the data values are around zero, that is, the data is sparsely distributed. From a coding perspective this implies that we can encode and decode the data with only a small percentage of the coefficients. For modeling the densities of the source coefficients neither Laplacian nor less kurtotic, logistic functions, are adequate for speech bases. The generalized Gaussian parameter set θ_{ik} approximates $p_{\mathbf{g}}(s_{ik}^t)$—the distribution of the k^{th} filter output of the i^{th} source.[2] The basis filters $\mathbf{w}_{ik}$, rows of $\mathbf{W}_i$, and the individual parameter set θ_{ik} for the distribution of the filter output are obtained beforehand by the generalized Gaussian ICA learning algorithm presented by Lee and Lewicki (2000), and used as prior information for the proposed source separation algorithm.

3. Maximum Likelihood Source Inference

Pearlmutter and Parra (1996) showed that the likelihood of the basis filters for a set of training data are maximized by ICA learning algorithm. Suppose we know what kind of sound sources have been mixed and we were given the sets of basis filters from a training set. Could we infer the learning data? The answer is generally "no" when $N < T$ and no other information is given. In our problem of single channel signal separation, half of the solution is already given by the constraint $y^t = \lambda_1 x_1^t + \lambda_2 x_2^t$, where x_i^t constitutes the basis learning data $\mathbf{x}_i^t$ (Figure 1-**B**). Essentially, the goal of the source inferring algorithm of this paper is to complement the remaining half with the statistical information given by a set of basis filters $\mathbf{W}_i$ and coefficient density parameters θ_{ik}. If the parameters are given, we can perform *maximum a posteriori* (MAP) estimation by optimizing the data likelihood computed by the model parameters.

The separation algorithm has two major features: it is *adaptive* and should perform all relevant adaptation *on a single sample basis,* which means that the solution is achieved by altering a set of unknowns gradually from an arbitrary initial values to a certain goal, and the number of unknowns to be estimated equals the number of samples. In Section 3.1 we formulate a stochastic gradient ascent adaptation algorithm for the problem. In Section 3.2 we derive detailed adaptation formulas for the source signals, which is done by the generalized Gaussian expansion of the coefficient pdf. Section 3.3 explains how to update the scaling factors λ_i. Finally Section 3.4 gives a step-by-step description of the proposed separation algorithm in terms of the derived learning rules. The evaluation of the derived separation algorithm and practical issues in actual situations are discussed in Section 4.

3.1 Formulation of Separation Algorithm

If we have probabilistic models for x_1^t and x_2^t by the sets of basis filters $\mathbf{W}_1$ and $\mathbf{W}_2$, and if two source signals are statistically independent, we can formulate the single channel signal separation by the following constrained maximization problem:[3]

$$\begin{aligned}\left\{x_1^{t\,*}, x_2^{t\,*} \middle|\, t = 1, \ldots, T\right\} &= \arg\max_{\{x_1^t, x_2^t\}} p(x_1^1, x_1^2, \ldots, x_1^T | \mathbf{W}_1) \cdot p(x_2^1, x_2^2, \ldots, x_2^T | \mathbf{W}_2)\,, \\ &\ \text{s.t. } y^t = \lambda_1 x_1^t + \lambda_2 x_2^t\end{aligned}$$

where x_i^t is a sampled value of the i^{th} source at time t, and T is the length of each source signal. Separation of two source signals from a mixture can be regarded as a mapping from y^t to $\{x_1^t, x_2^t\}$. Since the number of parameters is $2T$ given only T observations, it is mathematically intractable to

[2] In the remainder of the paper, we will drop the parameter set θ_{ik} of a generalized Gaussian pdf $p_{\mathbf{g}}(s_{ik}^t|\theta_{ik})$. When we refer to a generalized Gaussian pdf $p_{\mathbf{g}}(s_{ik}^t)$, we assume that it is conditioned on a set of parameters $\theta_{ik} = \{\mu_{ik}, \sigma_{ik}, q_{ik}\}$, where the subscripts $\{i, k\}$ imply that every source coefficient distribution has its own set of parameters.

[3] The pdf $p(x_i^1, \ldots, x_i^T | \mathbf{W}_i)$ should be also conditioned on a set of generalized Gaussian parameters $\{\theta_{ik}\}$. We will drop the parameter set in the remainder of the paper and implicitly assume its existence in the pdfs whenever the basis filter $\mathbf{W}_i$ is conditioned or the generalized Gaussian pdf symbol $p_{\mathbf{g}}$ appears.

evaluate true values of the source signals. Instead, the proposed algorithm tries to find the estimates of the sources to maximize the posterior probability given the basis filters $\mathbf{W}_1$ and $\mathbf{W}_2$. In an ordinary ICA, the learning algorithm optimizes data likelihood by altering a set of basis filters. The target of the proposed separation method is identical, but the values to be altered are the data, not the basis filters.

The initial constraint, $y^t = \lambda_1 x_1^t + \lambda_2 x_2^t$, reduces the number of the unknowns to T, according to the following alternative formulation:

$$\begin{aligned}\left\{ {x_1^t}^* \middle| t = 1, \dots, T \right\} &= \arg\max_{\{x_1^t\}} p(x_1^1, x_1^2, \dots, x_1^T | \mathbf{W}_1) \cdot p(x_2^1, x_2^2, \dots, x_2^T | \mathbf{W}_2), \\ &\quad \text{where } x_2^t = (y^t - \lambda_1 x_1^t)/\lambda_2 \,.\end{aligned} \tag{5}$$

Due to a large amount of probabilistic dependence along the time samples of the source signals, evaluating $p(x_i^1, x_i^2, \dots, x_i^T | \mathbf{W}_i)$ is not a simple matter. However, if we assume that the dependence does not exceed N samples, such that $x_i^{t_1}$ and $x_i^{t_2}$ are statistically independent when $|t_1 - t_2| > N$, the probability of the whole signal is approximated by the product of the probability of all possible windows of length N,

$$\begin{aligned} p(x_i^1, x_i^2, \dots, x_i^T | \mathbf{W}_i) &\approx p(x_i^1, \dots, x_i^N | \mathbf{W}_i) p(x_i^2, \dots, x_i^{N+1} | \mathbf{W}_i) \cdots p(x_i^{T_N}, \dots, x_i^T | \mathbf{W}_i) \\ &= \prod_{\tau=1}^{T_N} p(\mathbf{x}_i^\tau | \mathbf{W}_i), \end{aligned} \tag{6}$$

where $T_N = T - N + 1$ and $\mathbf{x}_i^\tau = [x_i^\tau \; x_i^{\tau+1} \; \dots \; x_i^{\tau+N-1}]'$ as defined in Section 2.1.[4]

Now we focus on evaluating the multivariate pdf $p(\mathbf{x}_i^\tau | \mathbf{W}_i)$. When we pass $\mathbf{x}_i^\tau$ through a set of linear basis filters $\mathbf{W}_i$, a set of random variables, $\{s_{ik}^\tau = \mathbf{w}_{ik} \mathbf{x}_i^\tau | k = 1, \dots, N\}$, where k is a filter number, emerge at the output. By virtue of the ICA learning algorithm, the probabilistic dependence between the output random variables is minimized; hence we approximate $p(\mathbf{x}_i^\tau | \mathbf{W}_i)$ to the multiplication of the univariate pdfs of the output variables:

$$p(\mathbf{x}_i^\tau | \mathbf{W}_i) \approx |\det(\mathbf{W}_i)| \cdot \prod_{k=1}^{N} p_{\mathbf{g}}(s_{ik}^\tau), \tag{7}$$

where $p_{\mathbf{g}}(\cdot)$ is the generalized Gaussian pdf introduced in Section 2.3. The term $|\det(\mathbf{W}_i)|$ gives the change in volume produced by the linear transformation (Pham and Garrat, 1997). We define the object function $\mathcal{L}$ of the separation problem as the joint log probability of the two source signals given the basis filters, which is approximated to the sum of the log probabilities of the output variables based on Equations 6 and 7:

$$\begin{aligned} \mathcal{L} &\stackrel{\text{def}}{=} \log p(x_1^1, x_1^2, \dots, x_1^T | \mathbf{W}_1) \cdot p(x_2^1, x_2^2, \dots, x_2^T | \mathbf{W}_2) \\ &\approx \log \prod_{\tau=1}^{T_N} p(\mathbf{x}_1^\tau | \mathbf{W}_1) \cdot p(\mathbf{x}_2^\tau | \mathbf{W}_2) \\ &\approx \log \prod_{\tau=1}^{T_N} \left\{ |\det(\mathbf{W}_1)| \prod_{k=1}^{N} p_{\mathbf{g}}(s_{1k}^\tau) \cdot |\det(\mathbf{W}_2)| \prod_{k=1}^{N} p_{\mathbf{g}}(s_{2k}^\tau) \right\} \end{aligned}$$

[4] We use different timing indices t and τ, for t^{th} sample of a signal and for the column vector of continuous samples of length N starting from τ, respectively.

$$\propto \quad \sum_{\tau=1}^{T_N} \sum_{k=1}^{N} \left[\log p_{\mathbf{g}}(s_{1k}^{\tau}) + \log p_{\mathbf{g}}(s_{2k}^{\tau})\right] . \tag{8}$$

In the last expression, $|\det(\mathbf{W}_1)|$ and $|\det(\mathbf{W}_2)|$ vanish since their values are constant over the change of x_i^t. To find an optimized value of x_1^t at $\forall t \in \{1, \ldots, T\}$, we perform a gradient ascent search based on the adaptation formula derived by differentiating $\mathcal{L}$ with respect to x_1^t:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial x_1^t} &= \sum_{\tau=1}^{T_N} \sum_{k=1}^{N} \left[\frac{\partial \log p_{\mathbf{g}}(s_{1k}^{\tau})}{\partial x_1^t} + \frac{\partial \log p_{\mathbf{g}}(s_{2k}^{\tau})}{\partial x_1^t} \right] \\ &= \sum_{\tau=1}^{T_N} \sum_{k=1}^{N} \left[\frac{\partial \log p_{\mathbf{g}}(s_{1k}^{\tau})}{\partial s_{1k}^{\tau}} \frac{\partial s_{1k}^{\tau}}{\partial x_1^t} + \frac{\partial \log p_{\mathbf{g}}(s_{2k}^{\tau})}{\partial s_{2k}^{\tau}} \frac{\partial s_{2k}^{\tau}}{\partial x_2^t} \frac{\partial x_2^t}{\partial x_1^t} \right] , \end{aligned} \tag{9}$$

where the three different derivative terms inside the summation have the following meanings

- $\frac{\partial \log p_{\mathbf{g}}(s_{ik}^{\tau})}{\partial s_{ik}^{\tau}}$: stochastic gradient ascent for the k^{th} filter output.
- $\frac{\partial s_{ik}^{\tau}}{\partial x_i^t}$: adjustment in change from k^{th} filter output to source i.
- $\frac{\partial x_2^t}{\partial x_1^t}$: adjustment in change from source 2 to source 1.

In the following section we evaluate the above three terms, and present the actual adaptation procedures considering the constraint, $y^t = \lambda_1 x_1^t + \lambda_2 x_2^t$.

3.2 Deriving Adaptation Formulas

The stochastic gradient ascent for the k^{th} filter output of the i^{th} source is

$$\frac{\partial \log p_{\mathbf{g}}(s_{ik}^{\tau})}{\partial s_{ik}^{\tau}} = \varphi(s_{ik}^{\tau}), \quad \tau \in \{1, \ldots, T_N\} , \tag{10}$$

where $\varphi(\cdot)$ is the component score function of the generalized Gaussian pdf defined in Equation 4. The derivative term $\partial s_{ik}^{\tau} / \partial x_i^t$ is the adjustment from source coefficients to the original time domain. Because x_i^t can be appeared at any of the N possible positions of the input vector $\mathbf{x}_i^{\tau}$ whose output is s_{ik}^{τ}, the adjustment is determined by t and τ. Figure 3 provides a conceptual explanation of the adjustment mapping. Each $\mathbf{w}_{ik}$ takes windows of N continuous samples starting from the τ^{th} sample out of the i^{th} source signal, $\mathbf{x}_i^{\tau} = [x_i^{\tau} \ \ldots \ x_i^{\tau+N-1}]'$, and produces the output coefficient s_{ik}^{τ}. Each sample of the source participates in the generation of N different inputs, and henceforth in the generation of N different output coefficients for each filter. The following matrix-vector expression of the basis filtering highlights positions of a sample x_i^t in all the possible input windows:

$$\begin{aligned} \left[\ s_{ik}^{t-N+1} \quad s_i^{t-N+2} \quad \cdots \quad s_{ik}^{t}\ \right] &= \mathbf{w}_{ik} \cdot \left[\ \mathbf{x}_i^{t-N+1} \quad \mathbf{x}_i^{t-N+2} \quad \cdots \quad \mathbf{x}_i^{t}\ \right] \\ &= \begin{bmatrix} w_{ik1} \\ w_{ik2} \\ \vdots \\ w_{ikN} \end{bmatrix}' \cdot \begin{bmatrix} x_i^{t-N+1} & x_i^{t-N+2} & \cdots & \boxed{x_i^t} \\ x_i^{t-N+2} & \cdots & \boxed{x_i^t} & x_i^{t+1} \\ \vdots & \boxed{x_i^t} & \ddots & \vdots \\ \boxed{x_i^t} & x_i^{t+1} & \cdots & x_i^{t+N-1} \end{bmatrix} , \end{aligned}$$

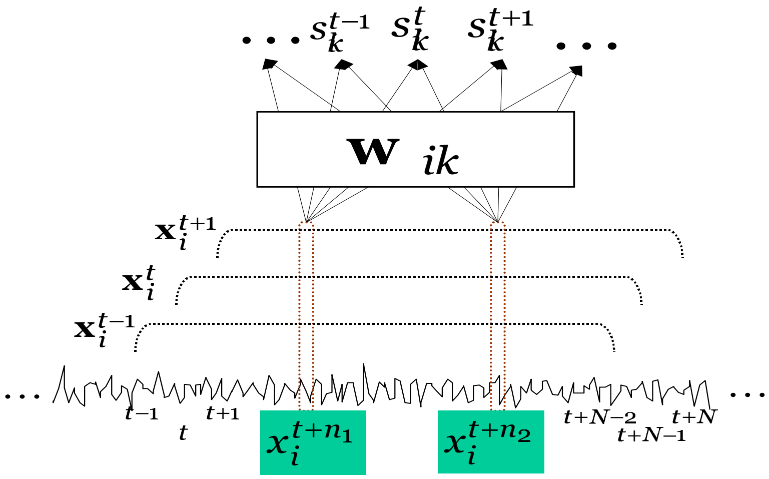


Figure 3: The participation of a sample in the source signal to the generation of each output coefficient. The input $\mathbf{x}_i^t$ is a vector composed of N continuous samples ranging from t to $t+N-1$ in the i^{th} source. The output coefficient s_{ik}^t is obtained by passing $\mathbf{x}_i^t$ through $\mathbf{w}_{ik}$. The middle of the figure shows that there exist N different possible input covers over a sample, which subsequently participate in the generation of N different output coefficients per filter.

where the scalar w_{ikn} is the n^{th} component of $\mathbf{w}_{ik}$. The indices of the windows containing x_i^t range from $t-N+1$ to t. We introduce an offset variable $n \in [1,N]$ so that s_{ik}^{t-n+1} may cover the range $[t-N+1,t]$. Then the partial derivative of the output at time $t-n+1$ with respect to the source signal at time t becomes a simple scalar value as

$$\frac{\partial s_{ik}^{t-n+1}}{\partial x_i^t} = \frac{\partial \left(\sum_{\tau=1}^{N} w_{ik\tau} \cdot x_i^{t-n+\tau}\right)}{\partial x_i^t} = w_{ikn} \,. \tag{11}$$

The summation from $\tau = 1$ to T_N in Equation 9 is reduced to the summation over only N relevant output coefficients, by using the offset variable n and Equations 10 and 11,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial x_1^t} &= \sum_{\tau=1}^{T_N} \sum_{k=1}^{N} \left[\frac{\partial \log p_{\mathbf{g}}(s_{1k}^{\tau})}{\partial s_{1k}^{\tau}} \frac{\partial s_{1k}^{\tau}}{\partial x_1^t} + \frac{\partial \log p_{\mathbf{g}}(s_{2k}^{\tau})}{\partial s_{2k}^{\tau}} \frac{\partial s_{2k}^{\tau}}{\partial x_2^t} \frac{\partial x_2^t}{\partial x_1^t} \right] \\ &= \sum_{n=1}^{N} \sum_{k=1}^{N} \left[\varphi(s_{1k}^{t_n}) w_{1kn} + \varphi(s_{2k}^{t_n}) w_{2kn} \cdot \frac{\partial x_2^t}{\partial x_1^t} \right] , \end{aligned} \tag{12}$$

where $t_n = t-n+1$. The first multiplier inside the summation, $\varphi(s_{ik}^{t_n})$, is interpreted as a stochastic gradient ascent that gives the direction and the amount of the change at the output of the basis filter, $s_{ik}^t = \mathbf{w}_{ik}\mathbf{x}_i^t$. The second term w_{ikn} accounts for the change produced by the filter between the input x_i^t and the output $s_{ik}^{t_n}$. The summation implies that the source signal is decomposed to N independent components.

The last derivative term inside the summation of Equation 12 defines the interactions between the two source signals, determined by the constraint $y^t = \lambda_1 x_1^t + \lambda_2 x_2^t$. At every time t every source signal can be expressed by the counterpart, $x_2^t = (y^t - \lambda_1 x_1^t)/\lambda_2$ and $x_1^t = (y^t - \lambda_2 x_2^t)/\lambda_1$. We represent the relationships between the sources by the following two equivalent differential equations:

$$\frac{\partial x_2^t}{\partial x_1^t} = -\frac{\lambda_1}{\lambda_2} \Leftrightarrow \frac{\partial x_1^t}{\partial x_2^t} = -\frac{\lambda_2}{\lambda_1}.$$

We evaluate the final learning rule for x_1^t as

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial x_1^t} &= \sum_{n=1}^{N}\sum_{k=1}^{N}\left[\varphi(s_{1k}^{t_n})w_{1kn} + \varphi(s_{2k}^{t_n})w_{2kn}\cdot\left(-\frac{\lambda_1}{\lambda_2}\right)\right] \\ &= \sum_{k=1}^{N}\sum_{n=1}^{N}\left[\varphi(s_{1k}^{t_n})w_{1kn} - \frac{\lambda_1}{\lambda_2}\cdot\varphi(s_{2k}^{t_n})w_{2kn}\right]. \end{aligned} \tag{13}$$

The second term inside the final summation can be interpreted as a stochastic gradient ascent for x_2^t scaled by $-\lambda_1/\lambda_2$. The denominator λ_2 normalizes the gradient, and the numerator λ_1 scales it to be added to x_1^t. The minus sign implies that adjusting x_2^t affects x_1^t in the opposite direction. Similar reasoning leads to the rule for the second source:

$$\frac{\partial \mathcal{L}}{\partial x_2^t} = \sum_{k=1}^{N}\sum_{n=1}^{N}\left[-\frac{\lambda_2}{\lambda_1}\cdot\varphi(s_{1k}^{t_n})w_{1kn} + \varphi(s_{2k}^{t_n})w_{2kn}\right]. \tag{14}$$

Updating the sources directly using these learning rules might lead to a violation of the initial constraint. To avoid the violation, the values of the source signals after adaptation must always satisfy

$$\begin{aligned} y^t &= \lambda_1(x_1^t + \Delta x_1^t) + \lambda_2(x_2^t + \Delta x_2^t) \\ &\Leftrightarrow \lambda_1 \Delta x_1^t + \lambda_2 \Delta x_2^t = 0. \end{aligned}$$

In the actual application of the adaptation rules, we scale Equations 13 and 14 appropriately and express the final learning rules as

$$\begin{aligned} \Delta x_1^t &= \eta \sum_{k=1}^{N}\sum_{n=1}^{N}\left[\frac{\lambda_2}{\lambda_1}\cdot\varphi(s_{1k}^{t_n})w_{1kn} - \varphi(s_{2k}^{t_n})w_{2kn}\right], \\ \Delta x_2^t &= \eta \sum_{k=1}^{N}\sum_{n=1}^{N}\left[-\varphi(s_{1k}^{t_n})w_{1kn} + \frac{\lambda_1}{\lambda_2}\cdot\varphi(s_{2k}^{t_n})w_{2kn}\right], \end{aligned} \tag{15}$$

where η is a learning gain. The whole dataflow of the proposed method is summarized in four steps in Figure 4. In step **A**, the source signals are decomposed into N statistically independent codes. The decomposition is done by a set of the given ICA filters, $\mathbf{s}_i^t = \mathbf{W}_i \mathbf{x}_i^t$. In step **B**, the stochastic gradient ascent for each filter output code is computed from the derivative of the log likelihood of the code (Equation 10). In step **C**, the computed gradient is transformed to the source domain according to Equation 11. All the filter output codes are regarded as being independent, so all the computations are performed independently. In step **D**, we add up all the gradients and modify them to satisfy the initial constraint according to Equation 15. The four steps comprise one iteration of the adaptation of each sample. The solution is achieved after repeating this iteration on the source signal x_i^t at every time t to a convergence from certain initial values.

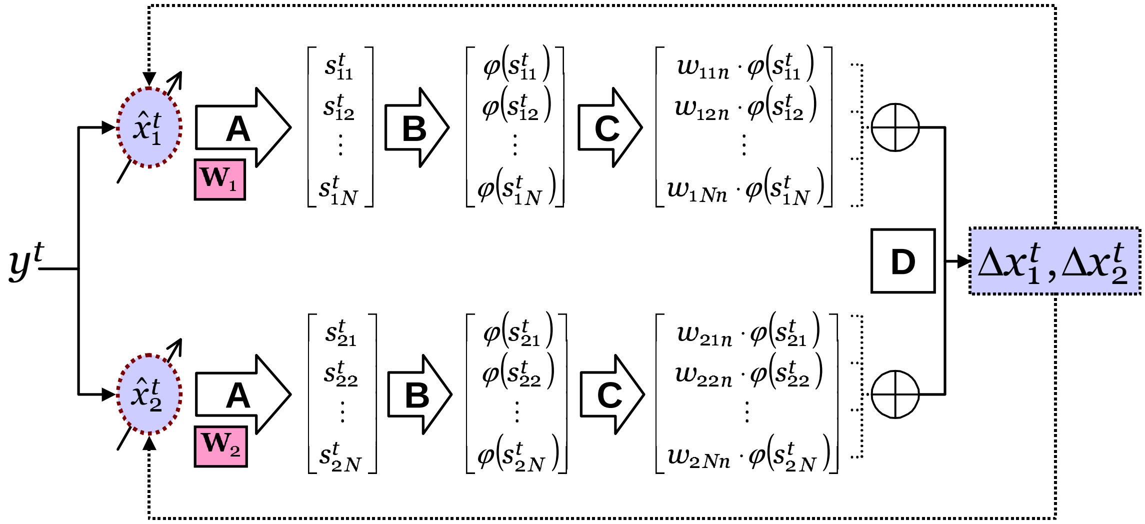


Figure 4: The overall structure and the data flow of the proposed method. In the beginning, we are given single channel data y^t, and we have the estimates of the source signals, $\hat{x}_i^t$, at every adaptation step. (**A**) At each time-point, the current estimates of the source signals are passed through a set of basis filters $\mathbf{W}_i$, generating N sparse codes s_{ik}^t that are statistically independent. (**B**) The stochastic gradient for each code is computed from the derivative of the log likelihood of each individual code. (**C**) The gradient for each code is transformed to the domain of source signal. (**D**) The individual gradients are combined and modified to satisfy the given constraints, and added to the current estimates of the source signals.

3.3 Updating Scaling Factors

Updating the contribution factors λ_i can be accomplished by finding the maximum *a posteriori* values. To simplify the inferring steps, we force the sum of the factors to be constant, such that $\lambda_1 + \lambda_2 = 1$. The value of λ_2 is completely dependent on the value of λ_1, so we need to consider λ_1 only. Given the current estimates of the sources x_i^t, the posterior probability of λ_1 is

$$p(\lambda_1 | x_1^1, \ldots, x_1^T,\ x_2^1, \ldots, x_2^T) \propto p(x_1^1, \ldots, x_1^T) p(x_2^1, \ldots, x_2^T) p_\lambda(\lambda_1),$$

where $p_\lambda(\cdot)$ is the prior density function of λ. Performing a log operation on the above equation yields

$$\log p(x_1^1, \ldots, x_1^T) p(x_2^1, \ldots, x_2^T) p_\lambda(\lambda_1) \ \cong\ \mathcal{L} + \log p_\lambda(\lambda_1)\,,$$

where $\mathcal{L}$ is the object function defined in Equation 8. If we assume λ_1 to be uniformly distributed in the range $[0, 1]$, $p_\lambda(\cdot)$ is considered as a constant, and vanishes in

$$\begin{aligned} \lambda_1^* &= \arg\max_{\lambda_1} \{\mathcal{L} + \log p_\lambda(\lambda_1)\} \\ &= \arg\max_{\lambda_1} \mathcal{L}\,. \end{aligned}$$

We differentiate $\mathcal{L}$ with respect to λ_1, and substitute $\lambda_2 = 1 - \lambda_1$:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \lambda_1} &= \sum_{\tau=1}^{T_N} \sum_{k=1}^{N} \left[\frac{\partial \log p_{\mathbf{g}}(s_{1k}^{\tau})}{\partial \lambda_1} + \frac{\partial \log p_{\mathbf{g}}(s_{2k}^{\tau})}{\partial \lambda_1} \right] \\ &= \sum_{\tau=1}^{T_N} \sum_{k=1}^{N} \left[\varphi(s_{1k}^{\tau}) \frac{\partial s_{1k}^{\tau}}{\partial \lambda_1} + \varphi(s_{2k}^{\tau}) \frac{\partial s_{2k}^{\tau}}{\partial \lambda_2} \cdot \frac{\partial \lambda_2}{\partial \lambda_1} \right] \\ &= \sum_{\tau=1}^{T_N} \sum_{k=1}^{N} \left[\varphi(s_{1k}^{\tau}) \frac{\partial s_{1k}^{\tau}}{\partial \lambda_1} - \varphi(s_{2k}^{\tau}) \frac{\partial s_{2k}^{\tau}}{\partial \lambda_2} \right] .\end{aligned}$$

From the constraint $y^t = \lambda_1 x_1^t + \lambda_2 x_2^t$, we might deduce that the value of $\lambda_i x_i^t$ is unaffected by the change of either λ_i or x_i^t, for all $i \in \{1,2\}$, $t \in [1,T]$. Because s_{ik}^{τ} is the output of x_i^t, $\lambda_i s_{ik}^{\tau}$ is also unaffected by λ_i or s_{ik}^{τ}. So the partial derivative of s_{ik}^{τ} with respect to λ_i becomes

$$\frac{\partial s_{ik}^{\tau}}{\partial \lambda_i} = \lambda_i s_{ik}^{\tau} \cdot \frac{\partial}{\partial \lambda_i} \left(\frac{1}{\lambda_i} \right) = -\frac{s_{ik}^{\tau}}{\lambda_i} .$$

The stochastic gradient ascent for λ_1 is then

$$\frac{\partial \mathcal{L}}{\partial \lambda_1} = \sum_{\tau=1}^{T_N} \sum_{k=1}^{N} \left[-\varphi(s_{1k}^{\tau}) \frac{s_{1k}^{\tau}}{\lambda_1} + \varphi(s_{2k}^{\tau}) \frac{s_{2k}^{\tau}}{\lambda_2} \right] . \tag{16}$$

In order to satisfy the constraint $\lambda_1 \in [0,1]$, we perform the update by

$$\lambda_1^{(new)} = h_{\lambda} \left(\lambda_1^{(old)} + \eta_{\lambda} \cdot \frac{\partial \mathcal{L}}{\partial \lambda_1} \right) , \tag{17}$$

where η_{λ} is a learning gain for λ_1 and the limiting function $h_{\lambda}(\cdot)$ is

$$h_{\lambda}(d) = \begin{cases} \varepsilon & \text{if } d < \varepsilon \\ 1-\varepsilon & \text{if } d > 1-\varepsilon \\ d & \text{otherwise} \end{cases} ,$$

where $\varepsilon \in [0,1]$ is a positive real constant. In our implementation η_{λ} is determined empirically, and and ε is set to less than 10^{-3}.

3.4 Iterative Source Separation Algorithm and Time Complexity Analysis

Using the adaptation formulas derived in the preceding sections, the optimization of Equation 6 can be accomplished by a simple iterative algorithm with the following form:

Algorithm: SINGLE CHANNEL SOURCE SEPARATION

```
Inputs
  Observations: $\{y^t | t = 1, \ldots, T\}$
  Model parameters: $\mathbf{W}_1$, $\mathbf{W}_2$, $\{\theta_{1k}, \theta_{2k} | k = 1, \ldots, N\}$
```

Outputs[5]

Source signal estimates: $\{\hat{x}_1^t, \hat{x}_2^t \,|\, t = 1, \dots, T\}$
Gain constants: $\hat{\lambda}_1$, $\hat{\lambda}_2$

Procedures

1. Take some initial values for the outputs.
   For example, $\hat{x}_1^t \Leftarrow y^t$, $\hat{x}_2^t \Leftarrow y^t$, $\forall t \in [1, T]$, $\hat{\lambda}_1 \Leftarrow 0.5$, $\hat{\lambda}_2 \Leftarrow 0.5$.
2. For all $i \in [1, 2]$, $t \in [1, T - N + 1]$, and $k \in [1, N]$,
   (a) Compute $\hat{s}_{ik}^t = \mathbf{w}_{ik}\hat{\mathbf{x}}_i^t$ where $\hat{\mathbf{x}}_i^t = [\hat{x}_i^t \; \hat{x}_i^{t+1} \; \dots \; \hat{x}_i^{t+N-1}]'$
   (b) Compute $\varphi(\hat{s}_{ik}^t)$ according to Equation 4 using the generalized Gaussian parameters θ_{ik}.
3. Update T samples of the source signal estimates at the same time according to Equation 15, to be precise

$$
\begin{aligned}
\hat{x}_1^t &\Leftarrow \hat{x}_1^t + \eta \sum_{k=1}^{N} \sum_{n=1}^{N} \left[\frac{\hat{\lambda}_2}{\hat{\lambda}_1} \cdot \varphi(\hat{s}_{1k}^{t_n}) w_{1kn} - \varphi(\hat{s}_{2k}^{t_n}) w_{2kn} \right], \\
\hat{x}_2^t &\Leftarrow \hat{x}_2^t + \eta \sum_{k=1}^{N} \sum_{n=1}^{N} \left[-\varphi(\hat{s}_{1k}^{t_n}) w_{1kn} + \frac{\hat{\lambda}_1}{\hat{\lambda}_2} \cdot \varphi(\hat{s}_{2k}^{t_n}) w_{2kn} \right].
\end{aligned}
$$

4. Update scaling factors according to Equations 16 and 17,

$$
\begin{aligned}
\hat{\lambda}_1^{(new)} &\Leftarrow h_\lambda \left(\hat{\lambda}_1^{(old)} + \eta_\lambda \cdot \sum_{\tau=1}^{T_N} \sum_{k=1}^{N} \left[-\varphi(\hat{s}_{1k}^{\tau}) \frac{\hat{s}_{1k}^{\tau}}{\lambda_1} + \varphi(\hat{s}_{2k}^{\tau}) \frac{\hat{s}_{2k}^{\tau}}{\lambda_2} \right] \right) \\
\hat{\lambda}_2^{(new)} &\Leftarrow 1 - \hat{\lambda}_1^{(new)}.
\end{aligned}
$$

5. Repeat steps from 2 to 4 until convergence.

The computational overhead of steps 2a and 3 dominates the time complexity of the algorithm. In step 2a, N multiplications are required to compute $2N(T - N + 1)$ output coefficients. In step 3, $2N^2$ terms are summed up to evaluate the gradient for each sample. The time complexity of the algorithm for one iteration is therefore $O(N^2 T)$ if $N \ll T$.

4. Evaluations

We now present some examples of single channel separation of artificial mixtures using speech signals and music signals. The separation performances with the basis filters learned by ICA are compared to those with other conventional bases—Fourier, fixed wavelet function, and data-driven principal component analysis (PCA) basis filters. To assess the limits of our method, we compared our method to Wiener filtering with real spectrograms. We then present the separation results of noise and speech recorded in a real environment.

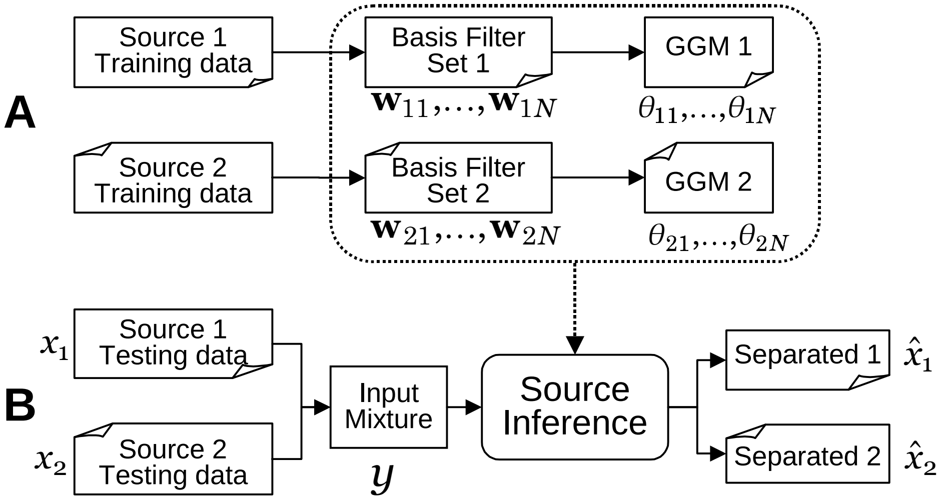


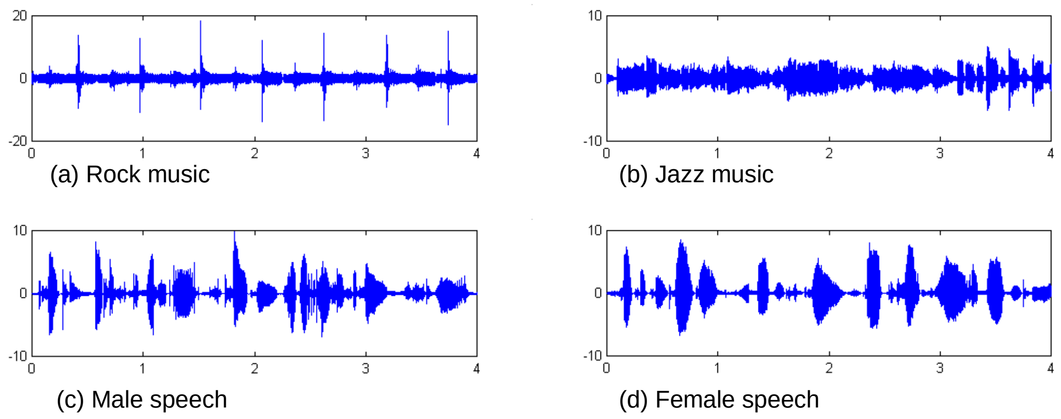
Figure 5: Simulation system setup. (A) Training phase: two sets of training data are used to obtain the basis filters and the generalized Gaussian parameters. (B) Testing phase: two source signals x_1 and x_2 are mixed into a monaural signal y. The proposed signal separation algorithm recovers the original source signals given the sets of the basis filters and generalized Gaussian pdf parameters.

4.1 Simulation Setup

We have tested the performance of the proposed method on single channel mixtures of two different sound types. The simulation system setup is illustrated in Figure 5. The simulation is divided into two phases. In the first phase, we prepare training data, and run the ICA learning algorithm to obtain basis filters $\mathbf{w}_{ik}$, and generalized Gaussian parameters (θ_{ik}) for modeling coefficient (s_{ik}^t) pdfs. The basis filters and pdf parameters are estimated separately for both source 1 and source 2. In the testing phase, two source signals x_1^t and x_2^t, which are not included in the training data sets, are mixed into a single channel mixture y^t, and we apply the proposed separation algorithm and recover the original sources.

We adopted four different sound types for our simulation experiment. They were monaural signals of rock and jazz music, male and female speech. We used different sets of sound signals for learning basis functions and for generating the mixtures. For the mixture generation, two sentences of the target speakers "mcpm0" and "fdaw0", one for each speaker, were selected from the TIMIT speech database. The training sets were designed to have 21 sentences for each gender, 3 each from 7 randomly chosen males and 7 randomly chosen females. The utterances of the 2 target speakers were not included in the training set. Rock music was mainly composed of guitar and drum sounds, and jazz was generated by a wind instrument. Vocal parts of both music sounds were excluded. Half of the music sound was used for training, half for generating mixtures. All signals were downsampled to 8kHz, from original 44.1kHz (music) and 16kHz (speech). The training data

[5] Variables with ˆ are the estimates of the true values and will be altered by the adaptation formulas. Inputs are fixed, so no ˆ is attached.



(a) Rock music

(b) Jazz music

(c) Male speech

(d) Female speech

Figure 6: Waveforms of four sound sources, from training sets. Audio files for the source signals are available at `http://speech.kaist.ac.kr/˜jangbal/ch1bss/`.

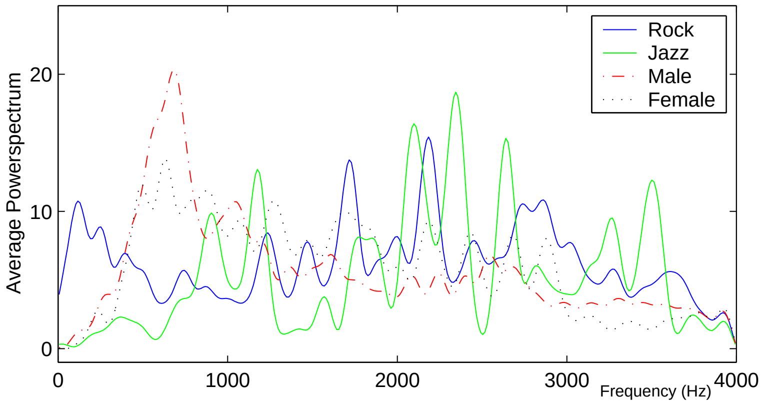


Figure 7: Average powerspectra of the 4 sound sources. Frequency scale ranges in 0∼4kHz (*x*-axis), since all the signals are sampled at 8kHz. The powerspectra are averaged and represented in the *y*-axis.

were segmented in 64 samples (8ms) starting at every sample. Audio files for all the experiments are accessible at `http://speech.kaist.ac.kr/˜jangbal/ch1bss/`.

Figure 6 displays the waveforms of four sound sources used for training—learning basis filters and estimating generalized Gaussian model parameters. We used different data for the separation experiments. Figure 7 compares the four sources by the average spectra. Each covers all the frequency bands, although they are different in amplitude. One might expect that simple filtering or masking cannot separate the mixed sources clearly.

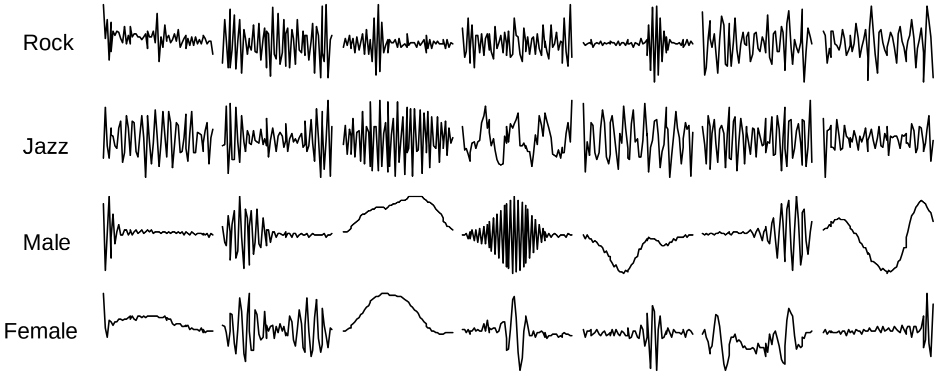


Figure 8: Basis filters learned by ICA. Only 7 basis filters were chosen out of complete sets of 64. The full set of basis filters is available at `http://speech.kaist.ac.kr/~jangbal/ch1bss/`. They are obtained by the generalized Gaussian ICA learning algorithm described in Section 2.2.

4.2 Learned Basis Filters

Subsets of the learned basis filters ($\mathbf{w}_{ik}$) of the four sound types are presented in Figure 8. The adaptation of the generalized Gaussian ICA learning started from a 64×64 square identity matrix, and the gradients of the basis functions were computed on a block of 1000 waveform segments. The parameter q_{ik} for each $p_{\mathbf{g}}(s_{ik}^t)$ was updated every 10 gradient steps. The learned basis filters are generally represented by the superposition of sinusoids of different magnitude and some of them reside only in confined ranges in the time domain. Speech basis filters are oriented and localized in both time and frequency domains, bearing a resemblance to Gabor wavelets (Gaussian-modulated sinusoids). More analysis on the difference between the male and female basis filters can be found in work by Lee and Jang (2001). Jazz basis filters are mostly stationary, but frequently show non-stationary behaviors in terms of amplitude changes in the time axis. Rock basis filters are less stationary and the “drum beats” of the rock music are characterized by abrupt changes in amplitude.

To show the advantage of achieving higher-order probabilistic independence over first-order independence (decorrelation), we performed comparative experiments with the basis filters obtained by PCA, which removes correlations between the output coefficients. Decorrelation is defined as transforming a zero mean vector $\mathbf{x}$ with a matrix $\mathbf{W}$, so that $\mathbf{Wx}$ has an identity covariance matrix. The PCA basis filters are orthogonal and can be obtained from the eigenvectors of the covariance matrix, $\mathbf{W}_p = \mathbf{D}^{-\frac{1}{2}}\mathbf{E}^T$, where $\mathbf{E}$ is a matrix with columns as eigenvectors of the $E[\mathbf{xx}^T]$. Figure 9 shows examples of PCA basis filters for each of the four sound sources. The bases are different from each other since the covariance matrices are from different sets of training data, but the differences are not as significant as those arising in the ICA bases. For speech bases, the PCA basis filters are much more stable in amplitudes and cover the whole time range like the Fourier basis, although the ICA basis filters are localized in time and similar to Gabor wavelets.

In contrast to the data-driven ICA and PCA bases, we also performed experiments with two kinds of basis filters that were fixed over all the sound sources: Fourier and wavelet basis. Speech basis filters learned by ICA behave like Gabor wavelets, and the other data-driven basis filters, except some of the rock basis filters, have similar behaviors to pure sinusoids. Therefore it is

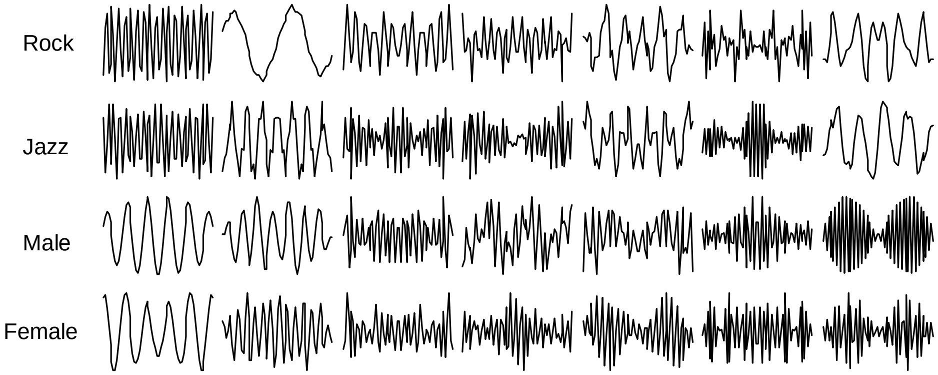


Figure 9: Basis filters obtained by PCA. Only 7 basis filters were chosen out of complete sets of 64. They are obtained by eigenvalue decomposition on the covariance matrix computed from the same training data as used in learning ICA basis filters.

valuable to assess the effectiveness of the real Fourier and the real Gabor wavelet filters to the proposed separation method. In Equation 2 we assumed that the basis filters are real-valued, and hence we adopted a discrete cosine transform (DCT) basis, which gives only real coefficients:

$$s(k) = \sum_{n=1}^{N} x(n) \cos \frac{\pi(k-1)}{2N} (2n-1) \, ,$$

where $k \in [1, N]$ is an index indicating center frequency of the basis filter. A real-valued 1-D Gabor wavelet is a planar sinusoid with a Gaussian envelope, defined by Loy (2002)

$$w(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \cdot \cos(\omega x)$$

where μ and σ respectively determines the position and the width of the Gaussian envelope, and ω is the frequency of the sinusoid. The values of σ and ω are gradually increased as the frequency grows for the set of all the filters to span the whole time-frequency space as it can be seen in the ordinary wavelet basis. Aside from scale, only the ratio between wavelength and the width of the Gaussian envelope can make two Gabor wavelets differ.

Figure 10 shows some examples of DCT and Gabor wavelet bases. DCT basis filters are spread over the time axis and are completely stationary, that is, each of the DCT filters is composed of a single sinusoid of unique frequency. Gabor wavelets are also stationary but reside only in confined ranges in the time domain. In Figures 8 and 9, ICA and PCA basis filters exhibit less regularity. PCA basis filters and Fourier basis filters show similar characteristics, and the ICA basis filters of the two speech signals and the Gabor wavelets also show great resemblance.

4.3 Separation Results of Simulated Mixtures

We generated a synthesized mixture by selecting two sources out of the four and simply adding them. The proposed separation algorithm in Section 3.4 was applied to recover the original sources from a single channel mixture. The source signal estimates were initialized to the values of mixture

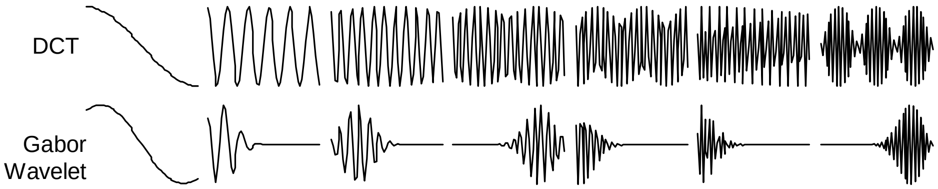


Figure 10: DCT basis filters (first row) and Gabor wavelet basis filters (second row). Only 7 basis filters were chosen out of complete sets of 64. The same set of basis filters are used for all the four sound sources.

signal: $\hat{x}_1^t = \hat{x}_2^t = y^t$. The initial $\hat{\lambda}_i$ were both 0.5 to satisfy $\hat{\lambda}_1 + \hat{\lambda}_2 = 1$. All the samples of the current source estimates were simultaneously updated at each iteration, and the scaling factors were updated at every 10 iterations. The separation converged roughly after 100 iterations, depending on the learning rate and other various system parameters. The procedures of the separation algorithm—traversing all the data and computing gradients—are similar to those of the basis learning algorithm, so their time complexities are likewise of the same order. The measured separation time on a 1.0 GHz Pentium PC was roughly 10 minutes for an 8 seconds long mixture.

The similarity between the original source signals and the estimated sources is measured by signal-to-noise ratio (SNR), which is defined by

$$\mathrm{snr}_s(\hat{s})\ [\mathrm{dB}] = 10 \log_{10} \frac{\sum_t s^2}{\sum_t (s - \hat{s})^2},$$

where s is the original source and $\hat{s}$ its estimate. To qualify a separation result we use the sum of the SNRs of the two recovered source signals: $\mathrm{snr}_{x_1}(\hat{x}_1) + \mathrm{snr}_{x_2}(\hat{x}_2)$. Table 1 reports SNR results for the four different bases. In terms of average SNR, the two data-driven bases performed better than

Table 1: SNR results of the proposed method. (R, J, M, F) stand for rock, jazz music, male, and female speech. '*mix*' column lists the symbols of the sources that are mixed to y, and the values in the other columns are the SNR sums, $\mathrm{snr}_{x_1}(\hat{x}_1) + \mathrm{snr}_{x_2}(\hat{x}_2)$, measured in dB. The first line of each column indicates the used method to obtain the basis filters. "GW" stands for Gabor wavelet. Audio files for all the results are accessible at `http://speech.-kaist.ac.kr/~jangbal/ch1bss/`.

mix	DCT	GW	PCA	**ICA**
R + J	0.7	1.3	6.9	**13.0**
R + M	3.0	2.1	4.7	**8.9**
R + F	3.0	1.8	5.8	**8.8**
J + M	7.2	5.8	9.3	**10.3**
J + F	8.1	5.9	10.9	**10.4**
M + F	4.8	3.3	4.9	**5.9**
Average	4.5	3.4	7.1	**9.6**

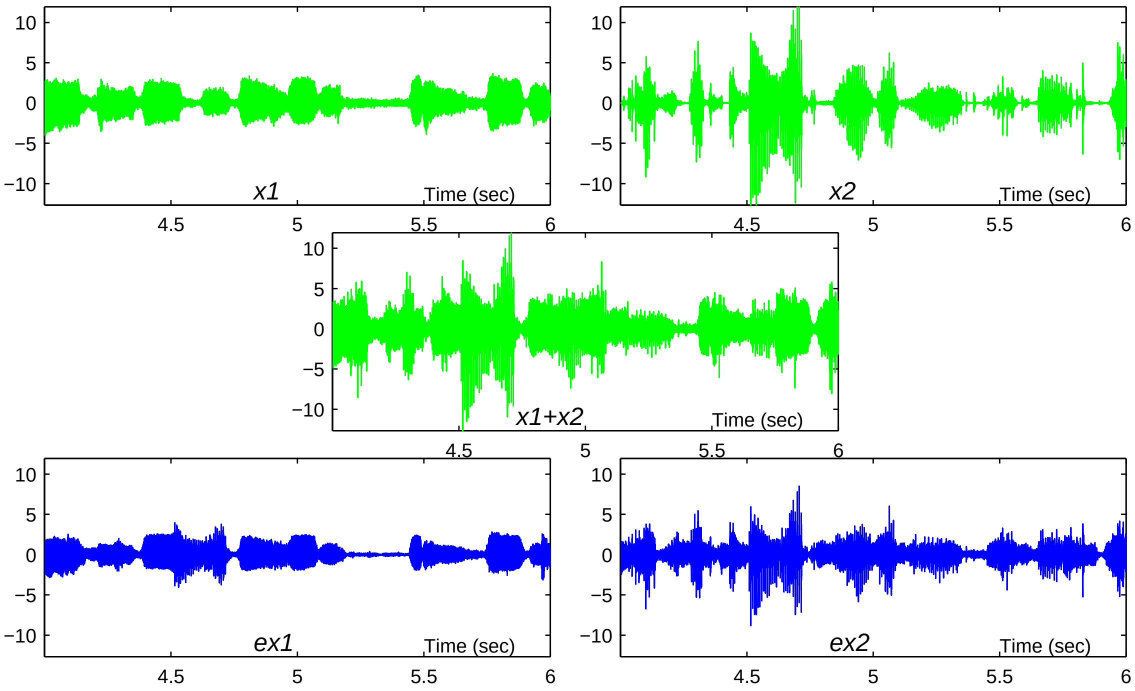


Figure 11: Separation results of jazz music and male speech. In vertical order: original sources (x_1 and x_2), mixed signal ($x_1 + x_2$), and the recovered signals.

the two fixed bases, and the ICA basis displayed the best performance. Moreover, the ICA basis guaranteed a certain degree of SNR performance for all the cases, whereas the performances of the two fixed bases and PCA basis varied greatly according to the mixed sound sources. The SNR of jazz and female mixture separation for the PCA basis was better than for the ICA basis, although the other mixtures were badly separated. DCT and Gabor wavelet basis showed very good SNRs for the mixtures of jazz music compared to the other mixtures. The likely explanation for this is that jazz music is very close to stationary, and as a result PCA and ICA induce jazz music basis filters of similar characteristics (Figures 8 and 9), and those basis filters resemble DCT basis filters. Although Gabor wavelet filters are localized in time, they are also from sinusoids, so they represent jazz music well in comparison with the other source signals. Generally, mixtures containing jazz music were recovered comparatively cleanly, and the male-female mixture was the least recovered. With regard to rock music mixtures, the SNR differences between ICA basis and the other bases were much larger than those of other mixtures. This is because the drum beats (abrupt changes in amplitude) are expressed well only in the ICA basis filters.

Figure 11 illustrates the waveforms of the original sources and the recovered results for the mixture of jazz music and male speech, and Figure 12 shows for the mixture of male and female speech. Their SNR sums were 10.3 and 5.9. The separation of speech-speech mixture was much poorer than those of music-speech mixtures. From the experimental results, we conclude that the demixing performance highly relies on the basis functions. The estimates of the source signals, mixed and observed in a single channel, are projected on each of the bases sets, and the sources are

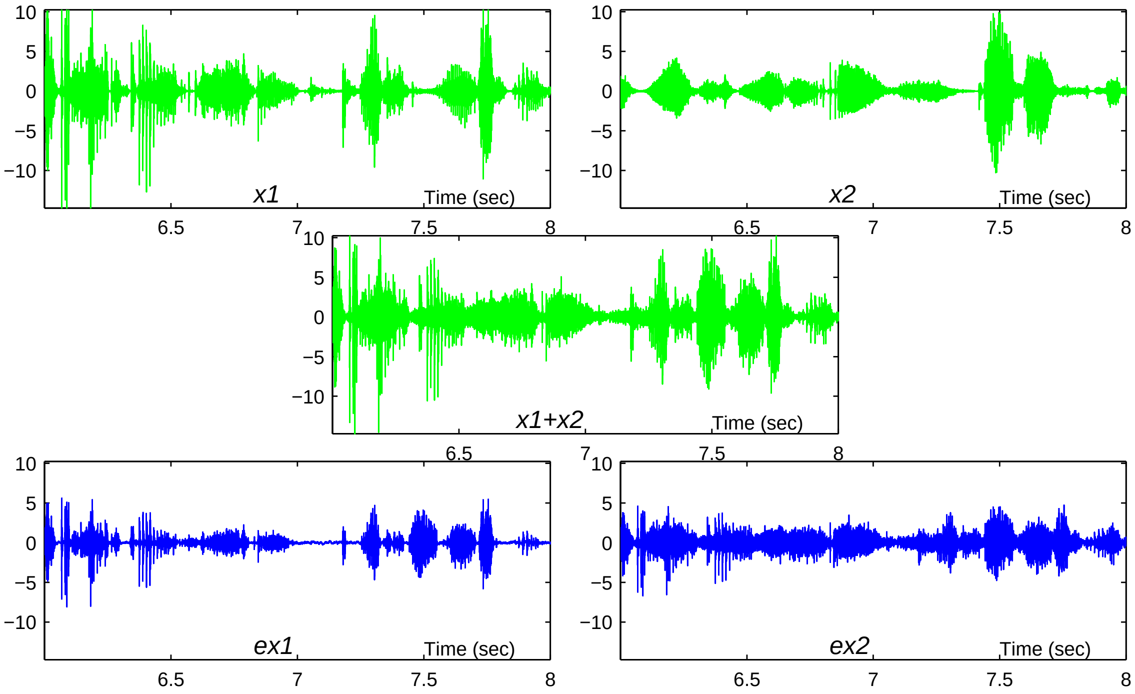


Figure 12: Separation results of male and female speech signals.

isolated by iteratively approaching the maximally probable projections. Although it is difficult to define a similarity between two sets of basis filters in a probabilistic sense, minimizing differences in the projecting directions of the basis filters is crucial for the success of the separation algorithm. Based on the learned basis functions shown in Figure 8, there is seemingly too much overlap in the signal space between two speakers for our algorithm to ever work for mixtures of speech. ICA found sets of bases that explain the class of the music signals well, but performed poorly in explaining the class of the speech signals. Speech basis functions vary in amplitudes frequently in the time domain, and the coefficient distributions are extremely sparse. These characteristics are caused by the nonstationary nature of the speech signal. In contrast, as can be seen in Figure 8, the amplitudes of the music signals are comparatively stable, and the basis functions cover a longer range in the time axis. The coefficient distributions are less sparse than those of speech basis functions, which is analogous to earlier findings, such as in Bell and Sejnowski (1995).

4.4 Comparison to Wiener Filtering

It is very difficult to compare a separation method with other CASA techniques; their approaches are vastly different in many ways such that an optimal tuning of their parameters would be beyond the scope of this paper. Instead, in order to assess the separability of our method, we performed experiments that show how close the separation performance of the proposed algorithm approaches the theoretical *"limit"*.

Hopgood and Rayner (1999) proved that a stationary signal—a finite bandwidth signal in the Fourier domain—can be described by a set of autocorrelation functions, in which only second-order statistics are considered. In that case, least square estimates of the source signals mixed in a single channel can be obtained by a linear time invariant (LTI) Wiener filter, and an optimal separation under stationarity assumptions is achieved when the Wiener filter is derived from the autocorrelations of the original sources. Although Wiener filtering has the disadvantage that the estimation criterion is fixed and depends on the stationarity assumptions, Hopgood and Rayner (1999) stated that time-varying Wiener filter formulation enables separation of nonstationary sources:

$$W_i(\omega,t) = \frac{\hat{X}_i(\omega,t)}{\hat{X}_1(\omega,t)+\hat{X}_2(\omega,t)},$$

where $\hat{X}_i(\omega,t)$ is the estimate of the powerspectum of source i at frequency ω and time t. When true source signals are available, Wiener filtering can be regarded as a theoretical upperbound of the frequency-domain techniques. Parra and Spence (2000) also stated that second-order statistics at multiple times capture the higher-order statistics of nonstationary signals, which supports that time-varying Wiener filters provide statistical independence of mixed source signals. The higher-order statistical structures of the sound sources are also captured by the ICA basis filters; therefore we compared the performances of the proposed method and the time-varying Wiener filtering.

The construction of the time-varying Wiener filter requires the powerspectra of true source signals. We use the powerspectra of the original sources in the mixture when computing Wiener filters, whereas the basis filters used in the proposed method are learned from training data sets that are not used in the mixture generation. The Wiener filter approach in this case can be regarded as a separation upperbound. The filters were computed every block of 0.5, 1.0, 2.0, and 3.0 sec. Their

Table 2: Comparison of the proposed method with Wiener filtering. '*mix*' column lists the symbols of the sources that are mixed to the input. (R, J, M, F) stand for rock, jazz music, male, and female speech. "Wiener" columns are the evaluated SNRs grouped by the block lengths (in seconds), and the filters are computed from the average powerspectrum of each block. The last column lists the SNRs of the proposed method, and the last row is the average. Note that the Wiener filters are from testing data that are actually mixed into input mixture, however the basis filters of the proposed method are from the training data that are different from testing data. The comparison is to show how close to an optimal solution the proposed method is. The performance of the proposed method was closest to Wiener filtering at the block length 1.0s. Audio files for all the results are accessible at `http://speech.kaist.ac.kr/~jangbal/ch1bss/`.

mix	Wiener				proposed
	0.5s	1.0s	2.0s	3.0s	
R + J	11.1	10.3	9.4	9.1	13.0
R + M	8.7	8.1	7.1	7.3	8.9
R + F	10.1	8.9	8.2	7.3	8.8
J + M	13.5	11.9	10.0	10.5	10.3
J + F	14.1	11.0	9.4	8.4	10.4
M + F	9.9	8.5	7.8	6.1	5.9
Average	11.2	**9.8**	8.6	8.1	**9.6**

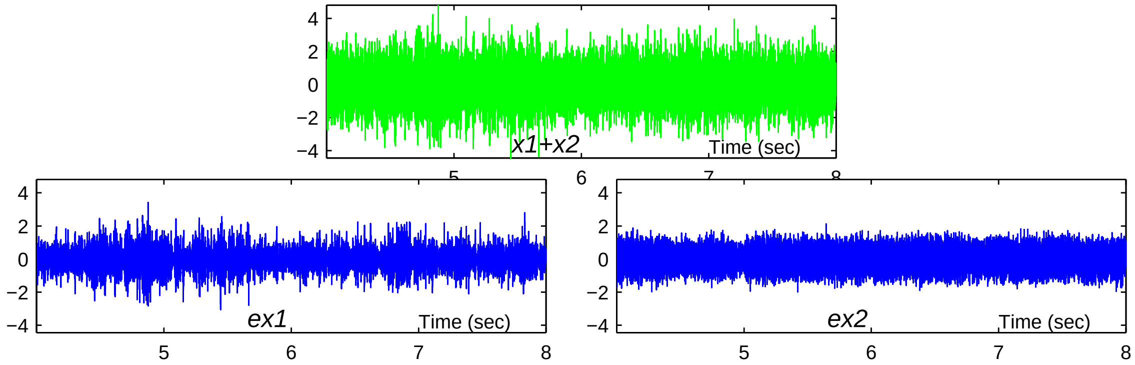


Figure 13: Separation result of real recording. Input signal is at the top of the figure, below are the recovered male speech and the noise.

performances are measured by SNRs and compared to the proposed method in Table 2. In terms of average SNR, our blind results were comparable in SNR with results obtained when the Wiener filters were computed every 1.0 sec.

4.5 Experiments with Real Recordings

We have tested the performance of the proposed method on recordings in a real environment. Data were recorded in a diffuse sound room. Four speakers were employed, one in each corner of the sound room. A signal played through these speakers produces a uniform sound field throughout the room. The recorded signals were composed of a male speech utterance on the background of very loud noise. The level of noise was so high that even human listeners could hardly recognize what was spoken. The estimated SNR was about -8 dB. The focus of this experiment is to recover human speech in real recordings, to assess the performance of the proposed separation algorithm in a real environment.

The basis functions of general TIMIT male speakers (Figure 8) are used for the recorded male speaker. Since we do not know exactly the characteristics of the noise source, we assumed two different types of well-known noisy signals: Gaussian white noise and pink noise. White noise has a uniform spectrogram all over frequency axis as well as over time axis. Pink noise is similar but the power decreases exponentially as the frequency increases. The spectrogram of the pink noise resembles the noisy recording in our case. (The noise sources, their basis functions and spectrograms, the recorded sound, and the separated results are available at the provided website: `http://speech.kaist.ac.kr/~jangbal/ch1bss/`.) The algorithm did not work with the white noise, but it successfully recovered the original sources with the pink noise; the waveforms are displayed in Figure 13. Although a "perfect separation" was not attained, it should be noted that the input was too noisy and we did not have the true basis functions. To achieve better separation in real environments, it is necessary to have a large pool of bases for various kinds of natural sounds and to find the most characterizing basis for a generic problem.

5. Discussions

This section investigates a number of issues: comparison to other methods, separability, and more detailed interpretations of experimental results. Future research issues such as dealing with more than two source signals are also discussed at the end of the section.

5.1 Comparison to other Separation Methods

Traditional approaches to signal separation are classified as either spectral techniques or time-domain nonlinear filtering methods. Spectral techniques assume that source signals are disjoint in the spectrogram, which frequently result in audible distortions of the signal in the regions where the assumption mismatches. Roweis (2001) presented a refiltering technique which estimates time-varying masking filters that localize sound streams in a spectro-temporal region. In his work sound sources are supposedly disjoint in the spectrogram and there exists a “mask” that divides the mixed multiple streams completely. A somewhat similar technique is proposed by Rickard, Balan, and Rosca (2001). They did not try to obtain the “exact” mask but an estimate by a ML-based gradient search. However, being based on the strong assumption in the spectral domain, these methods also suffer from the overlapped spectrogram.

To overcome the limit of the spectral methods, a number of time-domain filtering techniques have been introduced. They are based on splitting the whole signal space into several disjoint and orthogonal subspaces that suppress overlaps. The criteria employed by the former time-domain methods mostly involve second-order statistics: least square estimation (Balan et al., 1999), minimum mean square estimation (Wan and Nelson, 1997), and Wiener filtering derived from the autocorrelation functions (Hopgood and Rayner, 1999). The use of AR (autoregressive) models on the sources has been successful. Balan et al. (1999) assume the source signals are $\mathrm{AR}(p)$ processes, and they are inferred from a monaural input by a least square estimation method. Wan and Nelson (1997) used AR Kalman filters to enhance the noisy speech signals, where the filters were obtained from neural networks trained on the specific noise. These methods performed well with input signals well-suited to the AR models, for example speech signals. However that is also a major drawback to applying them to the real applications. Moreover they consider second-order statistics only, which restricts the separable cases to orthogonal subspaces (Hopgood and Rayner, 1999).

Our method is also classified as a time-domain method but avoids these strong assumptions by virtue of higher-order statistics. There is no longer orthogonality constraint of the subspaces, as the basis functions obtained by the ICA algorithm are not restricted to being orthogonal. The constraints are dictated by the ICA algorithm that forces the basis functions to result in an efficient representation, that is, the linearly independent source coefficients; both the basis functions and their corresponding pdfs are key to obtaining a faithful MAP based inference algorithm. The higher-order statistics of the source signal described by a prior set of basis functions capture the inherent statistical structures.

Another notable advantage is that the proposed method automatically generates the prior information. While the other single channel separation methods also require the prior information, their methods to characterize the source signals are dependent on the developer’s intuitive knowledge, such as harmonic structures or empirical psycho-acoustics. In contrast, our method exploits ICA for the automation of characterizing source signals. The basis functions can be generated whenever appropriate learning data are available, which may not be identical to the separation data. The training data and the test data are different in our experiments and the mixtures were successfully separated.

5.2 Comparison to Multichannel Signal Separation

The proposed method requires the basis functions of the mixed source signals. In order to expand its applications to real world problems, a "dictionary" of bases for various kinds of natural sounds and finding the most characterizing basis in the dictionary for a generic case are necessary conditions to achieve good separation performance. These requirements make it more difficult to apply the proposed method to a real world problem than the conventional BSS techniques, which neither make assumptions nor require information about the source signals. However BSS suffers from the necessity of multiple channel observations, at least 2 channel observations are required, while the proposed method deals with single channel observations. The role of the basis functions is in some sense a substitute for extra-channel input.

5.3 Separability

The problem stated here is one of finding sets of bases that explain one class of signal well. The estimates of the source signals, mixed and observed in a single channel, are projected on each of the bases sets, and sources are isolated by iteratively approaching the maximally probable projections. Although it is difficult to define a similarity between two sets of basis filters in a probabilistic sense, minimizing differences in the projecting directions of the basis filters is crucial for the success of the separation algorithm. Based on the learned basis functions, it seems that there is too much overlap in signal space between two speakers for our algorithm to ever work for mixtures of speech. One way around this obstacle would be to do the separation in some feature space where there is both better class separation, and the possibility of transformation back to signal space. Future work will be to find more distinguishable subspaces, and develop better criteria for learning the source signals. The current process of training the bases is non-discriminative. It would seem advantageous to train the sets of bases discriminatively. This would bring the separation results closer to the theoretical limitation.

5.4 More than Two Sources

The method can be extended to the case when $P > 2$. We should decompose the whole problem into $P = 2$ subproblems, because the proposed algorithm is defined only in that case. One possible example is a sequential extraction of the sources: if there is a basis that characterizes a generic sound, that is, which subsumes all kinds of sound sources, then we use this basis and the basis of the target sound that we are interested in extracting. The separation results are expected to be the target source and the mixture of the remaining $P-1$ sources. Repeating this extraction $P-1$ times yields the final results. Another example is merging bases: if there is a method to merge a number of bases and we have all the individual bases, we can construct a basis for Q sources and the other for the remaining $P-Q$ sources. Then we can split the mixture into two submixtures. Likewise repeating the split yields the final separation. In summary, the case $P > 2$ can be handled but additional research such as building a generic basis or merging different bases is required.

6. Summary

We presented a technique for single channel source separation utilizing the time-domain ICA basis functions. Instead of traditional prior knowledge of the sources, we exploited the statistical structures of the sources that are inherently captured by the basis and its coefficients from a training set. The algorithm recovers original sound streams through gradient-ascent adaptation steps pursuing the maximum likelihood estimate, computed by the parameters of the basis filters and the generalized Gaussian distributions of the filter coefficients. With the separation results of the real recordings as well as simulated mixtures, we demonstrated that the proposed method is applicable to real world problems such as blind source separation, denoising, and restoration of corrupted or lost data.

Our current research includes the extension of this framework to perform model comparisons to estimate the optimal set of basis functions to use given a dictionary of basis functions. This is achieved by applying a variational Bayes method to compare different basis function models to select the most likely source. This method also allows us to cope with other unknown parameters such the as the number of sources. Future work will address the optimization of the learning rules towards real-time processing and the evaluation of this methodology with speech recognition tasks in noisy environments, such as the AURORA database.

References

Samer A. Abdallah and Mark D. Plumbley. If the independent components of natural images are edges, what are the independent components of natural sounds? In *Proceedings of International Conference on Independent Component Analysis and Signal Separation (ICA2001)*, pages 534–539, San Diego, CA, December 2001.

Shun-Ichi Amari and Jean-Francois Cardoso. Blind source separation — semiparametric statistical approach. *IEEE Transactions on Signal Processing*, 45(11):2692–2700, 1997.

Radu Balan, Alexander Jourjine, and Justinian Rosca. AR processes and sources can be reconstructed from degenerate mixtures. In *Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation (ICA99)*, pages 467–472, Aussois, France, January 1999.

Anthony J. Bell and Terrence J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1004–1034, 1995.

Anthony J. Bell and Terrence J. Sejnowski. Learning the higher-order structures of a natural sound. *Network: Computation in Neural Systems*, 7(2):261–266, July 1996.

Anthony J. Bell and Terrence J. Sejnowski. The “independent components” of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

Pau Bofill and Michael Zibulevsky. Underdetermined blind source separation using sparse representations. *Signal Processing*, 81(11):2353–2362, 2001.

George E. P. Box and G. C. Tiao. *Baysian Inference in Statistical Analysis*. John Wiley and Sons, 1973.

Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge MA, 1990.

Albert S. Bregman. *Computational Auditory Scene Analysis*. MIT Press, Cambridge MA, 1994.

Guy J. Brown and Martin Cooke. Computational auditory scene analysis. *Computer Speech and Language*, 8(4):297–336, 1994.

Jean-Francois Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters*, 4(4):112–114, April 1997.

Jean-Francois Cardoso and Beate Laheld. Equivariant adaptive source separation. *IEEE Transactions on Signal Processing*, 45(2):424–444, 1996.

Pierre Comon. Independent component analysis, A new concept? *Signal Processing*, 36:287–314, 1994.

Daniel P. W. Ellis. A computer implementation of psychoacoustic grouping rules. In *Proceedings of the 12th International Conference on Pattern Recognition*, 1994.

David J. Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.

James R. Hopgood and Peter J. W. Rayner. Single channel signal separation using linear time-varying filters: Separability of non-stationary stochastic signals. In *Proceedings of ICASSP*, volume 3, pages 1449–1452, Phoenix, Arizona, March 1999.

Aapo Hyvärinen. Sparse code shrinkage: denoising of nongaussian data by maximum likelihood estimation. *Neural Computation*, 11(7):1739–1768, 1999.

Kunio Kashino and Hidehiko Tanaka. A sound source separation system with the ability of automatic tone modeling. In *Proceedings of Interational Computer Music Conference*, pages 248–255, 1993.

Jong-Hwan Lee, Ho-Young Jung, Te-Won Lee, and Soo-Young Lee. Speech feature extraction using independent component analysis. In *Proceedings of ICASSP*, volume 3, pages 1631–1634, Istanbul, Turkey, June 2000a.

Te-Won Lee, Mark Girolami, Anthony J. Bell, and Terrence J. Sejnowski. A unifying information-theoretic framework for independent component analysis. *computers & mathematics with applications*, 31(11):1–21, March 2000b.

Te-Won Lee and Gil-Jin Jang. The statistical structures of male and female speech signals. In *Proceedings of ICASSP*, Salt Lake City, Utah, May 2001.

Te-Won Lee and Michael S. Lewicki. The generalized Gaussian mixture model using ICA. In *International Workshop on Independent Component Analysis (ICA'00)*, pages 239–244, Helsinki, Finland, June 2000.

Michael S. Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356–363, 2002.

Michael S. Lewicki and Terrence J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12:337–365, 2000.

Gareth Loy. Fast computation of the gabor wavelet transform. In *Digital Image Computing Techniques and Applications*, Melbourne, Australia, January 2002.

David J. C. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. *Report, University of Cambridge, Cavendish Lab*, August 1996.

Hiroshi G. Okuno, Tomohiro Nakatani, and Takeshi Kawabata. Listening to two simultaneous speeches. *Speech Communications*, 27:299–310, 1999.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

Lucas Parra and Clay Spence. Convolutive blind separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, 8:320–327, May 2000.

Barak A. Pearlmutter and Lucas Parra. A context-sensitive generalization of ICA. In *Proceedings of ICONIP*, pages 151–157, Hong Kong, September 1996.

Dinh Tuan Pham and P. Garrat. Blind source separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Transactions on Signal Processing*, 45(7):1712–1725, 1997.

Scott Rickard, Radu Balan, and Justinian Rosca. Real-time time-frequency based blind source separation. In *Proceedings of International Conference on Independent Component Analysis and Signal Separation (ICA2001)*, pages 651–656, San Diego, CA, December 2001.

Sam T. Roweis. One microphone source separation. *Advances in Neural Information Processing Systems*, 13:793–799, 2001.

Eric A. Wan and Alex T. Nelson. Neural dual extended Kalman filtering: Applications in speech enhancement and monaural blind signal separation. In *Proceedings of IEEE Workshop on Neural Networks and Signal Processing*, 1997.

DeLiang L. Wang and Guy J. Brown. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks*, 10:684–697, 1999.

Michael Zibulevsky and Barak A. Pearlmutter. Blind source separation by sparse decomposition. *Neural Computations*, 13(4), 2001.