Blind Source Separation of More Sources Than Mixtures Using Overcomplete Representations

Te-Won Lee, *Member, IEEE,* Michael S. Lewicki, Mark Girolami, *Member, IEEE,*
and Terrence J. Sejnowski, *Senior Member, IEEE*

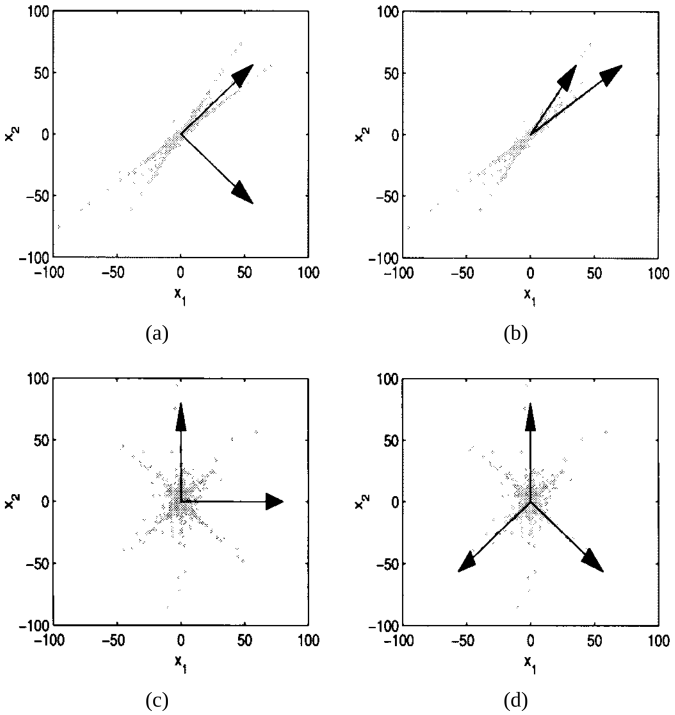
***Abstract*—Empirical results were obtained for the blind source separation of more sources than mixtures using a recently proposed framework for learning overcomplete representations. This technique assumes a linear mixing model with additive noise and involves two steps: 1) learning an overcomplete representation for the observed data and 2) inferring sources given a sparse prior on the coefficients. We demonstrate that three speech signals can be separated with good fidelity given only two mixtures of the three signals. Similar results were obtained with mixtures of two speech signals and one music signal.**

***Index Terms*—Blind source separation, independent component analysis, overcomplete dictionary, overcomplete representation, speech signal separation.**

I. Introduction

Recent advances in blind source separation by independent component analysis (ICA) have many potential applications including speech recognition systems, telecommunications, and medical signal processing. The goal of ICA is to recover independent sources given only sensor observations that are unknown linear mixtures of the unobserved independent source signals [3]–[5], [8].

The standard formulation of ICA requires at least as many sensors as sources. Lewicki and Sejnowski [9], [11] have proposed a generalized ICA method for learning overcomplete representations of the data that allows for more basis vectors than dimensions in the input. The goal of this method is illustrated in Fig. 1. In a two-dimensional (2-D) data space, the observations $\mathbf{x}$ in Fig. 1(a) and (b) were generated by a linear mixture of two independent random sparse sources. In this space, Fig. 1(a) shows orthogonal basis vectors (principle component analysis, PCA) and Fig. 1(b) shows independent basis vectors. If the 2-D observed data are generated by three sparse sources, as shown in Fig. 1(c) and (d), the complete ICA representation (c) cannot model the data adequately but the overcomplete ICA representation (d) finds three basis vectors that fit the underlying distribution of the data.



(a)

(b)

(c)

(d)

Fig. 1. Illustration of basis vectors in a 2-D data space with two sparse sources (top) or three sparse sources (bottom). (a) PCA finds orthogonal basis vectors. (b) ICA representation finds independent basis vectors. (c) ICA cannot model the data distribution adequately with three sources, but (d) the overcomplete ICA representation finds three basis vectors that match the underlying data distribution (see [11]).

In this letter, the learning rules for overcomplete ICA are briefly summarized in Section II, as derived by Lewicki and Sejnowski [11]. In Section III, simulation results are presented for speech signals and music signals. The discussion in Section IV covers related work and future research issues.

II. Learning Overcomplete Representations

The observed M-dimensional data $\mathbf{x} = [x_1, \cdots, x_M]^T$ may be modeled as a linear overcomplete mixing matrix, $\mathbf{A}$, $(M \times N)$[1] with additive noise.

$$x = \mathbf{As} + \mathbf{n} \tag{1}$$

Manuscript received August 31, 1998. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. K. Buckley.

T.-W. Lee is with Howard Hughes Medical Institute, Computational Neurobiology Laboratory, The Salk Institute, La Jolla, CA 92037 USA (e-mail: tewon@salk.edu; lewicki@salk.edu).

M. S. Lewicki was with Howard Hughes Medical Institute, Computational Neurobiology Laboratory, The Salk Institute, La Jolla, CA 92037 USA. He is now with Carnegie Mellon University, Pittsburgh, PA 15213 USA.

M. Girolami is with the Department of Computing and Information Systems, University of Paisley, Paisley PA1 2BE, U.K. (e-mail: giro0ci@paisley.ac.uk).

T. J. Sejnowski is with Howard Hughes Medical Institute, Computational Neurobiology Laboratory, The Salk Institute, La Jolla, CA 92037 USA, and also with the Department of Biology, University of California, San Diego, La Jolla, CA 92093 USA.

Publisher Item Identifier S 1070-9908(99)02461-X.

[1] In most ICA formulations, the matrix $\mathbf{A}$ is restricted $M \geq N$, which is not imposed here.

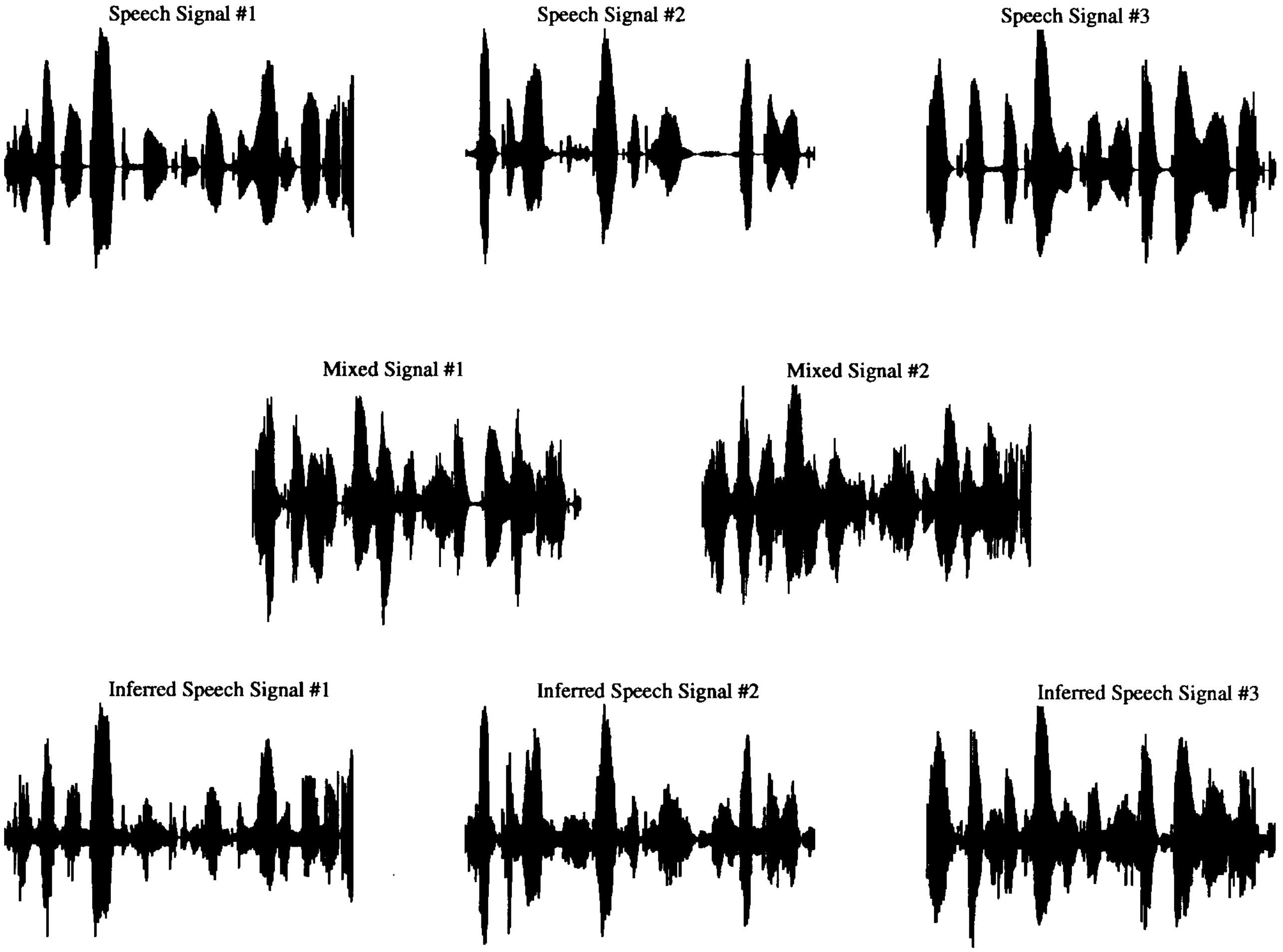


Fig. 2. Demonstration of the separation of three speech signals from two mixtures. Top row: time course of three speech signals. Middle row: two observations of three mixed speech signals. Bottom row: inferred speech signals.

where $\mathbf{s} = [s_1, \cdots, s_N]^T$ are the sources and $\mathbf{n}$ is assumed to be a white Gaussian noise with variance σ^2 so that

$$\log P(\mathbf{x}|\mathbf{A}, \mathbf{s}) \propto -\frac{1}{2\sigma^2}(\mathbf{x} - \mathbf{A}\mathbf{s}). \tag{2}$$

It is also assumed that the sources s_i are mutually independent, so that the joint probability distribution has the form $P(\mathbf{s}) = \prod_{i=1}^{M} P(s_i)$, and each source s_i has a sparse distribution, such as the Laplacian density $P(s_i) \propto \exp(-\alpha|s_i|)$.

Given the above model and assumptions, the goal is to infer both the basis vectors $\mathbf{A}$ and the sources $\mathbf{s}$ given the mixtures $\mathbf{x}$.

A. Inferring the Sources

Due to the additive noise and the rectangular mixing matrix $\mathbf{A}$, the solution for $\mathbf{s}$ cannot be found by the pseudo-inverse $\mathbf{s} = \mathbf{A}^{+}\mathbf{x}$. A probabilistic approach to estimating the sources is based on finding the maximum *a posteriori* value of $\mathbf{s}$:

$$\begin{aligned}\hat{\mathbf{s}} &= \max_{\mathbf{s}} P(\mathbf{s}|\mathbf{x}, \mathbf{A}) \\ &= \max_{\mathbf{s}} P(\mathbf{x}|\mathbf{A}, \mathbf{s})P(\mathbf{s}).\end{aligned} \tag{3}$$

Given basis vectors $\mathbf{A}$, and observation $\mathbf{x}$, (3) can be optimized by gradient ascent on the log posterior distribution [9], [11].

B. Learning the Basis Vectors

The objective for learning the basis vectors, $\mathbf{A}$, is to maximize the probability of the data.

$$P(\mathbf{x}_1, \cdots, \mathbf{x}_T|\mathbf{A}) = \prod_{i=1}^{T} P(x_i|\mathbf{A}) \tag{4}$$

which assumes temporal independence of the samples. Computation of the likelihood requires marginalizing over all possible sources

$$P(\mathbf{x}|\mathbf{A}) = \int P(\mathbf{x}|\mathbf{A}, \mathbf{s})P(\mathbf{s})\, d\mathbf{s}. \tag{5}$$

For general overcomplete bases, this integral is intractable. For the special case of zero noise and $\mathbf{A}$ invertible (a complete basis), the integral in (5) is solvable and leads to the standard ICA learning algorithm [3], [4]. Lewicki and Sejnowski [11] approximated (5) by fitting a multivariate Gaussian around $\hat{\mathbf{s}}$. The basis vectors were learned by performing gradient ascent on the log of (4) using the approximation of (5). The learning rule is

$$\Delta\mathbf{A} \propto \mathbf{A}\mathbf{A}^T \frac{\partial}{\partial \mathbf{A}} \log P(\mathbf{x}|\mathbf{A}) \approx -\mathbf{A}\big(\phi(\hat{\mathbf{s}})\hat{\mathbf{s}}^T + \mathbf{I}\big) \tag{6}$$

where $\phi(\hat{s}_i) = \partial \log P(\hat{s}_i)/\partial \hat{s}_i$ is called the score function, and $\mathbf{I}$ is the identity matrix. The prefactor $\mathbf{A}\mathbf{A}^T$ produces the

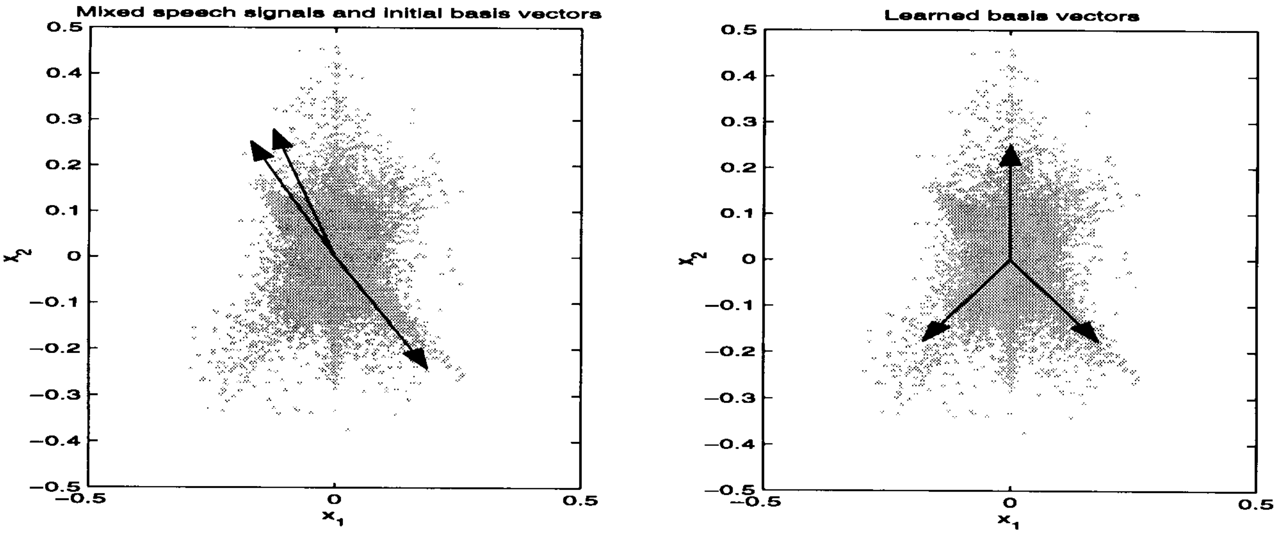


Fig. 3. Left: Two-dimensional scatter plot of the two mixed signals. The three basis vectors were randomly initialized. Right: After convergence, the learned basis functions are shown as arrows along the three speech signals. The learned basis vectors may be permutated and have a different sign.

natural gradient extension [1], [2] which speeds convergence. The matrix $\mathbf{A}$ in (6) is not restricted to be square and thus works for overcomplete representation. The derivation is described in Lewicki and Sejnowski [11]. Note that each gradient step requires computation of $\hat{s}$ as in (3).

III. Experimental Results

A. Blind Separation of Speech Signals

Speech signals with silent time segments are sparsely distributed and are approximated by a Laplacian model. Three speech signals from the same speaker, sampled at 8 kHz with 8 bits per sample, were taken from the TIMIT database and are shown in Fig. 2 (top). We mixed the three speech signals into two mixtures:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 1 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \\ s_3(t) \end{bmatrix}. \tag{7}$$

Fig. 2 (middle) shows the time course of the two mixed speech signals. The 2-D scatter plot (x_1 against x_2) in Fig. 3 (left) shows the three directions of the data. The three basis vectors of $\mathbf{A}$ were initially chosen randomly and were learned using (6). The learning process converged after 50 iterations. When more than three basis vectors were chosen, the amplitude of the redundant basis vectors converged to zero. The noise level l was set to three bits out of eight, i.e., the maximum amplitude of the noise signal was $2^3/2^8 \approx 3\%$ of the data range. Fig. 3 (right) shows the learned basis vectors. The sources were inferred using (3) and were recovered up to permutation and sign. Fig. 2 (bottom) shows the three inferred speech signals after reordering and sign correction. The signal-to-noise ratio (SNR) for the separation was 20, 17, and 21 dB, respectively. Experiments with different speech signals and different mixing matrices yielded similar results. Although the temporal structure of the speech signal was not taken into consideration in the model, the separation quality was good.[2]

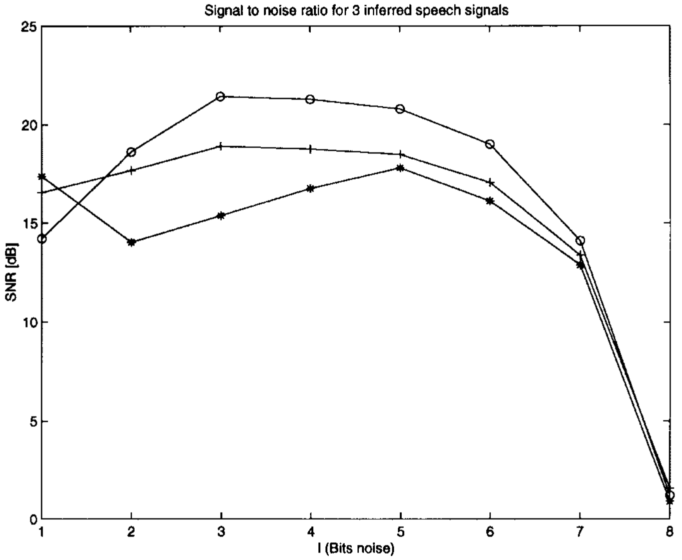


Fig. 4. SNR as a function of noise level l. * = speech signal 1; + = speech signal 2, and o = speech signal 3.

The assumed noise level, l, determines whether a data point should be considered as noise or as signal. A high noise level ignores a wide range of data points around zero and puts more weight on outliers when finding the basis vectors and when inferring the sources. This is significant in case of additive noise, where we may adjust the appropriate noise level to infer the sources. Fig. 4 shows the SNR as a function of the noise-level l. Reasonably good SNR results were obtained for noise levels up to 6 b (a maximum of the noise amplitude of 25%) and the performance degraded rapidly for a noise level of 7 or 8 b.

[2] The original, mixed, and inferred signals are available at http//www.cnl.salk.edu/~tewon/Over.

We also applied the method of Lin *et al.* [12] (see Section IV) to this dataset. They inferred the sources by assuming that there was only one nonzero source at a given time sample. Using this method, the SNR decreased by 4, 2, and 7 dB, respectively.

B. Preliminary Results with Other Mixtures

We performed further speech mixing experiments with varying number of sources and sensors. With two mixtures we extracted up to four mixed speech signals but the algorithm failed to find correct basis vectors when more than four sources were mixed into two observations. However, five speech signals were extracted from observations of three mixtures. We were also able to separate one music and two speech signals from two mixtures, although the Laplacian density model may be less accurate.

Overcomplete representations can be extended to learn structure in high-dimensional data space. For example, Lewicki and Olshausen [10] used a two times overcomplete representation to find 2 × 144 basis vectors for 12 × 12 patches of natural images.

The formulation used here may also be used to unmix signals that were mixed with additive noise as assumed for the model in (1). Preliminary results indicate that overcomplete ICA can recover highly noisy mixtures and obtain a reasonable SNR. For noise levels of 4 to 7 b, the ICA algorithm used here recovered two mixed speech signals with additive Gaussian noise with a 5 to 10 dB improvement in SNR compared to the standard ICA [3].

IV. Discussion

A. Comparison to Other Methods

The problem of separating more sources than observations has been treated by several other methods. For the special case of binary sources, Pajunen [13] used a maximum likelihood approach to reduce the problem to finding M clusters in the mixture space, and Hermann and Yang [7] applied self-organizing maps to find the clusters for binary sources. Lin *et al.* [12] proposed a method for continuous signals by employing image analysis tools to detect geometric structure of the 2-D mixture data locating the extremal density directions and thus finding the basis vectors. The sources were inferred by assuming that there was only one nonzero output at a given time, i.e., each data point was assigned to one source with the closest basis vector and all other sources were set to zero. In our experiments, this inference method gave poor SNR for the speech separation example. The overcomplete ICA approach can be applied to continuous signals and is not restricted to binary sources. Furthermore, the probabilistic framework allows more flexible models, which might lead to more accurate inferences. Another approach to finding the basis vectors is to use cumulants [6], which was not explored here.

B. Conclusions

We have shown here that overcomplete representations can be used for blind source separation when there are more sources than mixtures. Reasonably good separations were obtained for two mixtures of three speech signals and for two mixtures of two speech signals and one music signal. Overcomplete representations reduce to ICA when the number of mixtures is equal to or greater than the number of sources. We are currently investigating the use of overcomplete representations of EEG data for artifact removal and for neural signal detection from a small number of sensors.

References

[1] S. Amari, “Neural learning in structured parameter spaces,” in *Advances in Neural Information Processing Systems*, vol. 9. Cambridge, MA: MIT Press, 1997, pp. 127–133.

[2] ——, “Natural gradient works efficiently in learning,” *Neural Comput.,* vol. 10, pp. 251–276, 1998.

[3] A. J. Bell and T. J. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution,” *Neural Comput.,* vol. 7, pp. 1129–1159, 1995.

[4] J.-F. Cardoso, “Blind signal separation: Statistical principles,” *Proc. IEEE,* vol. 86, pp. 2009–2025, Oct. 1998.

[5] P. Comon, “Independent component analysis—A new concept?,” *Signal Process.,* vol. 36, pp. 287–314, 1994.

[6] P. Comon and B. Mourrain, “Decomposition of quantics in sums of powers of linear forms,” *Signal Process.,* vol. 53, pp. 93–107, Sept. 1996.

[7] M. Hermann and H. Yang, “Perspectives and limitations of self-organizing maps,” in *Proc. ICONIP’96.*

[8] T.-W. Lee, *Independent Component Analysis: Theory and Applications.* Boston, MA: Kluwer, 1998.

[9] M. Lewicki and T. J. Sejnowski, “Learning nonlinear overcomplete representations for efficient coding,” in *Advances in Neural Information Processing Systems,* vol. 10. Cambridge, MA: MIT Press, 1998, pp. 815–821.

[10] M. S. Lewicki and B. Olshausen, “Inferring sparse, overcomplete image codes using an efficient coding framework,” submitted for publication.

[11] M. S. Lewicki and T. J. Sejnowski, “Learning overcomplete representations,” *Neural Comput.,* to be published.

[12] J. Lin, D. Grier, and J. Cowan, “Feature extraction approach to blind source separation,” in *Proc. IEEE Workshop on Neural Networks for Signal Processing,* 1997, pp. 398–405.

[13] P. Pajunen and J. Karhunen, “A maximum likelihood approach to nonlinear blind separation,” in *ICANN*, Lausanne, Switzerland, 1997, pp. 541–546.